2021

A Two-Level Information Modelling Translation Methodology and Framework to Achieve Semantic Interoperability in Constrained GeoObservational Sensor Systems

Paul Stacey
Technological University Dublin, paul.stacey@tudublin.ie

A Two-Level Information Modelling Translation Methodology and Framework to Achieve Semantic Interoperability in Constrained Geo-Observational Sensor Systems.

By

Paul Stacey BEng (Hons), MPhil

A thesis submitted to Technological University Dublin, for the degree of Doctor of Philosophy.

Supervised by Dr. Damon Berry

School of Electrical & Electronic Engineering

TU Dublin – City Campus

Kevin Street, Dublin 8

June 2021

Abstract

As geographical observational data capture, storage and sharing technologies such as in situ remote monitoring systems and spatial data infrastructures evolve, the vision of a Digital Earth, first articulated by Al Gore in 1998 is getting ever closer. However, there are still many challenges and open research questions. For example, data quality, provenance and heterogeneity remain an issue due to the complexity of geo-spatial data and information representation.

Observational data are often inadequately semantically enriched by geo-observational information systems or spatial data infrastructures and so they often do not fully capture the true meaning of the associated datasets. Furthermore, data models underpinning these information systems are typically too rigid in their data representation to allow for the ever-changing and evolving nature of geo-spatial domain concepts. This impoverished approach to observational data representation reduces the ability of multi-disciplinary practitioners to share information in an interoperable and computable way.

The health domain experiences similar challenges with representing complex and evolving domain information concepts. Within any complex domain (such as Earth system science or health) two categories or levels of domain concepts exist. Those concepts that remain stable over a long period of time, and those concepts that are prone to change, as the domain knowledge evolves, and new discoveries are made. Health informaticians have developed a sophisticated two-level modelling systems design approach for electronic health documentation over many years, and with the use of *archetypes*, have shown how data, information, and knowledge interoperability among heterogenous systems can be achieved.

This research investigates whether two-level modelling can be translated from the health domain to the geo-spatial domain and applied to observing scenarios to achieve semantic interoperability within and between spatial data infrastructures, beyond what is possible with current state-of-the-art approaches.

A detailed review of state-of-the-art SDIs, geo-spatial standards and the two-level modelling methodology was performed. A cross-domain translation methodology was developed, and a proof-of-concept geo-spatial two-level modelling framework was defined and implemented. The Open Geospatial Consortium's (OGC) Observations & Measurements (O&M) standard was re-profiled to aid investigation of the two-level information modelling approach. An evaluation of the method was undertaken using

specific use-case scenarios. Information modelling was performed using the two-level modelling method to show how existing historical ocean observing datasets can be expressed semantically and harmonized using two-level modelling. Also, the flexibility of the approach was investigated by applying the method to an air quality monitoring scenario using a technologically constrained monitoring sensor system.

This work has demonstrated that two-level modelling can be translated to the geo-spatial domain and then further developed to be used within a constrained technological sensor system; using traditional wireless sensor networks, semantic web technologies and Internet of Things based technologies. Domain specific evaluation results show that two-level modelling presents a viable approach to achieve semantic interoperability between constrained geo-observational sensor systems and spatial data infrastructures for ocean observing and city based air quality observing scenarios. This has been demonstrated through the re-purposing of selected, existing geospatial data models and standards. However, it was found that re-using existing standards requires careful ontological analysis per domain concept and so caution is recommended in assuming the wider applicability of the approach.

While the benefits of adopting a two-level information modelling approach to geospatial information modelling are potentially great, it was found that translation to a new domain is complex. The complexity of the approach was found to be a barrier to adoption, especially in commercial based projects where standards implementation is low on implementation road maps and the perceived benefits of standards adherence are low. Arising from this work, a novel set of base software components, methods and fundamental geo-archetypes have been developed. However, during this work it was not possible to form the required rich community of supporters to fully validate geo-archetypes. Therefore, the findings of this work are not exhaustive, and the archetype models produced are only indicative. The findings of this work can be used as the basis to encourage further investigation and uptake of two-level modelling within the Earth system science and geo-spatial domain. Ultimately, the outcomes of this work are to recommend further development and evaluation of the approach, building on the positive results thus far, and the base software artefacts developed to support the approach.

Keywords: two-level modelling, archetypes, information modelling, GIS, geospatial, standards, internet of things, observations and measurements, semantics, GIScience, resource constrained devices, knowledge-based systems, interoperability.

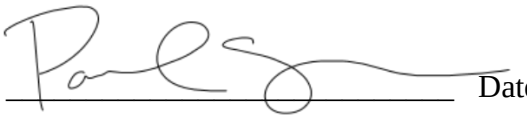
Declaration

I certify that this thesis which I now submit for examination for the award of Doctor of Philosophy, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work.

This thesis was prepared according to the regulations for graduate study by research of Technological University Dublin and has not been submitted in whole or in part for another award in any other third level institution.

The work reported on in this thesis conforms to the principles and requirements of TU Dublin's guidelines for ethics in research.

TU Dublin has permission to keep, lend or copy this thesis in whole or in part, on condition that any such use of the material of the thesis be duly acknowledged.

Signature  Date 02/06/2021

Acknowledgements

First and foremost, I would like to thank my supervisor Dr. Damon Berry for his enduring patience, sharing his in-depth knowledge and experience of two-level modelling and stimulating philosophical discussions, which have been invaluable throughout this PhD study. Damon's encouragement and guidance throughout the ups and downs of undertaking a PhD in a part-time mode have been pivotal in completing this work. Thank you!

I wish to thank my extended family, my wife Elaine and our five sons Luca, Wil, Ethan, Jonah, and Elijah for their patience and understanding throughout this work. I look forward to uninterrupted weekends, and earlier home arrivals to spend more time with you all!

When this work began the School of Electrical & Electronic Engineering at TU Dublin – City Campus, and the Department of Engineering at TU Dublin – Blanchardstown Campus belonged to sperate institutions but have now merged under the umbrella of TU Dublin. Both have provided me with continued encouragement and support throughout this work. I would like to thank the Environmental Health Sciences Institute (ESHI) on the Grangegorman campus for a quiet and pleasant place to work, and the space to just be a student.

I would like to thank Dr. Adam Leadbetter from the Marine Institute, Ireland for first introducing me to Al Gore's Digital Earth concept back in 2015 and his informative discussions and advice, including comments and feedback on the early draft of the Earth Science Informatics Journal manuscript.

I would like to acknowledge and thank the IoT FIT Lab for free use of their experimental wireless sensor network facilities.

Finally, I would like to thank Thomas Beale for sharing his perspectives and experiences on two-level modelling adoption within the health domain and providing invaluable guidance on translating two-level modelling to new domains in an all too short Skype discussion, and My main take away, *if you build it they will come*…. Eventually.

Abbreviations List

6LBR	6LoWPAN/RPL Border Router
ACDD	Attribute Conventions for Data Discovery
ADL	Archetype Description Language
AJAX	Asynchronous JavaScript and XML
AM	Archetype Model
AMM	Archetype Modelling Methodology
ANSI	American National Standards Institute
AOM	Archetype Object Model
API	Application Programming Interface
AQL	Archetype Query Language
$AQL_{contiki}$	Antelope Query Language
ARM	Advanced RISC Machine
BFO	Basic Formal Ontology
BNF	Backus–Naur form
CAP	Consistency Availability Partition (tolerance)
CCGI	Citizen Contributed Geographical Information
cADL	Constraint Archetype Description Language
CEM	Clinical Element Model
CEN	Comité Européen de Normalisation
CF	Climate Forecast
Chlfa	Chlorophyll-a
CIMI	Clinical Information Modelling Initiative
CKM	Clinical Knowledge Management System
CMEMS	Copernicus Marine Environment Monitoring Service's
CO	Carbon Monoxide
CoAP	Constrained Application Protocol
COP	Common Operational Picture
CRUD	Create Read Update Delete
CSML	Climate Science Modelling Language
CSS	Cascading Style Sheets
CSV	Comma Separated Values
DA	Data Assimilation
dADL	Data Archetype Description Language
DATAMEQ	Data Management Exchange and Quality Working Group

DBMS	Database Management System
DCM	Detailed Clinical Model
DCSM	Dutch Continental Shelf Model
DKM	Domain Knowledge Management System
DOLCE	Descriptive Ontology for Linguistic and Cognitive Engineering
DSL	Domain Specific Language
DUL	DOLCE Ultra Lite
DV	Data Value
ECHO	Earth Observing System Clearing House
EF	Environmental Monitoring Facilities
EHR	Electronic Healthcare Record
EMODnet	European Marine Observation and Data Network
EO	Earth Observation
ER	Entity Relationship
ESS	Earth System Science
EU	European Union
EXI	Efficient XML Interchange
FAIR	Findable, Accessible, Interoperable and Reusable
FIFO	First In First Out
FIT	Future Internet of Things
FOPL	First Order Predicate Logic
FOV	Field of View
FRAM	Ferroelectric Random-Access Memory
GAM	Generalized Additive Models
GCM	Generic Conceptual Model
GEHR	Good Electronic Healthcare Record
GEMET	General Multilingual Environmental Thesaurus
GEO-DAB	GEO Data Access Broker
GEOSS	Global Earth Observation System of Systems
GEOMS	Generic Earth Observation Metadata Standard
GIS	Geographical Information System
GOOS	Global Ocean Observing System
GOR	Geo Observations Record
GORM	Grails' Object Relational Mapping
GPS	Global Positioning System

GRA	GOOS Regional Alliances
GIRM	Generalised Identity Reference Model.
GSP	Groovy Server Pages
HATEOAS	Hypermedia as the Engine Application Stack
HTTP	Hypertext Transfer Protocol
IAB	Internet Architecture Board
ICS	International Commission on Stratigraphy
ICT	Information Communications Technology
IEEE	Institute of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force
IN STAC	In Situ Thematic Centre
INSPIRE	Infrastructure for Spatial Information in Europe
IOC	Intergovernmental Oceanographic Commission
IOC/IODE	IOC/ International Oceanographic Data and Information Exchange
IoT	Internet of Things
IP	Internet Protocol
IPCC	Intergovernmental Panel on Climate Change
IR	Implementing Rule
ISO	International Organization for Standardization
ISS	International Space Station
JAX-RS	Java API for Restful Web Services
JSON	JavaScript Object Notation
LD	Linked Data
MIG	Maintenance and Implementation Group
MIPS	Million Instructions Per Second
MSDI	Marine Spatial Data Infrastructure
MSFD	Marine Strategy Framework Directive
MSI	MultiSpectral Instrument
MVP	Model View Controller
NASA	National Aeronautics and Space Administration
NERC	Natural Environment Research Council
NetCDF	Network Common Data Format
NIST	National Institute of Standards & Technology
NMMP	National Marine Monitoring Programme
NOOS	North West European Shelf Operational Oceanographic System

NTNU	Norwegian University of Science and Technology
NWS	North West Shelf
O&M	Observations and Measurements
ODIP	Ocean Data Interoperability Platform
ODM	Observations Data Model
ODP	Open Data Portal
OF	Oceanographic geographical Features
OGC	Open Geospatial Consortium
OID	Object Identifier
OLCI	Ocean and Land Colour Instrument
OMG	Object Management Group
OO	Object Oriented
OPT	Operational Templates
OPTaaS	Operational Templates as a Service
OS	Operating System
OWL	Ontology Web Language
PSC	Planning Support Concept
QC	Quality Control
RAM	Random Access Memory
RDAC	Regional Data Acquisition Centres
RDF	Resource Description Framework
RDFS	Resource Description Framework Schema
RESTful	Representational State Transfer
RFC	Request for Comment
RM	Reference Model
RMSE	Root Mean Square
ROM	Read Only Memory
RPL	Routing Protocol for Low Power and Lossy Networks
SDGs	Sustainable Development Goals
SDI	Spatial Data Infrastructure
SDR	Sensor Data Record
SEEK	Singular Evolutive Extended Kalman Filter
SNOMED	Systematised Nomenclature of Medicine
SOAP	Simple Object Access Protocol
SOSA	Sensor, Observation, Sample, and Actuator Ontology

SOS	Sensor Observation Service
SPARQL	Simple Protocol and RDF Query Language
SQL	Structured Query Language
SRAM	Static Random-Access Memory
SSH	Secure Shell
SSNO	Semantic Sensor Network Ontology
SUMO	Suggested Upper Merged Ontology
SWE	Sensor Web Enablement
SWEET	Semantic Web for Earth and Environment Technology
TOM	Template Object Model
UDP	User Datagram Protocol
UML	Unified Modelling Language
UN	United Nations
UNESCO	United Nations Educational, Scientific and Cultural Organization
UNOOSA	United Nations Office for Outer Space Affairs
UNPF	United Nations Population Fund
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
US	United States (of America)
UUID	Universally Unique Identifier
WGA	Working Group on the Anthropocene
WLAN	Wireless Local Area Network
WMO	World Meteorological Organization
WoT	Web of Things
WSN	Wireless Sensor Network
WWW	World Wide Web
XHTML	Extensible Hypertext Markup Language
XMI	XML Metadata Interchange
XML	Extensible Markup Language
XSD	XML Schema Definition

Table of Contents

List of Figures

List of Code Listings

List of Tables

List of Equations

Chapter 1

"Yes the planet got destroyed. But for a beautiful moment in time we created a lot of value for shareholders"
(Tom Toro, New Yorker 2015)

1. INTRODUCTION

The world is experiencing a period of unprecedented and profound geographical and climatic change, which has the potential to be harmful and catastrophically disruptive to the Earth and all its occupants (Houghton et al., 1990) (Watts et al., 2019). The Earth sciences community is at the forefront of the global response to monitoring, understanding, and communicating this change (Solomon et al., 2007) (Edenhofer, 2014) (Pontin, 2020). This communication is critical, as it informs how society and those that govern society should react and adapt (Howarth, Parsons, and Thew, 2020).

In the future, society will increasingly rely on Earth sciences and Earth scientists to be able to make informed and critical decisions that consider the changing nature of the world around us. To enable the Earth science community to meet this global challenge, there is a need for high quality geospatial data and information.

Capturing, representing, processing, and analysing complex geospatial/geographical data and information is the domain of geographical information scientists (Goodchild, 2010). Geographical information scientists have a need to gather and combine data from many sources and in various ways to enable geospatial *convergence research* teams (Kedron et al., 2021), who in turn synthesize a new understanding of our physical world, producing new knowledge (Gahegan and Pike, 2006). In the future, geographical information scientists will increasingly need to extract knowledge from unstructured and

structured geospatial data to help meet the needs of the Earth sciences and scientists (Breunig et al., 2020).

Remote in situ environmental monitoring sensor deployments are one source of valuable environmental and geo-spatial observational data. Environmental monitoring sensor networks have the potential to transform Earth science (Hart and Martinez, 2020). However, these observational systems are often built in isolation, and their resultant data representations (metadata) are often not adequately designed for re-use and higher order knowledge generation. Knowledge relating sensed observational data captured by in situ sensor deployments is often hidden in sensor manuals and field operator logs (Fredericks and Botts, 2018). Also, remote in situ sensor systems are often technological constrained with limited power, communications, and processing ability. This is especially the case for deployments in harsh remote environments such as within a marine environment (Xu et al., 2019) or hazardous environments (monitoring volcanic process, landslides, avalanches etc.) (Hart and Martinez, 2020). These technical constraints often limit any kind of onboard rich data representation being applied at source. This lack of inherent interoperability in heterogeneous datasets produced by constrained sensor observing systems represents a missed opportunity for us all to benefit from the advancement of knowledge about our changing environment and planet.

The Berlin Declaration on Open Access to Knowledge in Sciences and Humanities (Borges, 2008) seeks to promote the Internet and Web as a functional instrument to promote and advance human knowledge. Open access to data and knowledge can also act as a key economic driver. Pooling existing resources can save significant amounts of public money. For example, the European Union green paper on Marine knowledge 2020 strategy (European Commission, 2012) estimates that a shared marine data infrastructure consisting of high-quality marine data collected by EU public bodies could save €1Billion

per year. However, as we shall see, there are many barriers to building such data infrastructures; marine or otherwise, and consequently discoverable and interoperable data are the focus of much research (Columbus Consortium, 2016).

There are many data standards that allow what is termed *syntactic* interoperability, and the sharing of remote and in situ sensor systems observational data, such as the Open Geospatial Consortium's (OGC) suite of standards[1]. However, data heterogeneity remains a pervasive problem in geo observational information infrastructures, and *semantic interoperability* (the next level beyond syntactic interoperability) remains a work in progress.

Data heterogeneity is characterised by the many different coding formats, constraint models and storage solutions used to capture, share and persist data. Data heterogeneity leads to a missed opportunity for organisations and businesses to create value leveraged off the fusion of rich datasets.

In complex domains such as such as health and Earth systems science-based sub domains (e.g. oceanography etc.) knowledge is constantly evolving. Capturing volatile domain specific knowledge concepts in an observational system and supporting information management infrastructures, invariably leads to a mismatch between the needs of the domain practitioner (marine scientist for example) and the versatility and expressiveness of the concepts represented. The core issue is the inflexible representation of domain concepts and how they are managed over time as they evolve.

This work investigates the approaches used to model and standardise geospatial data and information and the aspects that leads to inflexibility in concept representation. A proposal to adapt and translate an existing flexible modelling approach, known as *two-*

[1] The OGC is a worldwide consortium that develop and publish standards for location based technologies. https://www.ogc.org/

level modelling (Beale, 2002), from the equally complex domain of health to solve some of the issues identified within geospatial information infrastructures is investigated. Two-level modelling introduces *archetypes* to address core issues of interoperability, standardisation and flexible concept representation within health-based information systems.

Like all good stories, we start at the beginning with an outline of the background and motivations for this work, describing to the reader the inherent complexities within geospatial data and the need to investigate solutions. Also contained in this chapter are the research problem, hypothesis, research question, objectives, methodology and contributions arising.

1.1 Background and Motivations

Humans are currently experiencing a rare epochal event. We exist at the transition of geological timescales. Geological timescales relate geological strata (stratigraphy) to time (Stoppani, 1873 cited in Hamilton and Grinevald, 2015). This system of geological timescales is used by Earth scientists to map the relationship of events relating to Earth's history to a chronological period. Modern day life i.e. social structures, demographics and more besides have evolved within a geological time frame called the *Holocene*.

The Holocene is part of the Quaternary period. It arrived approximately 11,700 years ago, after the Pleistocene epoch (Williams et al., 1997). The Holocene has provided us with relatively stable and predictable patterns of climatic events, upon which modern agricultural methodologies and practices rely (Mayewski et al., 2004) (Wanner et al., 2008). Advances in agricultural practices have afforded humans the ability to greatly progress as a species at an increasingly impressive rate. The net effect of these advancements has been the ability for us to grow the human population. With this exponential growth, our ability to impact the Earth around us has increased.

The Industrial Revolution brought about the first measurable global impact of humans (Crutzen and Stoermer, 2000). Large amounts of carbon dioxide were released by the burning of fossil fuels to power the Industrial Revolution causing globally measurable deposits to occur. Over the intervening time, the quantity and quality of measurements has improved. Today, the effects of our rapid expansion on the Earth around us has become extensive and the systematic measurement of these effects has also increased.

Such is the extent of change driven by human activities, the Earth is crossing a new geological boundary. Humans have become such a significant geological force, contributing to a huge amount of geological change, beyond anything the Earth has experienced in its 4.5 billion- year history, the term "*anthropocene*", meaning the age of humans, is being used to define the current geological epoch (Crutzen, 2002) (Crutzen, 2006). Crutzen's claim of a new human-influenced epoch was further backed up by work done by Zalasiewicz et al. (2008). In August 2016, the British-led Working Group on the Anthropocene (WGA) declared its support for the new epoch by stating its belief that it began in 1950. The WGA's work gives weight to the likelihood of the anthropocene being ratified by the International Commission on Stratigraphy (ICS) in the future[2].

The investigations of the ICS on the merits of defining a new human influenced epoch (anthropocene) may continue for some time. In the meantime, what can be said is that current human activity has a great influence on the Earth's climatic and geological processes. As a result, it is more important than ever to understand the reciprocal nature of this relationship between human led processes and that of the Earth system. This need to take a holistic view of the Earth system gave rise to the super discipline called *Earth System Science* (NASA, 1986).

[2] As of October 2020, the WAG states that "The Anthropocene is not currently a formally defined geological unit within the Geological Time Scale; officially we still live within the *Meghalayan* Age of the *Holocene* Epoch" http://quaternary.stratigraphy.org/working-groups/anthropocene/

1.1.1 Earth System Science

Traditionally natural sciences investigated and attempted to understand physical, chemical and biological processes independently. Today a more planetary approach is the norm. This relatively new way to examine natural processes is referred to as Earth Systems Science (ESS). ESS portrays the Earth as an intricate network of interrelated entities (NASA, 1986). In ESS, Earth is viewed as a complex, evolving planet that is characterised by continuously interacting physical and biological change (Mackenzie, 2010). Changes within Earth's processes occur across a wide range of geo-spatial and temporal scales. Quantifying and understanding the extent of change between interrelated Earth processes in terms of time, space and scale is important for making higher-level decisions (for example relating to human populations and biological related industries such as agri/aqua-culture).

Areas near and around the Earth's surface are divided into categories called geospheres (Williams et al., 2012). There are four natural geospheres: *lithosphere*, *hydrosphere*, *biosphere*, and *atmosphere*. The four geospheres are named from derivations of their Greek meaning: stone (litho), air (atmo), water (hydro), and life (bio). As a result of humankind's evolution, another 5th sphere – the *anthroposphere* – is used to capture economic, political and social growth. Humankind's interaction with the Earth's surface to achieve growth in these areas affects the other four geospheres. In fact, the *anthroposphere* (previously referred to as the technosphere) conflicts with the other geospheres (Milsum, 1968).

Using the scientific method and the holistic approach of Earth system science, the complex functioning of the system of Earth can be evaluated. The Scientific method (as applied to ESS) seeks to present an understanding of Earth's phenomena that is reliable, consistent, and non-arbitrary. Four basic steps must be employed:

1) Observations & Description

2) Formulation of hypotheses.

3) Prediction of other phenomenon by using formulated hypothesis.

4) Performance of appropriate experimental tests of the predictions.

Observations and description (step 1 above) of Earth's phenomena can be achieved through gathering observable, empirical, and measurable evidence.

Typically, these phenomena events do not tend to occur in isolation. To understand the wider consequences and contexts of natural phenomena, it is also necessary to examine observations from multiple locations or historical events. Combining observed datasets in real-time allows for the derivation of higher-level information across a range of observations. Combining datasets in this way requires the formation of "data communities", and sometimes in an ad-hoc fashion. These data communities may be a combination of real-time data streams from multiple independent sources (sensor-webs for example) along with near-real-time and historical datasets. The ability to find and bind observational data regarding Earth's phenomena requires a - yet to be achieved - globally connected geospatial information cyber-infrastructure.

1.1.2 Digital Earth

In 1998 US Vice President Al Gore set out a vision for what he termed "Digital Earth" (Gore, 1998). Gore's vision was a challenge to a diverse global community to enable the increasing amount of raw geospatial data to be combined and processed into understandable information. To achieve the Digital Earth vision, Gore highlighted the need to break into the growing vast silos of geo-data and make these datasets accessible and suitable for *secondary use*.

In many ways the Digital Earth paradigm is motivated by the ideals of the prominent 20th century scientist Michael Polanyi. Polanyi once famously said "we know more than

we can tell" (Polanyi, 1941). Polanyi was referring to the difficulty in the transfer of knowledge between humans by verbal means. Through the sharing of information and knowledge, new knowledge can be derived; this is true for all domains.

Polanyi was also a proponent of the idea of spontaneous order within science. Spontaneous order refers to an environment where systems of researchers can form to tackle specific problems and discover new knowledge. As knowledge evolves within a community, researchers may adjust their direction or behaviour, forming new orders in response to change. It is argued that spontaneous - as opposed to structured - order is much more conducive to the production of new meaning.

A Digital Earth system would be of great benefit to Earth System Science based domain specialists. Within the Earth System Science domain multi-disciplinary ESS Scientists have a need to combine data and information from many sources and in various ways (ideally computable ways) to synthesize new understanding and document new knowledge (Di et al., 2002). A Digital Earth system represents a natural platform to enable Earth System Scientists to document and share knowledge, and conceivably form on-the-fly communities of practice (spontaneous order). However, the realisation of a Digital Earth as defined by Gore is difficult to achieve in practice; and is still a work in progress (Craglia et al., 2012) (Boulton, 2018). Dangermond & Goodchild (2019) describe Digital Earth as an instance of a *digital twin*. Digital twins are real world objects replicated in a digital environment. Earth's digital twin (digital Earth) should capture the earth visually but in principle should also replicate how the Earth works in all its complexity.

Today, Polanyi's statement: *we know more than we can tell* is still valid. Since the onset of the Information Revolution, the medium of how we tell has changed greatly. However, our ability to share human knowledge even via modern digital mediums is still a significant challenge. In many ways, the recording of human knowledge has still not

surpassed the mediums of old - the mighty book - and the recording of a narrative expressed in natural language. However, what has changed are the mechanisms to allow the sharing of recorded knowledge. The standardisation of information systems has greatly improved the ability to widely disseminate data and information. Adoption of standards has allowed the efficient sharing of information within participating communities, allowing large, diverse and geographically disparate knowledge communities to exist.

The need to share data is not unique to the Digital Earth paradigm. The United Nations Sustainable Development Goals (SDGs) (UN, 2016) require diverse communities to now work closely together to meet the 17 defined goals. Rahimifard and Trollman (2018) give an engineering perspective to the SDGs, highlighting the need to enable knowledge transfer between diverse disciplines:

> *The complex nature of such SDGs often necessitates solutions based on complex systems that will require wide-ranging skills, lateral thinking, and knowledge transfer between various social, life and physical sciences as well as engineering disciplines.*

There is a general trend towards generated (including sensed) data being available online using the Web as a mechanism for dissemination (Jirka and Stasch, 2018). The pooling of datasets allows richer knowledge and information to be derived across inter-related data gathering activities. However, one aspect that is particular to the area of this work, is the need to gather data about natural phenomena that occur in an ad-hoc fashion. Conversely, these phenomena and the associated recorded events do not tend to occur in isolation. Human-induced changes can also occur unexpectedly and within a short time scale, but a spatially large scale (example: Chernobyl disaster, 1986).

The publishing, sharing and combination of related data/information from different information systems within a domain has been shown to be invaluable in decision making, analysis etc. A classic example of how widespread adoption of standard information formalisms and access methods can be incredibly valuable is the World Wide Web (WWW) (Berners-Lee, 1992). As the Web grew in popularity and scale, Web content search engines emerged to enable users find content more effectively. Initially, the Web was indexed by hand but as the amount of Web content grew this was no longer practical. Although Internet search engines existed prior to the existence of the Web, the Archie search engine is often cited as the first Web search engine. Seymour, Frantsvog and Kumar (2011) document a comprehensive history of the evolution of the first Web search engines. Modern search engines utilise widely used and standardised metadata that are formatted according to the XHTML specification. The growth of online open data appears to be mirroring the evolution of the Web and Web tools such as search engines, although it is at a slower pace, perhaps due to the higher complexity of standardisation of open datasets. In September 2018 Google launched "Google Dataset Search"[3], Google's first search engine dedicated to quickly finding open datasets on the Web.

Much-heralded terms such as "Connected Data", "Sensor Net", "Ubiquitous Computing" or "Internet of Things" are accompanied by a desire to publish and enable the integration of multiple sensor data-streams, along with data from historical monitoring events; to facilitate the generation of higher-level information. Consequently, the approach for creating sensor data and information has changed over the past decade. This new emerging paradigm has the expressed goal of enabling standardised sensor service interfaces and standardised datasets. These standardised sensor interfaces and datasets enable real-time sensor data to become accessible and shareable in a uniform way.

[3] https://toolbox.google.com/datasetsearch

Much progress has been made in the past number of years in tackling these key areas. The Open Geospatial Consortium's Sensor Web Enablement (SWE) (Botts et al., 2008) framework was a starting point in dealing with the challenge of making data available in a uniform way. The SWE defines a suite of Web services interfaces and common protocols abstracting from the heterogeneity of sensor communication. This goes some way towards supporting the possibility of disparate geo-spatial knowledge communities for discovering, sharing, and analysing observed data. However, there are still many challenges to be addressed. The SWE does not describe in detail how to integrate sensors and their data on-the-fly with minimal human interaction. Substantial effort is required to make a sensor and its observations available on the Web. Furthermore, the challenges of interoperability within information systems goes beyond the syntactical approach offered by the SWE (Bröring et al., 2011).

One of the key remaining challenges is the issue of semantic heterogeneity of sensed geo-spatial data and information and any resulting recorded knowledge. Mechanisms to integrate and exchange recorded knowledge through the sharing of data and information which use different data models and different ontological schemes are still under development. The problem of integrating disparate geospatial observational data and information is a major barrier to achieving the vision of Al Gore's Digital Earth (Guo et al., 2020). Strengthening the role of semantics development and implementation is now seen as essential to realising the Digital Earth vision (Schade et al., 2020).

Geo-spatial information is diverse, as are datasets related to the broad domain of ESS. Within this complex network of information sources are quantitative sensor-generated geospatial information. These types of data are generally captured from in situ and remote sensing systems, deployed pervasively and constantly sensing and reporting information about phenomena in the world around us. As mentioned above, this type of information

is hugely important to the ESS community to better understand historical and current dynamic processes about the Earth's complex system. However, the pervasive and heterogeneous and sometimes ad hoc nature of sensor-based monitoring systems means integration of datasets is either not possible or else very difficult. Ideally, integrated data and the secondary use of sensed data should be considered from the start of systems development to allow larger datasets to be merged and a richer view of Earth's processes to be possible.

Required are agreed international standards to ensure interoperability. Organisations such as the Open Geospatial Consortium have been advancing this agenda for many years, but the work is complex and slow and there are still many problems to solve.

Having given the reader a broad overview of the background and motivations for this work, the research problem statement is defined next as well as the aims and objectives of this project. Throughout the remainder of this chapter several core ideas are briefly referred to. These will be covered in more detail in later chapters. However, in this chapter, the reader will be introduced to only what is necessary to understand the aims and objectives of the project.

1.2 Problem Statement

Organisations such as the Open Geospatial Consortium (OGC) define standardised interfaces such as the sensor web enablement (SWE) framework suite of standards to allow interoperability between heterogeneous sensor network systems. However, the information model used to represent core observational data is *semi-structured* (i.e. loosely defined to allow for broad usage).

Semi-structured models are an improvement on standard models, but introduce additional problems; strong typing is lost, the model is still partially concrete as assumptions about the information are made and encoded. These encoded assumptions

used to capture environmentally sensed data lead to *conflation*[4]. With conflation, differences within data can become lost as the data move up through the data value chain. This approach is limited in its ability to enable interoperability of geospatial observational data, and results in low quality datasets and becomes a barrier to enabling datasets to be combined within any form of Digital Earth system.

Interoperability of information is a problem within all domains where islands of information exist. Practitioners within the geospatial domain, although they deal with information that is unique in many ways, need to look beyond their own domain to examine solutions developed in other complex domains (Diviaccio and Leadbetter, 2017). In fact, the health domain provides a wealth of experience and techniques that may prove useful in solving the issue of semantic interoperability of geospatial observational data and systems (Stacey and Berry, 2015) (Diviaccio and Leadbetter, 2017).

Non-technical practitioners such as geographers, oceanographers or indeed any Earth system scientist need some mechanism to be able to contribute their knowledge and experience to the development of the information systems they use in their daily work. In the past object-oriented analysis and design processes have attempted to elicit requirements using these types of users, but as discussed later in chapter 3 these processes have limited success.

Within the health domain, the issues detailed above also exist. The health domain can be seen as an analogue to the geospatial domain in terms of its complexity of its information, diversity of its domain experts and the many islands of information that exist across a vast array of heterogenous information systems. One solution that has been

[4] Errors resulting from conflation occur when two concepts are not adequately described, and are assumed to be the same concept, leading to a merging of disparate concepts. Ambiguously recorded temperature datasets of the same feature of interest could be naively combined without correction to create an incorrect historical view of the temperature of the feature of interest in question.

developed in the context of health information is that of two-level information modelling (Beale, 2002).

Two-level modelling is a multi-level modelling technique that separates the standardisation process into distinctive levels with supporting user-friendly tools. Most information systems are based on a single level information model. Health informaticians had long recognised the issues associated with systems based on a singular information model and for several years investigated multi-level solutions (Ingram et al., 1995) (Grimson et al., 1996, 1998) (Heard and Beale, 1996) (Kalra, 1997) before ultimately developing the two-level modelling approach (Beale, 2002). In two-level modelling a second (knowledge) level model is introduced that is directly defined by non-technical practitioners such as clinicians. This second level allows domain experts to contribute their knowledge and experience directly to the information system's information definition and is in contrast to traditional techniques such as object-oriented analysis and design which is largely driven by technical experts. Two-level modelling will be described in detail later in chapter 3.

1.3 Hypothesis

Two-level modelling techniques developed by health informaticians for use in clinical settings have been shown to be very powerful in enabling semantic interoperability between heterogeneous clinical information systems. Clinical information systems and large geospatial information systems have comparable issues with the complexities of modelling and combining information within their domains. Therefore, translating and adapting two-level modelling approaches within geospatial information systems could lead to the same enhancements in data quality and semantic interoperability within geospatial systems as observed in e-health systems.

Once translated and deployed, two-level modelling could allow diverse Earth System Science domain experts to be the primary drivers of geo observational sensor based digital artefacts; this in turn allows a rich distributed, evolving and interoperable knowledge cosmos to exist beyond what is possible with deployed state-of-the-art spatial data infrastructures and geo data portals.

1.4 Research Question

Can two-level modelling be translated from the health domain to the geo-spatial domain and applied to technologically constrained observing scenarios to improve semantic interoperability within and between spatial data infrastructures beyond what is possible with current state-of-the-art approaches?

1.5 Research Objectives

This research work seeks to develop novel approaches to aid geographical/environmental data collection and usage activities through interoperability. By enabling larger datasets to be combined using highly flexible interoperability mechanisms, this work seeks to enable the automatic synthesis/discovery of new knowledge from geo-sensor networks.

Additionally, this work is focused on developing approaches that can be used in heterogeneous geo-sensor networks consisting of constrained sensor nodes. It is the expressed goal of the work to provide mechanisms to annotate data as soon as possible by pushing the data processing to the edge of sensor networks. This annotation will consist of adding context, lineage, semantics etc. close to the point of data capture subject to bandwidth and other resource constraints. Pushing data quality right to the point of data capture can reduce (unintentional) conflation as the data move up the data value chain.

The specific objectives within the wider context of the research work described in this thesis are listed and enumerated below. The research objectives are further mapped to

technical approaches defined to meet the research objectives later in section 1.8 and Figure 1.2.

1.5.1 Objective 1

Identify the technical tasks required to translate the two-level modelling methodology from the health domain to the geo-spatial and Earth System Science domain

1.5.2 Objective 2

Define a technical architecture to underpin a two-level model enabled spatial data infrastructure.

1.5.3 Objective 3

Investigate to what extent two-level modelling can act as a solution for geo-observational sensor systems semantic interoperability.

1.5.4 Objective 4

Develop and make publicly available a library of geo-archetypes that can act as a proof-of-concept of two-level geospatial modelling and thus enable further exploration and adoption of two-level modelling within the geo-spatial community.

1.5.5 Objective 5

Investigate mechanisms to enable a two-level modelling approach to be applied to the *edge* and beyond of technological constrained in situ geo-observational sensor systems.

1.6 Methodology and Project History

This research was conducted during the period 2014 – 2019 in a part-time mode of study. The research output forms the basis for a new research agenda within the School of Electrical & Electronic Engineering at TU Dublin.

This work was conducted using three research approaches, outlined below. Firstly, a theoretical approach was used to examine the state-of-the-art in knowledge representation

within the geomatics domain (reported in Chapter 2 of this thesis). Knowledge representation techniques were examined, and appropriateness assessed for the problem domain (reported in Chapter 3 of this thesis).

1.6.1 Research Design

The research design approach taken within this work was informed by a "design science" paradigm, often used in Information Systems research. The design-science paradigm focuses on producing useful & innovative artefacts (Henver et al., 2004). Henver et al., (2004) have developed a useful conceptual framework to aid the understanding, execution, and evaluation of research similar to that of this work.

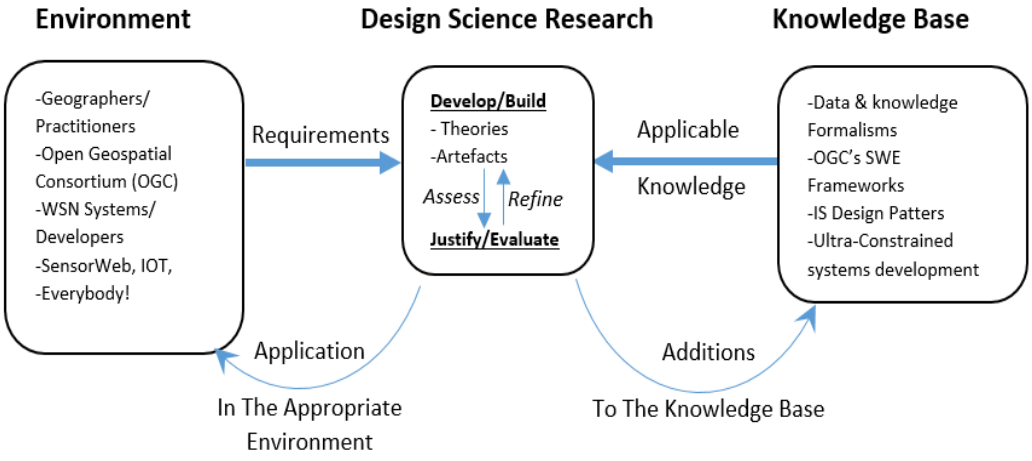


Figure 1.1 Design Science Paradigm, and Framework adapted from Henver et al. (2004).

Using the design science approach, the main research environment was identified and represented within a design science framework (Figure 1.1). The environment contains a broad range of stakeholders, listed under "environment" in Figure 1.1. Essentially all citizens are potential stakeholders, as given modern data mobile technology all citizens may have at some point a desire to generate and or share geo-information. However, this work is primarily aimed at providing solutions to non-technical geo-spatial domain experts and systems developer. The defined environment was used to refine a set of system requirements which are detailed in chapter 5.

Using the design science approach, the knowledge base was also defined (Figure 1.1). The knowledge base identifies the current communities of knowledge that informed the basis of this research work. The knowledge base also identified the communities or stakeholders that will benefit from the work carried out through additions to the knowledge base.

Using the environment and knowledge base, a set of theories were developed. A design process was then initiated, which was a cycle of "build and evaluate". The building refers to the building of design artefacts based on developed theories (detailed in Chapters 4, 5 and 6). Design artefacts are constructs, models, methods, and software instantiations (Henver et al., 2004). Design as an artefact was a primary method used within this work. The resulting artefacts were evaluated through the development of a real-world case study/action research (observational and analytical) approach.

The knowledge modelling methods and supporting infrastructure were applied to two key application domains. The focus here has been on applying the methods developed within ocean observing scenarios. Focusing on ocean observing scenarios was due to the background of the author in previously working within the marine monitoring area and due to the advanced nature of current ocean data portals, which allows the work to be framed within an area of Earth system science with advanced spatial data infrastructures in place. Also, the area of marine monitoring has been chosen by the EU under the INSPIRE framework as one of three areas to run standardisation pilots (discussed in more detailed in Chapter 2, section 2.5.2). However, the application domain could have just as easily have been applied to other areas such as land or atmospheric sensing. As such a basic example of the approaches developed applied to atmospheric sensing was also performed. Also, over the timeframe of this work the Internet of Things paradigm has gathered pace and matured significantly. In order to demonstrate the wide applicability

of the approach and its relevance to the emerging Internet of Things, the approach has also been considered in the context of the Internet of Things domain as applied to atmospheric sensing, specifically what is termed in the thesis the *Smart Cities scenario* (air quality sensing) in Chapter 6. The Smart Cities scenario is quite limited in scope. The primary goal of this study was to show the wider applicability of the technique to technologically constrained observing platforms. This scenario demonstrates the flexibility of the data modelling approach by applying the technique to an air quality monitoring use-case for a smart city project. The key element of this work was to show how emerging standards within the Internet of Things (IoT) field can benefit from the approach defined here and as constrained observing platforms move towards using more and more IoT innovations how this work can remain relevant in an age of IoT enabled remote in situ sensing systems (objective 5).

Having focused on the use of two-level models "at the edge" there is also a need to consider two level models in the context of aggregation and sharing of information across the scientific community. The ocean observing scenario seeks to demonstrate how the techniques developed as part of this work can improve the interoperability of data generated through ocean observing deployments. In this scenario the focus was on harmonising existing historical ocean observation datasets and applying a hindcasting technique to show the benefits of the approach for enhancing ocean chlorophyll-a estimation models within the Southern North Sea area.

1.6.2 Model & Experimental

The initial exploratory phase identified key requirements for an ideal interoperable geo-observational system model. This model formed the basis for a number of experimental simulations to be run to understand the ramifications on data management within the various sensor architectures that make up geo-sensor networks. A final "ideal" model was

derived from the results of simulations (see technical approaches 1.1, 1.2 and 1.3, Figure 1.2 below).

1.6.3 Build

The building of a proof-of-concept system was undertaken. A novel design incorporated the outcomes from model simulations and a systematic review of current state-of-the-art systems into a sensor-based architecture to show how the hypothesis advanced from the research question, functions in terms of an overall system. The proof of concept system aimed to show how real-time data streams can be successfully integrated with historical datasets and to support the efficient finding and binding of disparate datasets (see technical approaches 1,2,3 & 4, and evaluations 1 & 2, Figure 1.2 below).

1.6.4 Formal Methods

The ultimate outcome of research objectives (1) & (2) were a set of specifications to realise a novel approach to achieve semantic interoperability within geo-observational sensor systems. The correctness and quality of the overall solution was evaluated using a proof of concept build and its application to real-world scenarios using a use-case based evaluation method (see technical approaches 3 & 4, and evaluations 3 & 4, Figure 1.2 below). The outcome of these evaluations was analysed using a comparative analysis method to assess overall approach solution on meeting the research objectives.

1.7 Thesis Outline and Reader Guidance

A research canvas giving a broad overview of the research work is provided in Figure 1.2 below to assist the reader. This thesis provides the reader with a broad synthesis of the relevant literature from several disciplines - required to answer the research question.

It is necessary to deal with each of these in turn within the thesis as they have been instrumental in performing the requirements analysis for resultant design solutions and

developing a robust translation methodology for the adoption of two level modelling (see technical approaches 1 in Figure 1.2 below). Consequently, the thesis may have broad interest from health informaticians to Earth system scientists, Smart City architects, standardisation consortiums, embedded system engineers and others besides.

This section provides guidance for the individual reader so they may decide how best to navigate the content. Depending on the background of the reader, they may wish to focus on certain aspects of the work while passing over other sections. Figures 1.2, 1.3 and Figure 1.4 below provide a visual overview of the thesis to aid navigation.

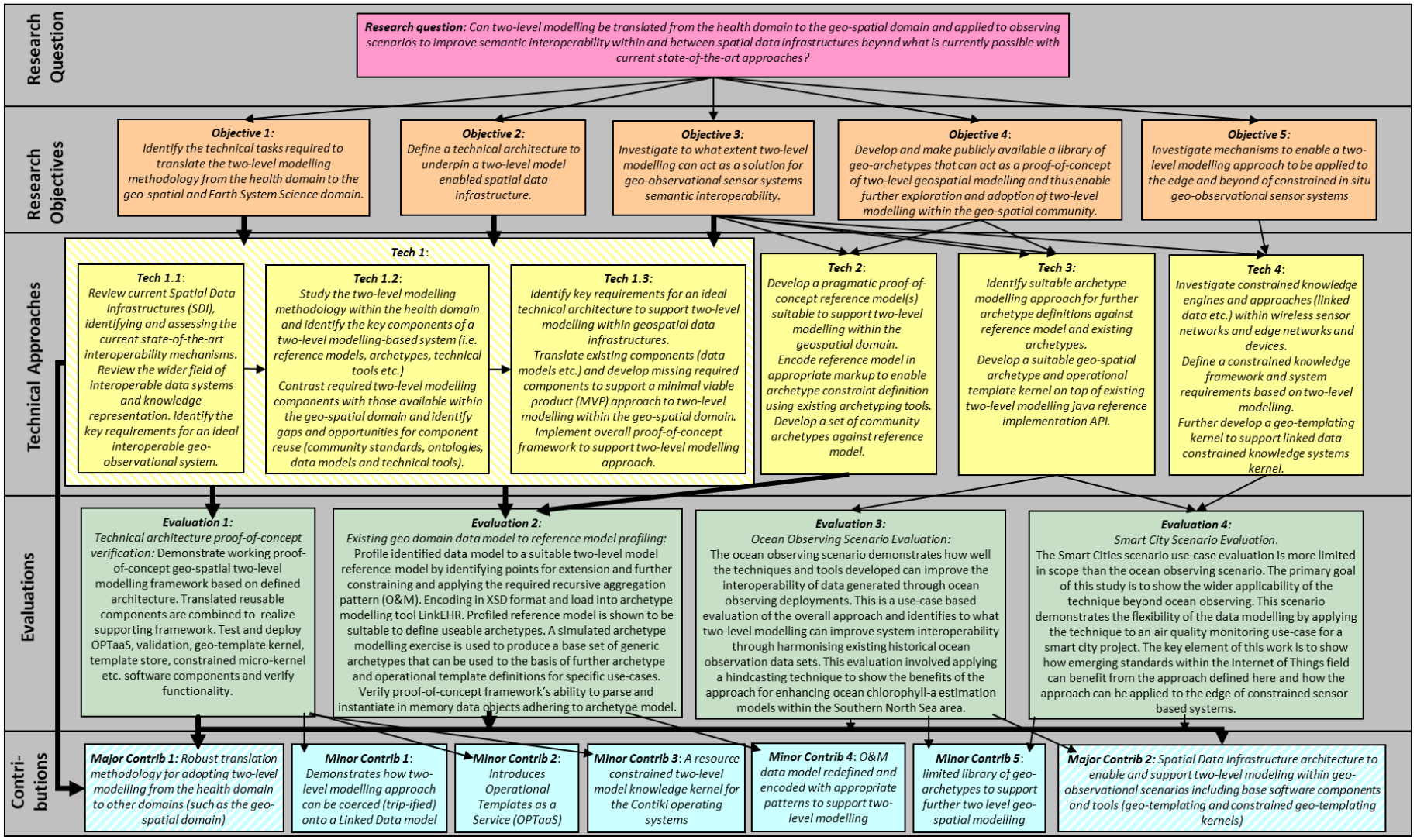


Figure 1.2 Research Canvas, Canvas elements are overlaid and mapped to Thesis chapters in Figures 1.3 and 1.4 below

The remainder of this thesis is organised as follows:

Chapter 2 This chapter gives a broad background and further context to the motivations for this work by providing a brief literature review of the history of Geography and Geographical Information Systems (GIS) before reviewing current state-of-the-art of Spatial Data Infrastructures (SDI) and the current state-of-the-art of geo-observational sensor-based systems and the application of semantics on constrained computing platforms. Chapter 2 and chapter 4 contrasts the required two-level modelling components with those that are already available within the geo-spatial domain and identifies gaps and opportunities for component reuse (community standards, ontologies, data models and technical tools).

This review contributes to research objectives (1) and (2). The reader is also given a more in depth introduction to several concepts and ideas mentioned earlier in this chapter such as the *Digital Earth* and domains such as *Earth System Science;* these concepts are necessary to understand the complexity of the problem domain being investigated.

Chapter 3 provides the reader with a review of semantic interoperability and formal representation of knowledge, culminating in an introduction and overview of *two-level modelling*. In this chapter a review of the wider field of interoperable data systems and knowledge representation is performed. Key requirements for an ideal interoperable geo-observational system are identified. Also, in this chapter a study of the two-level modelling methodology within the health domain is presented. Understanding two-level modelling and the use of archetypes are key to this work and it is recommended that the un-familiar reader dedicates some time to this section of the thesis. The review and requirements analysis presented in this chapter also contributes to research objectives (1) & (2).

Chapter 4 describes the approach developed to translate two-level modelling to the geo-spatial domain, which is key to answering the overall research question (defined above). The key components of a two-level modelling-based system identified in earlier chapters are contrasted with that available within the geo-spatial domain. Gaps and opportunities for component reuse (i.e. community standards, ontologies, data models and technical tools). An ideal technical architecture to support two-level modelling within geospatial data infrastructures is defined. Key system requirements are presented and pre-existing two-level modelling system components (data models etc.) are identified. The missing required components to support a minimal viable product (MVP) approach to two-level modelling within the geo-spatial domain are identified. Subsequently an overarching proof-of-concept framework to support two-level modelling approach is defined.

Chapter 5 describes the definition, design and implementation of a constrained two-level knowledge-based framework, necessary to address research objectives (1) (2) & (3). A pragmatic proof-of-concept reference model(s) developed to validate the two-level modelling translation methodology within the geospatial domain is described and the process of developing the first set of community archetypes against the reference model defined in chapter 4 is presented. The reader should note that chapter 5 describes a large array of implementation technology. The details are included not to distract from the primary objectives of the research work but are provided to highlight the complexity in adoption of two-level modelling approaches and to give evidence to the veracity of the validation of the methods described in chapter 4.

Chapter 5 also describes the work done on developing a constrained knowledge engine and the approaches employed (linked data etc.) to achieve a two-level modelling approach within wireless sensor networks and edge networks and devices. A constrained

knowledge framework and system requirements based on two-level modelling are defined. The development of the software kernel required to support linked data constrained knowledge systems kernel is presented. Chapter 5 describes the work performed to achieve research objectives (4) & (5).

Chapter 6 describes the application of the overall translation technique and supporting infrastructures in two specific evaluation scenarios (ocean observing and smart cities), including an analysis and synthesis of the approach. Chapter 6 also addresses the main research question and presents the ultimate results of testing the research hypothesis.

Chapter 7 articulates the final conclusions and implications of this research and sets of future directions of the research work including a summation of the contributions of this work and to what extent the research objectives 1-5 have been met.

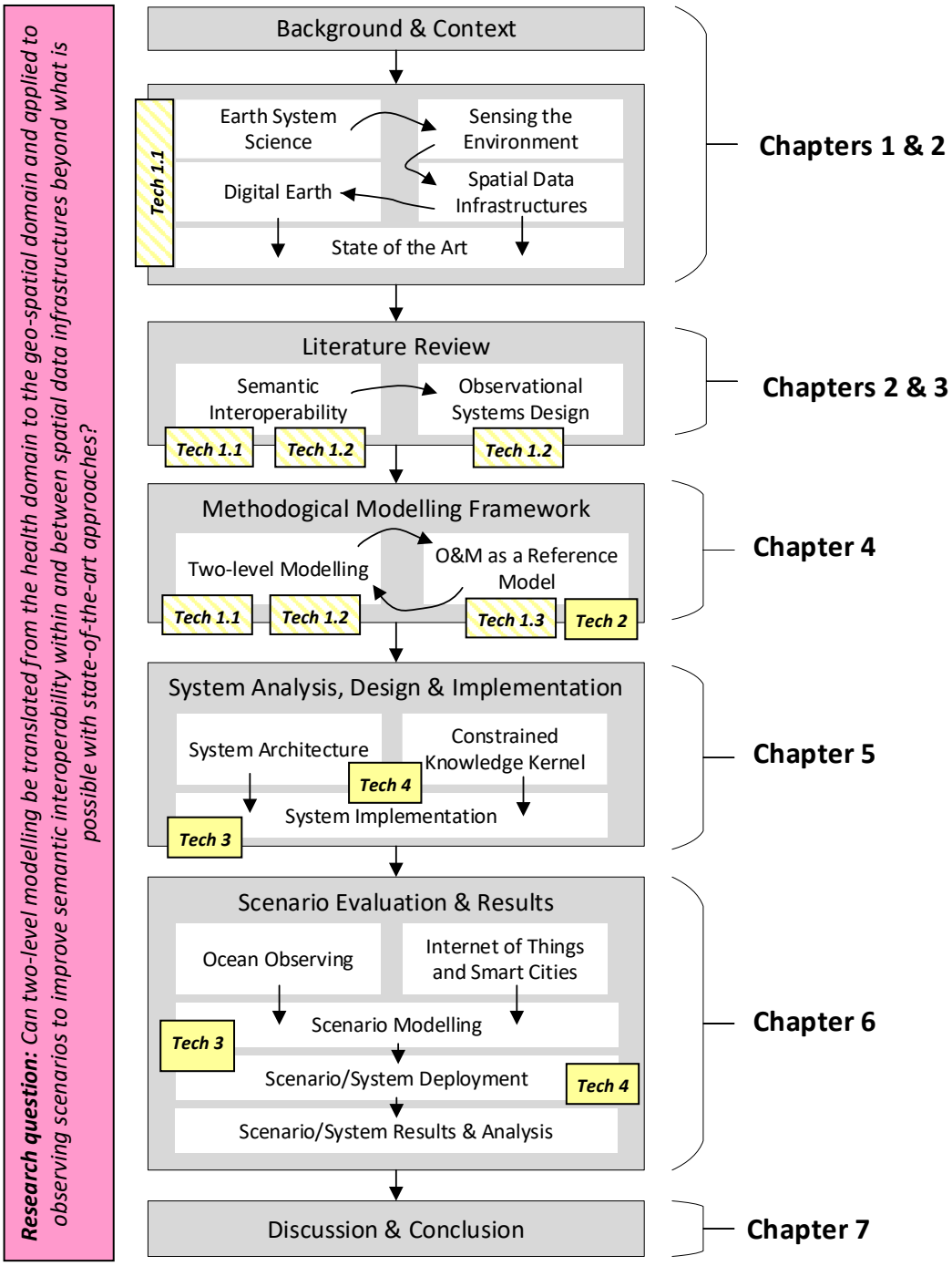


Figure 1.3 Thesis outline, showing technical approaches evolution to address specific research objectives. Technical approaches (Tech 1,2,3,4) mapped from research canvas (Figure 1.2)

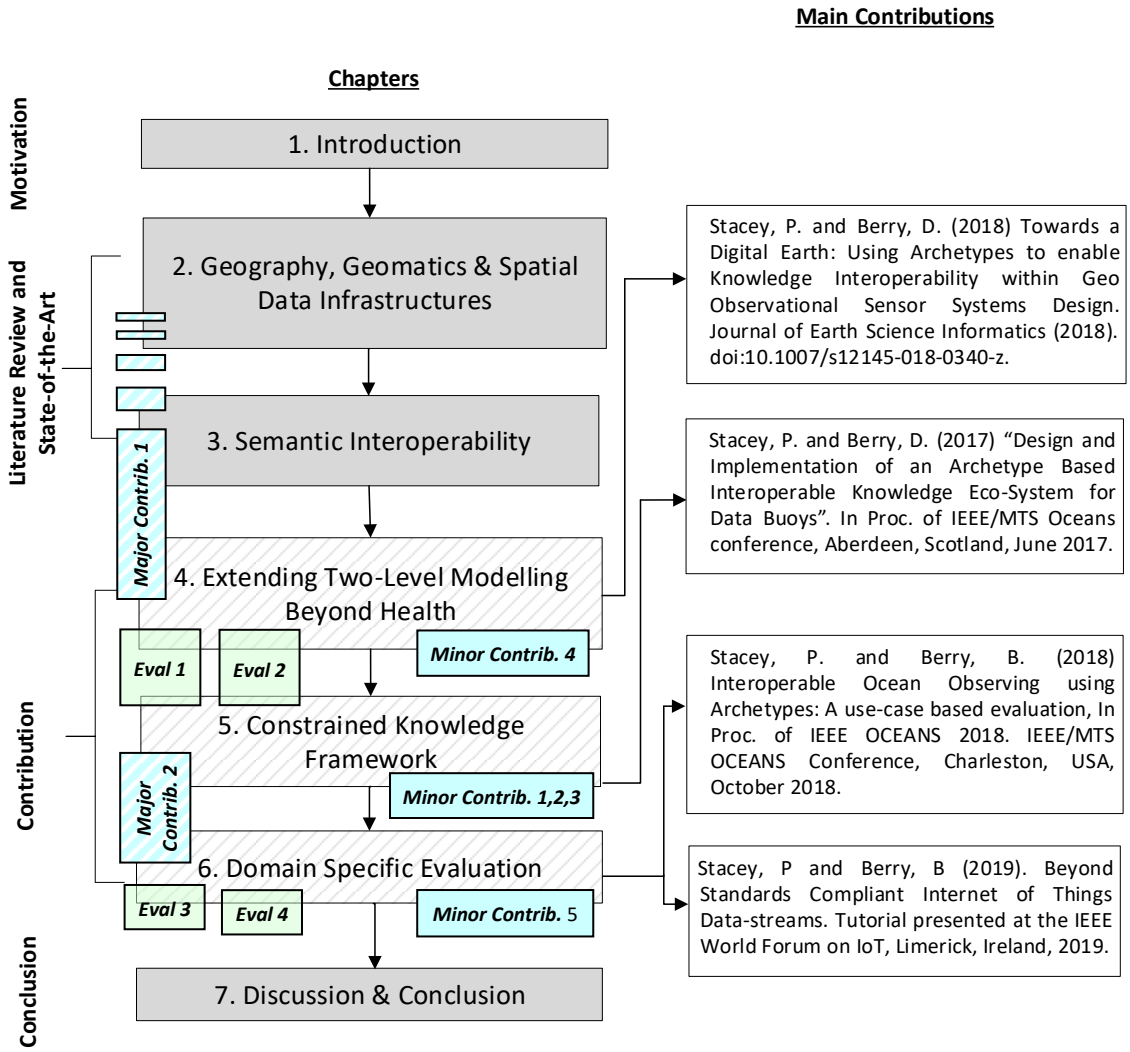


Figure 1.4 Thesis chapters, with evaluations & key contributions mapping. Evaluations (Eval 1,2,3,4) and Contributions (major/minor 1,2 etc.) mapped from the research canvas Figure 1.2

(NOTE: additional communications are listed at the end of this thesis)

Chapter 2

"Clearly, the Digital Earth will not happen overnight"
(Al Gore, 1998)

2. GEOGRAPHY, GEOMATICS & SPATIAL DATA INFRASTRUCTURES

Chapter Overview: Chapter 1 introduced the five main objectives of this work (section 1.5). To meet research objectives 1 and 2 (section 1.5.1 and 1.5.2), a comprehensive review of the main facets of geographic data and information are required i.e. the collection, distribution, storage, analysis, processing and presentation of geographic data or geographic information. ISO/TC 211 (2011) defines this collection of facets as the geomatics discipline.

This chapter provides the reader not only with a review of geomatics, but also a historical perspective of geomatics in relation to geography, including technologies pertaining to geomatics, such as remote sensor systems and geographic information management systems such as spatial data infrastructures (SDIs). As will be seen later (chapter 4) this review has informed the approach used in this work to translate two-level modelling for use within the geo-spatial domain i.e. the technical approach used to answer the research question (see section 1.2 and 1.3).

Firstly, the reader is presented with an overview of the intertwined evolution of geography and geographical information systems. It is important to understand this evolution, as by examining GIS' contentious association with particular branches of geography, one can get a good understanding of the challenges that exist when developing information systems for complex multi-disciplinary environments (such as

Earth system science). It is important to gain this perspective in the context of this research to ensure a comprehensive review of the inherent complexities within geomatics, while also assessing whether two-level modelling is appropriate for solving issues of interoperability in the geo-spatial domain and ultimately within Earth systems science and a Digital Earth.

Later in this chapter an overview of geo-observational sensor platform technologies is presented and discussed. The Internet of Things is introduced, and its relevance to the area of geomatics for providing solutions for the collection and distribution of sensor based observational data to SDIs is reviewed and discussed. Limitations of current technologies and techniques are also discussed.

2.1 Geography & GIS

Geography derives from the Greek γεωγραφία – geographia (Douglas, 2017), meaning to "describe or write about the Earth". Bartholomaus Keckerman, a theologian who lived from 1572 to 1609, can be credited as the founder of modern geography (Bonnett, 2008). Kerkerman distinguished between graphica generalis (which takes a global view of Earth) and geographica specialis, which focuses on particular regions (Livingstone, 1988). More recently, we can identify many different branches of geography (Bonnett, 2008):

- Physical Geography
- Human Geography
- Integrated Geography
- *Geomatics*
  - Spatial Analysis
  - Cartography
  - *Geographical Information Systems (GIS)*
  - *Remote Sensing*
  - Global Positioning System (GPS)
- Regional Geography

This research work lies within the branch of *geomatics*, and more specifically between the sub-branches of *geographical information systems* and *remote sensing*, discussed next.

2.1.1 GIS, a Geographer's Best Friend?

In his 1960s book 'Applied Geography', Dudley Stamp presented many applications of Geography in the real world (Stamp, 1960). Many of the ideas presented showed how Geography could be used across other disciplines. However, without the ability to readily share geographical information and knowledge, many of Stamp's applications were not realised or even possible until recently. The lack of geographical information systems at the time meant that useful geographical knowledge remained in the realm of geographers.

Since the early 1990s, the discipline referred to as geographical information science (GIScience) has sought solutions to the adequate representation of the uncertainty that exists within geographical information (Goodchild, 2010) (Goodchild, 2020). GIScience has also sought solutions to effectively share Geographical information and knowledge in a computable way, thus realising many of Stamp's original ideas. The activities of Geographical Information Scientists in the early 1990s enabled the development of the first modern day geographical information systems. Clarke (1997) defined Geographical Information Science as *"the discipline that uses geographic information systems as tools to understand the world"*. Geographical information systems are practical tools, whereas GIScience addresses the fundamental question of how data, space and the digital world relate (Geographical Science Committee, 2005).

Geo-Information Scientists have had - at times - a contentious relationship with another group of professionals in the scientific community that they most closely operate alongside, traditionally referred to as Geographers. In the early 1990s, GIS became the focal point of an academic debate about the merits of such systems. In his 1990 editorial

"GKS" (Taylor, 1990), Peter Taylor began what some refer to as the "GIS wars" (Schuurman 2000). Taylor suggested that while GIS had certain merits in managing and handling geographical information, GIS lacked the ability to generate knowledge through meaningful analysis. Taylor's editorial gave a voice to a growing discontent that had been brewing amongst what are referred to as "human geographers".

In 1991, M. F. Goodchild (a prominent GIS researcher) published a counter argument (Goodchild, 1991). In deference to Taylor's criticisms, Goodchild acknowledged the inadequacies of GIS, while also making the point that GIS was intended to be used alongside experts in the field; a tool to enhance and aid knowledge construction. This reassuring declaration from Goodchild - that GIS was only to be used by knowledge experts - does not hold true today, for several reasons.

The argument at the time was that knowledge could not be generated, and therefore GIS was only based on facts. To put it simply, it was believed that GIS tools should only be used by geographical experts, within a specialist sub domain. Only experienced practitioners would have the ability to interpret and analyse the data & information captured within GIS responsibly.

However, the firm embedding of GIS within an overall ESS Information Systems framework invariably means that geographic information will be shared with non-geographic experts. In fact, this sharing of information beyond the realm of geography would be a core goal of modern-day information system frameworks. It is therefore incumbent on all ESS domain-specific Information Scientists to ensure that information systems adequately capture the knowledge and intricacies of the domain information, representing it in a sharable, interoperable, and ideally reusable way.

It is important to consider and understand the origins of discontent amongst geographers towards the increasing digitisation of Geographical information. Gaining an

understanding of why there was a backlash against GIS from (typically non-technical) domain experts is an essential step in developing any environmental or geographical knowledge system. This research is primarily aimed at providing non-technical domain experts with tools to enable them to become the main drivers of how geographical information is defined. Therefore, understanding the end user requirements is key.

The book 'Ground Truth' (Pickles, 1995) provides a comprehensive record of the discourse at the time. Ground Truth has been attributed with causing a major shift in how geographic data should be modelled and represented. A publication by Goodchild, captured the mood ten years on from the publication of Ground Truth (Goodchild, 2006) and is a recommended divergence for the interested reader. Since then, the discourse continues (Thatcher et al., 2016) (Singleton and Arribas-Bel, 2019), albeit in a more unified way.

As is evident today, there is pervasive access to geographical and environmental data through the Internet, Web and mobile applications. In fact, non-experts (i.e. not geographers) now make important decisions based on geographical and environmental data every day. It could be argued that the "middle-men" (domain experts) have been to some degree cut out of the equation. It could also be argued that (much like news organisations and the advent of social media and *fake news*) we have entered a dangerous period in our appetite to disseminate geographical and environmental data. Without expert analysis, interpretation, and context these data could be described as incomplete, and unsuitable for leading to meaningful decision making. M.F. Goodchild noted in 2006, regarding geographic data:

> *"the average researcher, and increasingly the average citizen, clearly needs to know far more about the context, lineage, and meaning of data."* (Goodchild, 2006)

Also, with the advent of *Big Data* within the spatial data domain, the need for adopting high quality data science approaches within geography and geographic analysis is becoming critical (Singleton and Arribas-Bel, 2019).

Today, the wide applicability of traditional Geography can be observed in its intertwined relationship with the integrative super-discipline of Earth System Science. This expansionism of Geography as a discipline, when many sciences have become reductionist (Pitman, 2005), presents additional challenges to the ability of the domain's information systems to share knowledge to a wider super-discipline such as ESS. Again, this adds weight to the argument for the need to employ robust data science approaches to achieve semantically interoperable geographic data and information (see Chapter 1, section 1.5.3 objective 3), and avoid mis-interpretation, representation and conflation of data by non-experts.

Chapter 3 will return to the more philosophical complexities of information representation and semantic interoperability. Here the reader is presented with GIS from a systems technical architectural perspective. A review of GIS architectures, as will be seen later, has informed the technical architecture defined in this work (objective 2) and is part of the technical approaches used within this work to address the main research objectives (Chapter 1, Figure 1.2, tech 1.1).

2.2 GIS, More than Maps

Typically, mapping services are the primary focus for spatial data. Indeed, much of the GIS technology available today is optimised for mapping services and the rendering of geospatial layers on top of base mapping technologies. For example, GeoServer is OGC compliant for several Web mapping standards (GeoServer, 2019). However, there is not as much support for data related services relating to monitoring of physical phenomena. Despite there being many useful standards in this area. It is therefore reasonable to assert

that GIS has typically been map focused. This has driven the focus on achieving interoperability of mapping services, raster/vector data types and fusing of data layers. However, there are now many examples of ongoing initiatives to increase the number tools that implement data related standards through proof-of-concept implementations and library tools (Brodeur et al., 2019). A GIS product is only as good as its raw materials, and in the same way a GIS is only as good as the geospatial information that it manages and presents to users. So, the quality of the information gathering process that underpins GIS is of critical importance.

The 52° North open-source initiative tests implementations of open-source standards (Kraak et al., 2005). Several reference implementations have been released by 52° North, which includes the OGC SWE initiative. 52° North's focus since its foundation has been on the interoperable integration of geosensor data, specifically on standardisation of interfaces and data encodings for data from environmental sensing activities such as flood gauges, air pollution, space and air borne Earth imaging devices.

Data captured from geosensor deployments and traditional spatial data are not mutually exclusive, but complementary. However, the adoption of standards related to mapping services has had broader uptake compared to geosensor data that are related to Earth observation. It must also be noted that map making, historically has had more value to wider society, and its origins can be traced back way beyond that of Earth observation.

The balance of importance placed on mapping services over Earth observations within GIS systems is shifting and will continue to shift (Goldberg, 2014). This is happening for many reasons and is only set to accelerate with the growing pressure on all of society to become more knowledgeable regarding climate breakdown and the changes that are taking place within the natural processes that surround us (Fraisl et al., 2020). Increased

Earth observation is also driving this shift as more data products become available for consumption.

2.2.1 Earth Observation

As noted in Chapter 1, in situ remote sensor deployments and satellite-based Earth Observation (EO) systems monitoring environmental phenomena are two important sources of computable data for Earth Scientists (Hart and Martinez 2006). The main research question (Chapter 1, section 1.4) focuses on the application of two-level modelling approaches right to the point of capture[7] on technologically constrained sensor systems (i.e. in situ remote sensor platforms). Here, Earth observation is defined in more detail and a differentiation between the different Earth observing systems deployed in space and on land is provided.

The term Earth observation (EO) refers to any form of observations of the Earth. Although in certain communities EO often refers to remote sensing exclusively (i.e. satellite based sensing). In general, EO encompasses remote and in situ, including air borne sensing of the Earth's processes. The Group on Earth Observations (GEO) which includes over 100 member countries uses the term EO in the broader sense. Throughout this thesis, this broader definition of EO is also adopted.

The activity of gathering Earth observational data using remote sensing techniques can be traced back to World War 1 (Eyres, 2017). Using ordinary cameras mounted onto reconnaissance aircraft, remote observations of the position and strength of enemy forces were captured. This was the precursor to modern Earth Observation (EO).

In the digital age, vast amounts of Earth observational data have been collected and persisted in digital format. These datasets have been invaluable in helping humans study

[7] The reasons for this have been discussed briefly previously in chapter 1 and relate to conflation etc., this is dealt with in more detail in the next chapter.

Earth's processes. As our understanding of the complex interworking of Earth's many processes through an Earth Systems Science approach has increased, the benefits of combining Earth observational datasets are becoming clearer. Arguably, the ability to combine these disparate datasets is now essential in the context of a human influenced geological epoch (as discussed in Chapter 1).

Several techniques to capture Earth observations have been reported in the literature. Each technique provides a different perspective on the Earth and the goal of any truly comprehensive digital Earth system should be to ultimately harmonise and integrate their observations. Hence, they are considered next.

2.2.1.1 Satellite and Air Borne Remote Sensing

The Copernicus programme is Europe's eyes on the environment, bringing together data collected in space, on the ground, in the sea and in the air for the benefit of Europe's environment and its citizens. Copernicus includes space services and in situ components. The space component comprises 80% of the total Copernicus budget (Showstack, 2014).

In 2014 the European Space Agency began to launch its fleet of Sentinel satellites. Satellite data from the Copernicus Sentinels is made available on a full, free and open basis and serves as one of the main inputs into the production of the six thematic Copernicus Services: Land Monitoring, Marine Environment Monitoring, Atmosphere Monitoring; Climate Change, Emergency Management and Security. Specifically, the Sentinel 2 satellites focus on land, and Sentinel 3 satellites focus on marine (Copernicus, 2017). Data products are created and may be accessed from the Copernicus open data hub[8].

Satellites for environmental monitoring are normally equipped with a range of sensing equipment. Sentinel 2 satellites are somewhat limited compared to other environmental

[8] https://scihub.copernicus.eu/

monitoring satellites due to their focus on land coverage. Specifically, they have been deployed to monitor Polar Regions. Both deployed sentinel satellites are fitted mainly with a MultiSpectral Instrument (MSI), which has a 290km Field of View (FOV)[9].

Sentinel 3 satellites are dedicated to ocean monitoring and contain the following instrument payload[10]:

- An Ocean and Land Colour Instrument (OLCI).
- Sea and Land Surface Temperature Radiometer instrument.
- A dual frequency SAR altimeter.
- A Microwave Radiometer.

The sentinel 3 on board instruments provide accurate real-time ocean observing capabilities to monitor several ocean-based geographic features. For example, the OLCI equipment can detect harmful algae blooms and is used to supplement existing water quality monitoring processes (ocean observing for the detection and prediction of harmful algae blooms is discussed in more detail in Chapter 6 as part of the validation approach for this work).

Satellite-based Earth observations have notable limitations. For example, observational ability may be diminished with cloud cover. Also, at the level of the space component, satellite sensors need to be calibrated, and their data products validated, using independent on the ground or in situ data sources (known as ground truthing) meeting specific requirements. The Copernicus services rely on the availability of a wide variety of in situ data. These data are used both for production and validation (Copernicus, 2017), but also to augment coverage data and provide higher resolution of datasets on Earth and reduce the need for interpolation (Figure 2.1).

[9] https://sentinel.esa.int/web/sentinel/missions/sentinel-2/satellite-description
[10] https://sentinel.esa.int/web/sentinel/missions/sentinel-3/instrument-payload

Figure 2.1 September 2020, Dublin based company TechWorks Marine *tweets* the deployment of in situ "ground truthing" marine observation systems to validate satellite-based Earth Observation[11].

Large-scale satellite deployments like the Copernicus Sentinel missions are expensive operations. Today, new satellite and air borne remote sensing platforms are being deployed by both public and private organisations. Microsatellites and unmanned aerial vehicles (UAVs) are now opening new possibilities for augmenting already existing monitoring programs. Smaller and more cost-effective deployments such as *nano* and *pico* (cubesats) satellites are now using off the shelf electronic components to provide low cost specific Earth Observations (Heidt et al., 2000).

Cubesats can be deployed from platforms such as the International Space Station (ISS) (Figure 2.2). However, a new generation of space business-based start-ups are now providing design and launch services, further increasing the amount of heterogeneous monitoring activities and resulting datasets. As of August 2020, the United Nations Office for Outer Space Affairs (UNOOSA) returns 8615 registered objects launched into outer space[12].

[11] https://techworks.ie/en/
[12] http://www.unoosa.org/oosa/osoindex/search-ng.jspx

Figure 2.2 Astronaut Serena Maria Auñón-Chancellor talks live via live link from the International Space Station (ISS) with attendees of the IEEE Oceans 2018 conference and demonstrates how cubesats are launched from the ISS launch hatch. *Photo credit: author*.

2.2.1.2 In situ Sensing

In situ Earth observation/sensing typically refers to physical environmental monitoring systems being deployed on the ground, air, in or on water. In situ sensing may also be carried out by individuals taking samples by hand, with later processing of samples in a laboratory environment.

New and novel sources of in situ data, such as imagery gathered by drones and information collected by crowds of volunteer contributors (Goodchild, 2007) or citizen scientists (crowdsourcing) also fall under the in situ umbrella. This work focuses on in situ sensing systems deployed on technologically constrained observational platforms on land or sea, and as such in situ geo observation sensor-based systems are dealt with in more detail later in this chapter (section 2.4). Before that, a review of some of the techniques used to represent environmental data and geographical data collected by earth observational activities is presented.

2.3 Environmental and Geographical Data

This section provides a brief overview of some of the pertinent aspects of environmental data before these datasets are considered within the context of spatial data infrastructures later in section 2.5. This review sets out the state of the art in environmental data formats and representation. Also, this section presents to the reader some of the complexity associated with environmental data representation and the limitations of some common formats. For example, one of the most common data formats used to publish scientific datasets is the netCDF format (Rew and Davis, 1990). netCDF is pervasive in environmental data products (used for example in disseminating Copernicus EO data products). However, netCDF only acts as a container format but does not define data at the more fine-grained syntactic level, a minimum requirement for data interoperability (discussed in more detail in Chapter 3). Thus when attempting to combine netCDF based datasets from heterogeneous Earth observing systems typically the contained data are not harmonised into a singular standardised format, or where a data standard is employed there are many inconsistencies which make data fusion difficult or impossible.

Environmental data are normally collected (through observation and measurement) or inferred through statistical approaches to represent the state-of-the environment. However, determination of environmental state normally requires the grouping of several data, these groupings are referred to as environmental indices (Ott, 1978). A good example of environmental indices are quality indexes, such as *air quality index* or *water quality index*.

Geographical data can be divided into geometric data or attribute data. Geometric data are geometry data made up of points, lines or area. Attribute data can be sub-divided into qualitative (for example specifying the type of object) or quantitative (comprising ordinals, ratios or intervals). Geographical data are largely captured as either raster of

vector file formats. However, with the increased interest in geographical data for uses other than that of mapping (see section 2.2 above), image-based formats have been supplemented with many new formats that are more suitable to environmental data representation. For example, ArcGIS supports up to 36 file formats in addition to numerous raster formats and netCDF. In terms of attribute data, ArcGIS primarily uses netCDF. However, as was mentioned above and will be discussed in more detail below (section 2.5.2), netCDF has many limitations regarding interoperability of *attribute data*.

Uncertainty in geographic data occurs at different levels of abstraction. Position and temporal errors describe uncertainty in a metric sense. Completeness and consistency represent more abstract concepts that relate to coverage and reliability. These are more problematic to describe. So, how is uncertainty modelled in data as the data are transformed through different models of geographic space? As early as 1978, Sinton (1978) highlighted the problem of information structure as a barrier to analysis within GIS systems.

2.3.1 Geographic Objects

Geo-spatial knowledge representation predominantly takes an *object-field* conceptual view of geographical space (Cova and Goodchild, 2002). Here, objects are considered.

Taking a planetary scale view, the Earth is one object with a defined boundary. At the sub-level, Earth is made up of other objects with their own well-defined boundaries such as oceans and continents. *Tangible geographic objects* (for the most part) have broadly acceptable boundaries and properties (name, status) that do not generate much discourse. These geographic objects are referred to as discrete geographic objects. *Geographic phenomena* boundaries and properties on the other hand are more difficult to define. The boundaries of phenomena are normally continuous. For example, temperature as a naturally occurring phenomenon is continuous. It can also vary continuously in time and

space. As such, the boundary of phenomena such as temperature tends to be represented in a fuzzy arbitrary way due to the continuous variability of the phenomena.

Geographic objects tend to attract the interest of diverse stakeholders, all with different viewpoints. Getting all stakeholders to agree a consensus on the boundary and the properties of phenomena are represented or captured is difficult to achieve as each stakeholder will bring their own perspective and requirements to the discussion.

Next, boundary objects are considered in a little more detail. The point here is to illustrate the difficulty in achieving a shared world view of geographic data and the need for more inclusive, flexible, and complex frameworks to enable consensus-based shared world views of objects. For it is this inability that hampers interoperability efforts.

2.3.1.1 Boundary Objects

Star and Griesemer (1989) note that in general, scientific work is heterogeneous and requires cooperation. Due to divergent viewpoints, tensions exist while attempting to arrive at generalised findings. In their highly cited paper, Star and Griesmer examine this problem from a sociological perspective and articulate the importance of *boundary objects*.

Boundary objects are used to integrate scientific and technological classifications, while at the same time separating any opposing classifications. The boundary object construct was used by Harvey and Chrisman (1998) to examine the social negotiation that takes place within GIS systems development. They note that any time in which negotiations lead to the stabilisation of GIS technology, boundary objects have been at play.

Thus far, this chapter has reviewed the broad Earth observing systems in existence and some of the pertinent complexities inherent in the capture and representation of geographic observational data, due to the complex domain that these systems contribute

data and information to. The review now focuses on the Earth observing systems relevant to this work: in situ sensor based, technologically-constrained-systems (see research objectives, chapter 1, section 1.5).

2.4 Geo-Observational Sensor-based Systems

In 1999, Neil Gross predicted the exponential growth in planetary wide sensing:

> "*In the next century, planet Earth will don an electronic skin. It will use the Internet as a scaffold to support and transmit its sensations. This skin is already being stitched together."* (Neil Gross, 1999)

Over 20 years on from Gross' prediction we have now reached the point where billions of sensors are deployed globally for countless sensing applications. 2020 had been mooted for a long time as a watershed moment for the deployment of sensors and embedded devices to gather data about all aspects of our physical environment[13].

This section provides the reader with a review of current sensor based geo-observational systems available to monitor environmental phenomena. One key aspect of these systems is their limited computing power, which constrains their ability to process, store and communicate observational datasets. Limited computing power is typically a design choice due to the remoteness of their deployments, where access to reliable power sources is limited. Later in this section, these technologically constrained in situ observing platforms are discussed in the context of pervasive computing platforms such as IoT frameworks and *sensorWebs* (Delin and Jackson, 2001).

As in every aspect of this work, interoperability and standardisation are core considerations. Constrained systems present many challenges to interoperability, primarily due to their inability to handle the additional metadata requirements associated

[13] https://www.gartner.com/en/doc/463441-predicts-2020-as-iot-use-proliferates-so-do-signs-of-its-increasing-maturity-and-growing-pains

with standardisation and other interoperable solutions such as semantic mark-up (discussed in more detail in chapter 3).

2.4.1 Earth Observational Systems

In situ remote Earth observational systems are often built in isolation, and the data representations and associated documentation systems - where they exist - are often not adequately designed for secondary use, and higher order knowledge generation.

In recent times, many countries and jurisdictions have established their own remote EO systems and infrastructures (Westerbeeke et al., 2006) (GEO ,2016). NASA's Earth Observing System Clearing House (ECHO) (Pfister et al., 2001; ECHO, 2005) and the European Earth observation programme Copernicus (EO/Copernicus, 2016) are examples of how heterogeneous EO systems are being developed. In addition to these relatively monolithic satellite-based remote sensing systems, there is an even larger number of heterogeneous in situ remote sensing systems for capturing and publishing useful data. In general, there are a plethora of heterogeneous Earth-related monitoring systems with different access protocols, syntax, data types, identifiers, coding systems and metadata models. These deployed monitoring systems in their current state do not provide any clear mechanism for interoperability, even at the most fundamental data representation level. Examples of this fundamental problem are pervasive in scientific communities.

Interoperability mechanism deficiencies are particularly problematic in scenarios that require the consumption of data from many different heterogeneous sources. Solutions have been available for particular use cases for some time. For example, the Generic Earth Observation Metadata Standard (GEOMS) provides metadata definitions for a broad range of instrument types to allow the validation of satellite instruments from independent observations (Retscher et al., 2011). However, the flood of new systems providing Earth Observations in recent years has driven the need for standards to support the access and

processing of data from sensors from an even wider number of observing platforms (Khalsa, 2020)

As an example of issues related to interoperability outside of satellite instrument and data product validation, ongoing work at the Norwegian University of Science and Technology (NTNU) highlights the tangible consequences of the lack of interoperability. At NTNU work is ongoing to integrate data from various in situ sensor deployments to develop a common operational picture[14] (COP) to be able to better coordinate and manage operations in emergency situations such as an oil spill or combat operations (Osen et al., 2017). Their work has attempted to fuse ad-hoc data streams from all available relevant observing activities within an area of interest. For example, attempting to fuse water quality data from *ferry boxes* on ships passing within the area of an oil spill. NTNU's work has found that the key barrier to realising a COP is the lack of standardised geo-sensor-based data streams. Their solution is to develop their own integration services. However, this implementation is designed for a specific use-case, and consequently the aggregated data are not particularly suitable for secondary (re)use. These types of solutions are typical of the non-standardised bespoke approaches used on a per scenario basis within deployed systems.

Additional to the issue of standardised data streams is that of the quality of the sensed data within sensor data streams. There are many issues that can affect the quality of sensor data output: physical damage, lack of selectivity, non-linear performance, baseline drift, biofouling (Hayes et al., 2009) (O'Hare et al., 2009). As such, any observational data stream should have metadata describing the quality of the data from the sensor.

[14] "A common operational picture (COP) is a single identical display of relevant (operational) information shared by more than one Command" (DoD, 2007).

Standards do exist to enable the structured mark-up of quality data associated with the actual sensor such as the ISO standard 19157 which provides a standard data quality representation within geographic information (ISO, 2013).

2.4.1.1 Sensor Networks

A sensor network is a network of small sensing devices called motes or nodes which all collaborate on a common task (Verdone et al., 2010). In 1999 Estrin et al., published a highly cited paper highlighting the challenges for sensor networks heading to the 21st century (Estrin et al., 1999). One of the main challenges identified was that sensor network design could not rely on traditional wired network approaches as sensor networks would typically be data-centric and application specific. 10 years later, Nittel (2009) identified four key areas that presented research challenges for the advancement of Geo-Sensor Networks and Dynamic Environmental monitoring:

1) Programming of sensor networks is cumbersome and complex. User friendly API's are required to allow a user-friendly experience and facilitate experiments to be setup.
2) The problem of power consumption and supply. Novel algorithms need to be developed that detect and monitor and track environmental phenomena "in the network" instead of pulling the data to a centralised GIS system for data analysis.
3) To process both sensor network data as well as traditional geo-sensor data in real-time, a sensor data stream paradigm needs to be used for data management.
4) With continuously wider use of geo-sensor platforms, the problem of non-standardised sensor-data integration is of key importance to enable the so called "Sensor-Web" (sensorWebs are discussed in more detail below in section 2.4.4).

Nittel's four challenges listed above validate Estrin et al.'s hypothesis of data-centric and application specific sensor networks and the challenges they presented. These challenges

still exist today as sensor networks and single node deployments still tend to be application specific, which in turn means the data representation tends to be heterogenous (challenge 4 above).

2.4.1.2 Geo Sensor Networks

One of the primary functions of a GIS is to perform spatial data analysis. However, as the processing complexity of in situ sensing platforms increases it is possible that GIS systems will begin to disappear as a centralised analysis tool for raw sensor data (Nittel, 2009). This further compounds the need to have high quality data representation at the point of capture. Duckham (2008) proposed that sensor networks will ultimately become the GIS, bringing about ambient spatial intelligence.

Geospatial information is increasingly recognised as the common denominator between today's Web 2.0 dynamic social networking paradigm and that of the Web 4.0 (sensorWebs) (Carswell and Yin, 2012). A SensorWeb consists of a system of wireless, intra-communicating, spatially distributed sensor pods that can be easily deployed to monitor and explore new environments (Bizer, 2009) (Delin, 2001).

The SensorWeb is a framework that allows management & access to real-time heterogeneous datasets (Delin, 2001). The SensorWeb is a type of sensor network. However, sensorWebs are inherently different to sensor networks or a distributed set of communicating sensors. The goal of the SensorWeb is to extract and distribute Knowledge. Nodes or pods operating in a SensorWeb can modify behaviour based on data collected by other SensorWebs.

SensorWebs need enabling standardised service interfaces in order to create real-time accessible sensor data, this is similar to information on the WWW. Bröring et al. (2011b) notes that:

"Substantial effort is required to make a sensor and its observations available on the Sensor Web, since methods and mechanisms to automate this process are missing".

With the advent of the Internet of Things (IoT) (Atzori et al., 2010) (Andreev and Koucheryavy, 2012), the concept of a sensorWeb has been amalgamated with the concept of a Web of Things (WoT). In any case, the end goal of a sensor web or Web of Things is to extract knowledge from the individual data gathered by their constituent sensors and make this knowledge accessible in real-time. In terms of sensor-webs, this accessibility may or may not be through the WWW approach. Conversely, a geo-sensor network is a specific type of sensor network used to collect data about the physical world. Sensor Webs of geo-sensor networks seek to make datasets and streams available to support the geo-science research community. SensorWebs go beyond the IoT much in the same way the traditional Web provides a standardised documentation system on top of the traditional Internet.

2.4.1.3 Semantic SensorWeb

A semantic sensor network requires declarative specifications of sensing devices, the network, services, and the domain and its relation to the observations and measurements of the sensors and services (Compton, 2009). A core feature of the semantic sensor web is the use of *ontologies*. Ontologies are used to organise data into information and knowledge in a standardised way. Many ontologies have been developed to aid interoperability (Obrst, 2003).

In the Earth sciences domain NASA has defined the Semantic Web for Earth and Environmental Terminology (SWEET) ontology (Raskin and Pan, 2005). The SensorML based OntoSensor has also been defined (Goodwin and Caleb, 2006). The Semantic Sensor Networks Incubator Group which is part of the World Wide Web Consortium

(W3C) has developed the Semantic Sensor Network Ontology (SSNO) (Compton et al., 2012). The SSN ontology is aligned with classes in the DOLCE Ultra Lite (DUL) upper ontology (Masolo et al., 2003) (see Chapter 3). This alignment facilitates reuse and interoperability. Many ontologies do not align themselves, which makes interoperability difficult. SSNO is gaining wide acceptance and usage in the semantic sensor web community. A recent revamp of SSNO, which included lessons learnt from the original SSNO release also contains a realignment of SSNO concepts with OGC based concepts, which further increases its attractiveness as an ontology of choice for sensing applications (Taylor et al., 2019). Once aligned ontologies can be combined to provide more powerful semantics. For example, since SSNO's revamp, it can be also combined with the ontology SOSA (Sensor, Observation, Sample, and Actuator), SOSA provides additional rigour for individual axioms in sensing applications if needed (Janowicz et al., 2019).

Ontologies and related concepts are discussed in more detail in the next Chapter (Chapter 3). Before that, the remainder of this chapter presents the reader with an overview of the challenges presented in achieving interoperability within geo-spatial data and geospatial data infrastructures.

2.4.2 Technical Challenges & Interoperability Considerations

Achieving interoperability within geo-spatial data-centric geo-sensor networks is fundamental to address the research challenges described in chapter 1. Geo-sensor networks are highly subject to network churn. Network churn refers to the turnover rate of nodes interacting with the network. Reducing churn is necessary to ensure efficiency within geo-sensor networks (Pruteanu et al., 2011). *Micro-sensing* can be employed independent of a centralised server. Micro-sensing occurs at the edge of a sensor network where a collection of nodes coordinates to achieve a larger sensing task. For example, a deployment of water quality monitoring nodes along a river section may interact and

process data within a local mesh network without communicating to a backend server. In general, there is also a move towards decentralised IoT architectures, and thus the question of enabling semantic interoperability mechanisms at the edge of sensor networks is an area of growing research (Le-Tuan et al., 2020).

Within the computational field of geo-spatial information science there is a need for the development of algorithms for decentralised spatial computation, collaboration and event processes, including the detection of events between co-located sensor nodes. Typically, spatial information science-based algorithms are tailored to sparse sensor deployments and powerful computers. As the paradigm of how sensor data & information are made available has changed, intelligent and adaptive sensor platforms are needed, for measuring dynamic phenomena. Therefore, light weight in network data analysis needs interoperable data and information. Within geo-sensor networks there exists three levels of Interoperability

- Syntactic
- Semantic
- Process

A core feature of the semantic sensor web is the use of ontologies. However, ontologies in themselves present an integration issue that is particularly pertinent to multidisciplinary domains such as Earth sciences. Cooperation between multiple disciplines generally leads to a need to integrate multiple ontologies. The process of integration of ontologies is called ontology alignment. Ontology alignment is defined as the process of bringing ontologies into mutual agreement by the automatic discovery of mappings between related concepts (Martínez-Costa et al., 2010).

Data and information interoperability challenges and solutions (such as terminologies and ontologies) are discussed in more in chapter 3. For now, it is enough to highlight that

most solutions to interoperability typically require additional computing power to be employed to realise the solution. However, as mentioned above, typically in situ remote sensor based observing platforms and geo-sensor networks are technologically constrained in terms of battery power, processing power and communications ability. The next section provides the reader with a review of the types of technical constraints typically found in systems that are used to build observing platforms and clarifies the term *constrained system* used throughout this thesis.

2.4.2.1 Constraints at the point-of-capture (the sensor node)

16-bit MSP430 microcontrollers have typically dominated sensor mote platforms. Normally, during "sleep" they draw 1.3-2µA[15]. In contrast an ultra-low power 32-bit architecture (ARM cortex M3) draws 950µA[16]. Given that most of the operational life of a mote is spent asleep, current draw during sleep is a big consideration for system specification.

In 2005 Levis et al. predicted that there was no expectation for motes to move beyond a typical specification of approximately a 1-MIPS processor and tens of Kilobytes of storage. It was predicted that the benefits of Moore's Law would be applied to reduce size and cost, rather than increase capability (Levis et al., 2005). This prediction was somewhat naïve given the impending explosion of platforms driven by the hype of the IoT. However, such ultra-constrained sensor platforms are still pervasive today for many application areas, especially for geo-observational deployments where light weight geo-sensor network nodes are required. In other deployments such as ocean observing platforms ARM cortex A profile-based boards are more common.

[15] https://www.ti.com/lit/gpn/MSP430FG6425
[16] https://www.arm.com/products/silicon-ip-cpu/cortex-m/cortex-m3

On board computational processing is a major draw of battery power on sensing platforms. The chosen firmware (*bare-metal*) or operating system solution employed is a major contributing factor to the lifespan, development complexity and data processing capabilities of platforms. Operating Systems to enable the efficient development of applications on ultra-constrained mote platforms began to be investigated by the research community in the early 2000s. The focus of these *smaller* operating systems was to enable sensor networks to communicate and coordinate through standardised communication protocols such as Zigbee (Zigbee Alliance, 2006).

Levis et al. (2005) listed the main requirements for an operating systems design for sensor networks as focusing on:

- Limited resources
- Concurrency
- Flexibility
- Low Power

For brevity two key historically significant sensor network operating systems relevant to this work are presented, TinyOS and Contiki-NG. The latter OS is used as the OS of choice for the evaluation of this work (see chapter 5 for justifications and further discussions). However, it should be noted the area of sensor (or IoT) node operating systems is advancing at a fast pace and there are many other operating systems in existence.

TinyOS (Hill et al., 2000) (Levis, et al., 2005) emerged from the academic research community in 2000 on the back of a surging interest in sensor network research. Academics at UC Berkeley developed the sensor network operating systems in the first instance as a set of Perl scripts. After a number of revisions, TinyOS was re-written in the NesC language (Gay et al., 2003), a dialect of C. TinyOS is at the heart of its own

ecosystem that spans not just the research community, but also commercial systems such as Cisco's smart grid systems. TinyOS is considered "discouraging" to new users (Levis, 2012).

When compared with other embedded frameworks, TinyOS tends *not* to be the chosen solution for simpler sensing applications. TinyOS's evolution has always been with two major goals to the fore: minimising resource use and the prevention of software bugs. The later goal is a particularly problematic aspect of embedded systems development where debugging is not as fluid as in larger systems. In terms of remote deployments OS stability is also of primary concern as power cycling of platforms to achieve system reset tends to be difficult. The choice of NesC as the OS's native language was with bug minimisation in mind, meaning that it became difficult to write bugs into the software with the knock-on effect that it became difficult to write code for TinyOS platforms.

Despite its high entry learning-curve, TinyOS had been the de-facto OS choice for constrained sensor nodes for some time. This popularity appears to be waning in recent times and there has not been a major release of TinyOS since 2012 (TinyOS 2.2). However, development activity is ongoing (TinyOS Alliance, 2017) also TinyOS is still prevalent within the literature up to December 2020 (Queiroz, 2017) (Ahad et al., 2020) (Ali and Aslam, 2020).

Reusing (2012) highlights the overriding differences in TinyOS and Contiki. These differences are summarised next. TinyOS is suited to especially constrained hardware resources and Contiki offers a more flexibility when the hardware platform is not overtly scarce. TinyOS tends to cope better with limited resources as Contiki is a more complex operating system. TinyOS uses an event driven approach to concurrency where Contiki (which is also event driven) offers different levels of multithreading. Contiki offers more

flexible software replacement than TinyOS once deployed. TinyOS is more energy conservative (Reusing, 2012).

Other notable sensor network operating systems are: Mantis (Bhatti et al., 2005), SOS (Han et al., 2005), LiteOS (Cao, 2006). MansOS (Elsts, 2012) and RiotOS (Baccelli et al., 2013).

It is difficult to traverse the myriad of operating systems when deciding on a platform of choice. Each OS comes with optimisations for different purposes. For example, the purpose of LiteOS is to significantly reduce the learning curve for developers outside the sensor networks circles. Whereas configurability is the primary motivation and goal of SOS. The choice of OS is highly application specific, which is problematic when developing applications for a wide audience and even wider set of hardware platforms.

For this work the required processing power and associated software stacks available are a key consideration. What is found in the literature is that longevity, stability, and community support should be the main considerations where the specific technological considerations become somewhat arbitrary. For that reason, Linux should always be a primary consideration. Outside of Linux – and during this work - Contiki NG (Duquennoy, 2017) began showing promise as a platform to consider. Contiki NG is dealt with in more detail below.

Contiki-NG is a *fork* of the popular OS Contiki mentioned above (Duquennoy, 2017). The Contiki-NG project began in 2017 to improve a number of perceived short comings of the original Contiki operating system. The goal of Contiki-NG was to modernise the existing Contiki structure, configuration, logging and platforms to enable the OS to focus on dependable standard-based IPv6 communication and also to focus on modern IoT platforms, specifically 32-bit platforms such as the ARM Cortex M3/M4 and A8 profiles. It should be noted that Contiki-NG is a separate OS to Contiki and is maintained by a

separate community. The community support for Contiki-NG aims to take a more agile approach to development and streamline new feature adoption with periodic updates and releases[17].

To date, the Contiki-NG community has kept to their goals. Comparing the commit activities of both Contiki-NG and Contiki on their main Github branches (comparison performed by the author December 2020) shows that Contiki-NG is much more active with ongoing commit activity, whereas the last commit to the Contiki main branch was November 2018. This would suggest that Contiki-NG has now developed a richer development environment and is perhaps the best choice when beginning a *new* project. The discussion regarding embedded operating systems is continued within the evaluation section of this thesis, chapter 5.

Moving from the sensor node, observed data are typically transported from the observing platform using some form of communications network to ultimately be processed by some form of information management system. Having reviewed the technologies that exist at the point of observation capture (the sensor node) the discussion now moves to a review of these information management infrastructures. Modern infrastructures are used to manage not just in situ remote sensor based observational data but all Earth observational data and geo-spatial data. These large-scale systems are called spatial data infrastructures.

2.4.2.2 Knowledge Exchange in Pervasive Environments

The cloud computing paradigm suffers scalability challenges in large-scale deployments with many reporting nodes. To tackle the issue of scalability additional computing paradigms have emerged to complement cloud computing. Fog computing has been growing as a scalable distributed deployment solution in recent years (Iorga, 2017).

[17] https://github.com/contiki-ng/contiki-ng/wiki/More-about-Contiki%E2%80%90NG

Furthermore, fog computing layers themselves become saturated as the quantity of data grows and the network becomes unable to analyse and process the data. A new paradigm referred to as Mist computing, is emerging to deal with this challenge.

The US based National Institute of Standards and Technology (NIST) provides the following definition of mist computing:

> *Mist computing is a lightweight and rudimentary form of computing power that resides directly within the network fabric at the edge of the network fabric, the fog layer closest to the smart end-devices, using microcomputers and microcontrollers to feed into fog computing nodes and potentially onward towards the cloud computing services.* (Iorga, 2017).

Within mist computing limited computation is performed at the extreme edge of the network within the embedded nodes themselves. The mist computing paradigm has been shown to decrease latency while increasing the autonomy of nodes from the fog and cloud layers (Orsini et al., 2015). In pervasive environments, individual nodes interact and must share knowledge. Due to the deeply-embedded nature of these nodes, lightweight knowledge exchange mechanisms must be employed.

Sheth and Larson (1990) define the term *federated database* as a collection of database systems that are diverse autonomous but cooperate. They also differentiate between distributed database systems from federated database systems by stating that in distributed systems data are deliberately distributed to take advantage of distribution (increased availability and reliability), however in a federated system the distribution is a consequence of the existence of multiple databases systems before federation, a situation that also results in heterogeneity (Sheth and Larson, 1990). This is certainly the case within many pervasive systems and also within this work. However, here the goal is to resolve the heterogeneity of the data by fine-grained standardisation of the data models

across the federation and thus enable semantic interoperability to enable the exchange of standardised data, information and knowledge within pervasive systems.

Knowledge-based systems enable advanced levels of functionality as they form meaning from data. This meaning in a pervasive environment can in turn allow computing systems or individual nodes to extract facts from data. The work done here facilitates the possibility of light-weight knowledge exchange between observing platforms in remote in situ and constrained pervasive monitoring environments. This is largely achieved through the use of a *geo-templating kernel* which is described in detail in chapter 5.

2.5 Spatial Data Infrastructures (SDI)

Spatial data infrastructures (SDI) are online systems that serve spatial data in an efficient way. Coordinating agreements on technology standards provide key support for SDIs (Kuhn, 2005). Many SDIs only exist within singular jurisdictions; however, the real value of SDIs is realised when they are transnational. SDIs are typically comprised of many GIS systems. GIS systems act as singular nodes within a larger SDI. Modern SDIs have also been indicated as a practical cost-effective way to report on the progress of the UN sustainable development goals (Elenabaas, 2018).

Today the European Commission is advancing the goal of access to open data in a transparent way using systems such as SDIs. This goal has prompted several initiatives such as the INSPIRE directive (INSPIRE, 2007). The European Commission emphasise the role of standards in achieving its industrial policies and seeks to ensure all standardisation forces in Europe pull in the same direction (Simonis, 2019). INSPIRE is fundamental to facilitating the agreements that are necessary to achieve EU wide transnational SDI infrastructures.

The open data movement, in addition to supporting interoperability, has enabled the realisation of numerous data portals. For example, the European data portal[18] acts as a data sink by harvesting metadata from many public sector data portals. In Ireland, the Irish open data portal (ODP) was recently launched[19]. Ireland's ODP contains diverse sets of data from finance to health but has a sizeable geo-spatial component from various data publishers and is a good example of how information systems can contribute to the publishing and sharing of important scientific data for secondary use. These data portal go beyond spatial data. Here, for brevity, only data portals and infrastructure relating to spatial data are considered.

2.5.1 INSPIRE

INSPIRE is a European directive that seeks to harmonise spatial data across Europe. The INSPIRE directive sets the *minimum* conditions for interoperable sharing and exchange of spatial data and leverages standardisation outputs of the OGC. INSPIRE is primarily for spatial data and there are several specific data specification thematic areas, called *annex themes*[20]. Within some INSPIRE annex themes the annex's scope extends past just basic spatial information to include measured or sensed data about the real world i.e. observational data. For example, INSPIRE mandates the OGC's observations & measurements (O&M ISO/DIS 19156) (ISO, 2011) standard for the representation of observed data in annex 3, theme environment monitoring facilities. As such it is important to review INSPIRE in the context of this work.

The INSPIRE directive provides technical guidance to member states in how to implement certain identified technologies and standards. The INSPIRE directive also

[18] https://www.europeandataportal.eu/
[19] https://data.gov.ie/
[20] A list of up to date INSPIRE annex themes can be found here: https://inspire.ec.europa.eu/Themes/Data-Specifications/2892

includes some legally binding rules called *implementing rules* (see Figure 2.3), which includes mandating the use of O&M by EU member states for several themes. Therefore, the O&M must be considered central to this work to ensure its relevance to observational data collection within the EU.

It is worth exploring the way in which the INSPIRE directive specifies how the O&M standard should be employed by EU member states as well as some real-world implementations of the INSPIRE directive within an Irish context for sensed data.

2.5.1.1 INSPIRE Annex II

The INSPIRE directive Annexes I, II & III provide data specifications within INSPIRE. As mentioned above, each annex deals with a specific theme. Annex III deals with the largest number of themes and includes *Environmental Monitoring Facilities* (EF). These themes in turn provide technical guidance on implementation. For example INSPIRE document D2.8.II/III.7 provides technical guidelines for specifically implementing the environmental monitoring facilities specification (INSPIRE, 2013a). The INSPIRE document D.28 provides detailed implementing rules regarding the EF specification. The different processes around INSPIRE's implementing rules and technical guidance are illustrated in Figure 2.3 below.

The INSPIRE Environmental Monitoring Facilities (EF) data specification is part of the environmental monitoring and observations thematic cluster. Key to the EF data specification is the adoption of the O&M data model. The full UML model for EF and other INSPIRE related models are published in UML format within the INSPIRE consolidated UML model[21]. Among all INSPIRE themes, The EF theme makes the heaviest use of O&M (INSPIRE, 2013a). The adoption of O&M has implications for the work presented in this thesis. As will be seen later, O&M has now become a well-

[21] https://inspire.ec.europa.eu/data-model/approved/r4618-ir/html/index.htm?goto=2:3:6:1:1:7980

established standard within the environmental monitoring community. O&M is also indicated in a number of themes that reside in both Annex II & III. INSPIRE provides further guidance on the use of O&M for all adopting themes in INPIRE document D2.9 (INSPIRE, 2016).

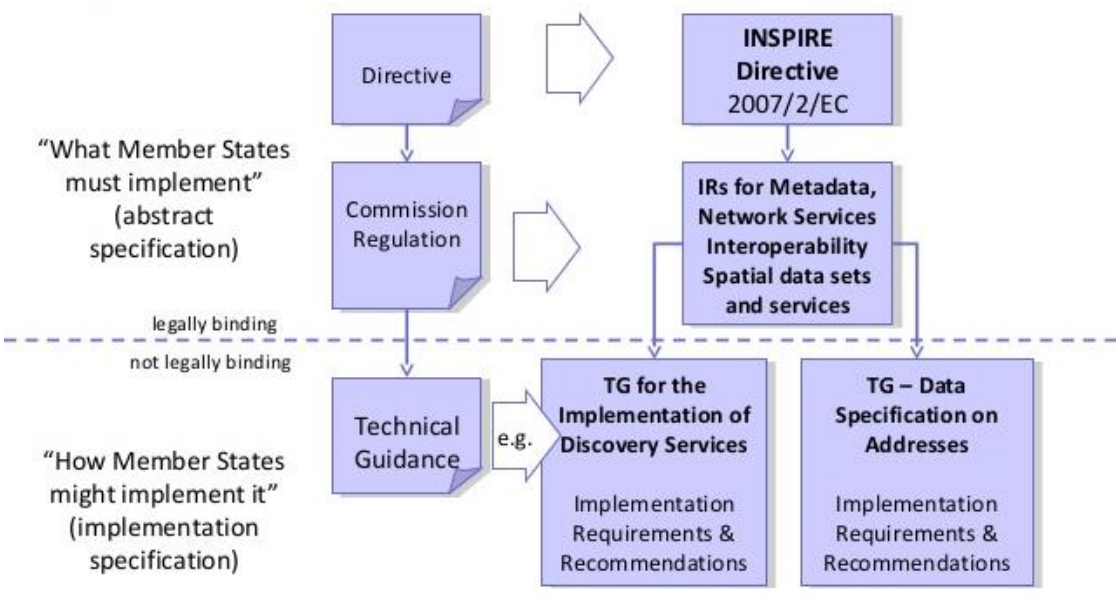


Figure 2.3 INSPIRE Implementing Rules vs. Technical Guidance (INSPIRE, 2007)

The adoption of O&M within INSPIRE implementing rules for several themes elevated the O&M standard in terms of its importance within geo-observation systems design. O&M now serves as the de facto standard to use when reporting observation and measurement data, especially within indicated themes such as environmental facilities monitoring.

2.5.1.2 INSPIRE Pilots

INSPIRE has also run several pilots in key policy areas related to INSPIRE to facilitate up take. As of December 2020, three pilot studies have been undertaken:

- INSPIRE Energy Pilot[22]
- INSPIRE Marine Pilot[23]
- INSPIRE Transportation Pilot[24]

This work is primarily focused on marine use cases, and specifically ocean observing activities. Therefore, the INSPIRE marine pilot is of primary interest within the thesis. In fact, the marine pilot approach is the basis for the evaluation method described in the Chapter 6. The next section presents a review of ocean observing based SDIs. For the ocean observing based SDIs the implementation status of INSPIRE is also noted.

2.5.1.3 INSPIRE Implementation Status

The European Commission is the body who over sees the INSPIRE road map[25] and adherence to implementing rules against key dates. The Joint Research Council (JRC) regularly publishes reports highlighting the implementation status of INSPIRE. The most recent report was published in 2017 (Cetl, 2017). The implementation status report details each countries progress in implementing the INSPIRE directive's implementing rules. Ireland's progress has been mixed when compared to other EU countries. In the 2017 report, Ireland's overall implementation status and trend was rated as "made some progress but still far from being complete, outstanding issues are significant". The report also noted that the lack of interoperable pan-European information products limits the use of the data beyond INSPIRE communities. The committee found many non-interoperable datasets that cannot be used in cross border applications.

Earth observation is a vast activity, and to be useful, this work is applied to one specific domain, while aiming to be easily translatable to any Earth system science area or Earth

[22] https://inspire.ec.europa.eu/pilot-projects/inspire-energy-pilot/440
[23] https://inspire.ec.europa.eu/pilot-projects/inspire-marine-pilot/438
[24] https://inspire.ec.europa.eu/pilot-projects/inspire-transportation-pilot/439
[25] https://inspire.ec.europa.eu/road-map-graphic/32443

observing activity. As mentioned in Chapter 1, section 1.6.1 (Research Design) the main application area of this work is for ocean observing scenarios and one of the evaluation activities involves data that is taken from ocean observing activities. Therefore, next a review of ocean observing data portals and SDIs is presented. The next section also reports on the level of compliance of ocean observing SDIs to the INSPIRE directive.

2.5.2 Ocean Observing SDIs

Within the ocean observing community EMODnet (European Commission, 2010), SeaDataNet (Schaap and Lowrt, 2010), JericoNEXT (Antonie, Sandrine and Jean-Valery, 2017) and AtlantOS (Fischer, 2016) have emerged as key spatial data infrastructures to manage the vast amounts of ocean data. These initiatives subsequently advance a complementary international goal of interoperable and open ocean data. For example, SeaDataNet contributes to the Ocean Data Interoperability Platform (ODIP) (Glaves et al., 2014). ODIP brings together all the key ocean data management organizations from the EU, US and Australia. ODIP in turn is promoted by IOC/IODE (UNESCO, 2018) and other international consortia to help achieve global ocean data interoperability. Through ODIP, EU projects such as INSPIRE are having a global impact. For example, the adoption of the Observations & Measurements standard within INSPIRE has seen O&M become a key component of the GEO-DAB discovery and access broker (Nativi and Bemmelen, 2016), this further highlights the importance of O&M as a data standard. GEO-DAB connects more than 150 international providers of high-quality Earth Observations. The continued investment in open and interoperable ocean spatial data infrastructures (SDI) around the world is beginning to realize dividends. However, there are still many challenges to overcome.

The Columbus project (Columbus Consortium, 2016) has also performed a broad review of ocean data portals. Their work is not exhaustive but highlights the wealth of

available SDIs and portals. The Columbus review is unique as its goal is to create measurable growth in the blue economy. It is also tasked with monitoring the implementation of the Marine Strategy Framework Directive (MSFD) (Olenin et al., 2010). Thus, the focus is on the ability of marine spatial data infrastructures to encourage and enable end users develop value added services and products. In their analysis it was found many marine data portals are built from a developer's perspective on the intended purpose, and not the end user. Therefore, ease of use and user friendliness of data sharing facilities can impede the wider sharing of collected data (Columbus Consortium, 2016).

2.5.2.1 Ocean Data Portals

Downstream services such as EMODnet-physics greatly enhance the ability of end users to consume high quality marine data products. New applications arising from the availability of high-quality data need to be cognizant of the EU Inspire Directive. With a combination of Copernicus Marine Environment Monitoring Service's (CMEMS) (Von Schuckmann et al., 2016) In Situ Thematic Centre (IN STAC) (Copernicus, 2018) and EMODnet users have access to harmonized open access data that has under gone automatic and manual data quality checks, and have been augmented with additional metadata. EMODnet's gateway contains seven thematic data portals.

The EMODnet-physics data ingestion process allows data providers to contribute their dataset directly to the EMODnet operational oceanography data exchange. Data providers will typically collect, control and distribute their data based on their own rules (EMODnet, 2018). EMODnet provides regional coordinators to work with data providers to enable the setup of new data flows. Where data providers are not in the position to harmonize their datasets with the EMODnet system, regional coordinators perform the task of data harvesting and harmonization.

EMODnet-physics acts as a downstream service for CMEMS-INSTAC and SeaDataNet. The CMEMS-INSTAC service performs the harmonization and automatic quality control on datasets at one of five regional centres. Quality checks are defined by the EuroGOOS Data Management Exchange and Quality Working Group (DATAMEQ) (Pouliquen, 2011). A conversion to a unique netCDF format is performed at Regional Data Acquisition Centers (RDAC) by trained staff. INS-TAC uses the OceanSITES netCDF format (OceanSites, 2015). OceanSITES netCDF is Climate and Forecast (CF) standard (Gregory, 2003) compliant and is recommended by CMEMS and EuroGOOS. INS-TAC produces quality-controlled aggregations of in situ observational data using OceanSITES netCDF. To aid this process, CMEMS provides the oceanotron server to manage the dissemination of data collections (Copernicus, 2017). The data model employed by oceanotron is based on the Climate Science Modelling Language (CSML) (Woolf et al., 2005) and aims to be compliant with O&M and CF (Climate Forecast) discrete sampling feature. CSML is in fact a specialist profile of O&M. CSML 3.0 is based on O&M and is aligned with binary CF netCDF.

2.5.2.2 NetCDF-CF

The netCDF standardized data model is domain independent (Rew and Davis, 1990). NetCDF specifies that datasets should be self-describing. However, netCDF files are not mandated to be self-describing. NetCDF files contain both array-oriented data and metadata. Due to its generic nature, netCDF is not specific to any domain, and so has wide applicability. Also, due its generic data model, further metadata standards are usually employed within a domain to ensure data served in netCDF are interoperable. As is the case with OceanSITES netCDF mentioned above, the CF metadata standard is often combined with netCDF to describe in further detail how to encode oceanographic and other geographical feature-based datasets. CF enables additional constraints to be applied

to netCDF datasets in terms of space, time, units and standard naming conventions etc. CF conventions require implementing datasets to contain sufficient self-describing metadata so that each variable has an appropriate level of descriptive metadata.

One of the core advantages of using the CF conventions to describe data is the CF standard-names controlled vocabulary (Eaton et al., 2003). The standard names are used when describing geophysical quantities. For example, sea water temperature is standardized to the entry id sea_water_temperature. CF standard names include associated units and a description of the represented quantity. For example, to further describe sea water temperature at a particular depth, a vertical coordinate variable should also be included in the dataset. There has been some criticism of CF conventions, as many attributes are optional. This means that data providers have typically omitted the attributes that are needed to fully understand the meaning of the structure of the data (NASA, 2019).

CF conventions are based on an open governance model with a bottom up standards process. This means that any community member can propose changes to the conventions. One central point that is relevant to this work is that the community consensus approach employed by CF conventions have been key to its success. This approach has allowed the bridging of a diverse group of earth system modelling communities. CF conventions are documented in online resources. However, these resources do not allow for immediate discovery and integration of datasets. The netCDF-LD extension (Car et al., 2017) seeks to allow the creation of netCDF compliant files that can also support linked open data principles. Implementing CF conventions with Attribute Conventions for Data Discovery (ACDD) (Davis, 2005) can also enhance data linking and data discovery when processing datasets.

2.5.2.3 INSPIRE & Oceansites netCDF Format

Within INSPIRE IR Requirement Annex IV (INSPIRE, 2013b) it states that any data related to the theme oceanographic geographical features (OF) shall be made available using a number of types, such as:

- PointObservation
- PointTimeSeriesObservation

All types listed in (INSPIRE, 2013a) and above are constraints to the O&M model. INSPIRE maintains a managed code list of recommended terms including the CF standard names. The INSPIRE ocean geographical features theme uses the O&M standard to ensure consistent encoding of observations. Observations can be measured, modelled and simulated. As O&M is a generic model, INSPIRE provides numerous extensions. One important extension to O&M is the complex properties model (INSPIRE, 2013a). The complex properties model allows system developers to produce interoperable observational data with the necessary fine-grained detail to describe the properties of the observation. However, Leadbetter and Vodden (2016) argue that the existing INSPIRE complex properties extension is too abstract in terms of real-world implementation. Highlighted is the fact that ocean observations typically require a quantity and a mathematical approach to describe the observed property. The initial captured quantity may undergo statistical transformation and adjustment before being encoded in the data stream. However, the details of the statistical process used is not captured in the dataset. This is typically important information, needed for re-use of the processed data and is a particular limitation for achieving fine grained interoperability of published datasets. As INSPIRE sets the minimum standards for interoperability additional approaches are needed. This is the central theme of the work presented in this thesis.

OceanSITES includes a quality check (QC) metadata for each data item. The reported QC indicator is typically on a simple scale (0-6 for example). However, the more detailed process of how the QC indicator was arrived at is not automatically linked with the actual dataset. It has been proposed that netCDF-LD can provide a solution to this, allowing provenance to be captured in the metadata, separate to the actual data and thus reducing the overhead of quality information tied to datasets.

By the end of 2020 all INSPIRE obligations must be implemented by EU member states. EMODnet aims to use INSPIRE standards. However as noted by Millard (2015) EMODnet may require solutions that diverge from INSPIRE, again providing additional argument for the need for additional solutions. Again, this forms part of the core research objectives of this work (chapter 1, section 1.5). EMODnet (2018) gives a detailed overview of EMODnet compliance with INSPIRE, which is overly detailed for the purposes of discussion to be included here. More relevant is that EMODnet has conducted a number of pilots (mentioned above in section 2.5.1.2) such as the real-time oceanography data exchange pilot using SWE (ODIP, 2018) (discussed in more detail below, section 2.5.3.2). These pilots have informed the evaluation approach used within this work (described later in chapter 6).

2.5.3 State-of-the-Art in Standards Implementation

There is a myriad of examples of deployed state-of-the-art and best in class SDIs. However, all SDIs differ in their prioritisation of implementation features. Gomes et al. (2020) provide a useful overview and comparison of seven new generation SDIs for big Earth observation data management and analysis. Even within these new generation of SDIs, standardisation efforts often exist on the periphery of many (SDI) implementation agendas. This is to be expected given tight budgets and deployment deadlines for the scenario of use. This is evidenced within the review provided by Gomes et al. (2020)

where data interoperability capabilities do not form part of their review. To encourage uptake, in 2013 INSPIRE established the maintenance and implementation group[26] (MIG).

The MIG adopt a supportive (rather than punitive) approach that encourages the sharing of implementation experiences and practices among those impacted by INSPIRE. As part of the work of the MIG, a useful toolkit[27] is maintained to aid INSPIRE implementers.

Here the current state-of-the art in standards adoption within marine spatial data infrastructures is considered. These are considered from an interoperability perspective. Firstly, the Global Ocean Observing System is considered as an exemplar of the global effort to combine ocean observing SDIs. Then the INSPIRE marine pilot is considered to show the current state-of- the art in standards implementation beyond that of GOOS and similar.

2.5.3.1 GOOS

The effort to realise a global ocean observing system (GOOS) can be traced back to 1993 when a memorandum of understanding was signed between the Intergovernmental Oceanographic Commission (IOC) and the World Meteorological Organisation (WMO) and others (Flemming, 1995). Nicholas Flemming (referred to as "Father GOOS") made an economic case for GOOS in Flemming (1995). There, Flemming noted that local observing systems had short time horizons and that a patchwork of these systems may in fact be more expensive to deploy but would produce less operational benefits.

[26] https://inspire.ec.europa.eu/inspire-maintenance-and-implementation/46
[27] https://inspire.ec.europa.eu/inspire-tools

Today, GOOS is a rich collection of in situ networks, satellite systems, governments, UN agencies and individual scientists[28]. However, interoperability of datasets and information are still a work in work progress.

In May 2019 GOOS published its 2030 strategy, which contained a key commitment to system integration and delivery, and specifically to ensuring GOOS ocean observing data and information are *findable, accessible, interoperable and reusable*, with appropriate quality and latency (GOOS, 2019).

2.5.3.2 INSPIRE Marine Pilot

The INSPIRE Marine Pilot has been used as the basis to develop the ocean observing use case evaluation approach for this work (evaluation 3, research canvas, Chapter 1, Figure 1.2). This pilot is therefore central to this work and is explained in more detail here. An overview is provided here for context and to add specificity to the discussion regarding state-of-the-art SDIs (the actual evaluation approach adopted based on the INSPIRE marine pilot is detailed later in Chapter 6).

The INSPIRE Marine Pilot crosses 6 different themes, including the EF (Environmental-monitoring Facilities) theme. The primary aim of this pilot was to investigate INSPIRE in the context of the Marine Strategy Framework Directive (MSFD), while developing tools to facilitate INSPIRE uptake to meet the MSFD obligations. The pilot shows by way of a number of datasets examples of how INSPIRE may be adopted within a marine environment. The pilot focused on chlorophyll-α datasets. A use-case evaluation approach was adopted. The use case was intended to:

- Harmonize the data of chlorophyll-α concentrations in the cross-border area of the Denmark, Germany, and the Netherlands
- Create the metadata for the data

[28] https://www.goosocean.org/

- Publish the metadata, and share the data using INSPIRE services thus fulfilling the requirements of MSFD Art. 19; and Use the services in an application that does some analysis on the harmonised data from the three countries.

Time series information is required to provide a sequence of data points/areas, measurements made over a time interval, linked to the sampling station (or area divided into grid) within their location. Time series data linked to the monitoring station (or area) has unchanged location during a monitoring period. Each monitoring station is related to at least one but could be related to more than one monitoring programme/sub-programme. The same location could be used for sampling on various indicators related to the different *quality descriptors* (QD) such as chlorophyll-a, nutrients[29] (sub-programmes of eutrophication-QD-5) and heavy metals (sub-programme of concentrations of contaminates QD-8).

QD5 Human induced eutrophication are identified by the following groups:

- Nutrients concentration
- Nutrient ratios
- Chlorophyll concentration
- Water transparency
- Dissolved oxygen

These types of spatial data are mandated to be modelled using application schemas based on Oceanographic geographical features (OF) that represents the physical or chemical (including chlorophyll a, as estimated on the physical property - ocean colour) properties of a sea.

[29] https://www.britannica.com/science/eutrophication

The OF model is based on the ISO 19156 Observations and Measurements (O&M) framework for consistent encoding of measured, modelled, or simulated data. For the purposes of interoperability in INSPIRE, the O&M model is profiled to add further precision about the types of processes, observable properties and features of interest that are used. O&M is profiled into Specialized Observations Types that differs grid, point, multipoint and trajectory observations, including the time series for each of the sampling geometries, that are common to Atmospheric Conditions/ Meteorological geographical features theme and are part of INSPIRE Generic Conceptual Model (GCM).

The results of the marine pilot were a requirements analysis, data model recommendations that align with EMODnet and INSPIRE including tools to implement data flows that has been key to defining the main evaluation of this work (described later in chapter 6).

2.6 Discussion & Conclusion

Thatcher et al. (2016) and Singleton & Arribas-Bel (2019) show how the GIS-wars of old have now led to a much more cohesive, community approach to the digitisation of geographical data. This increased collaboration between domain expert and informatician (or GI Scientist) has born such organisations as the OGC, which has in turn greatly progressed the development of critical data models such as O&M and standardised infrastructure access interfaces such as SOS.

The open data movement has broken down many of the barriers that individual jurisdictions and private organisations have faced in the past when seeking to publish datasets to publicly accessible data portals as free and open data. However, despite these advances, the exploitation of geospatial data portals and open geospatial data remains low.

Retrieving data from current spatial data infrastructures can be a cumbersome process. For example, current ocean-based SDI implementations do not allow for easy automatic discovery and federation of ocean observational data flows. There are many reasons for this, one aspect is the issue of the consistency of data formatting and data quality representation within these datasets, beyond that of data container formats such as netCDF. This hinders the development of data consuming applications, as the development of software in the face of large-scale heterogeneity becomes difficult and laborious, as software must be hard-coded and *hacked* for each dataset. Data harmonisation then becomes difficult or impossible and this ultimately results in *hidden knowledge*. Hidden knowledge is pervasive in all information management environments. Email is often cited as a good example of hidden knowledge. Within SDIs this may be geospatial data that is not accessible or searchable due to non-accessible storage, or inadequate metadata representation. Hidden geospatial knowledge is a missed opportunity to apply this knowledge to help solve the complex anthropogenically induced problems of our time (discussed in chapter 1).

Semantic search can be used to mine large datasets and expose hidden knowledge. However, semantic search can only be enabled when semantic interoperability mechanisms are employed within SDIs, which is often not the case. The description of the CMEMS-INSTAC service earlier highlights the manual effort that is required to ensure data are at the very least syntactically harmonised and interoperable. Semantic interoperability is the next level and is still very much a work in progress. Even current approaches to semantic interoperability are still deficient (discussed in more detail in the next chapter) in the representation of data and information, as the processes used to develop semantic annotations are not conducive to capturing the domain practitioners knowledge, a key complaint of the GIS-wars. The battle between geographers and GI

scientists may be over, but as the catalyst for the war is not won, but today a collaborative approach is now underway to solving these shortcomings.

Moreover, the trend towards the integration of geo-sensor networks, EO systems and sensed data into large scale spatial data infrastructures requires mechanisms for the sharing and processing of data across highly heterogeneous sensor-based systems. Standardisation of sensed data is essential, and initiatives such as the INSPIRE directive, which employ standards developed by the OGC, are helping to realise the implementation detail needed. However, employing data specifications at the node level is not always possible due to the constrained nature of the platforms from a processing, storage, power, and transmission perspective. This is especially evident when employing spatio-temporal semantics (Dukham et al., 2010). These are all significant barriers to achieving Gore's vision of a Digital Earth as was discussed in chapter 1.

Ongoing work aimed at solving the core problems of semantic interoperability in geospatial data and information is accelerating. Indeed, these issues are not unique to the geographic domain, and are the focus of much research in other complex domains. Diviacco and Leadbetter (2017) highlighted the need for Earth system science domains to investigate solutions to semantic interoperable systems that occur on the fringes of Earth system science. To understand the fundamental issue of semantic interoperability and the potential for fringe domain solutions to be used within the geospatial domain to tackle the semantic interoperability problem, a full review of methods used to represent information and knowledge is needed, including a review of methods applied in other domains.

The next chapter introduces the reader to the core concepts of semantic interoperability, including a review of the current state-of-the-art relating to achieving

semantic interoperability within geo-spatial infrastructures and other fringe complex domains such as health (health informatics).

Chapter 3 ultimately deals with the progenitive question to the research question (posed in chapter 1): *are there more advanced semantic interoperability methodologies within other complex domains that could be adopted within GIScience and SDIs to help improve semantic interoperability?*

Chapter 3

"We know more than we can tell"
(Polanyi 1941)

3. SEMANTIC INTEROPERABILITY

Chapter Overview: Chapter 1 discussed how interoperability and semantic interoperability remains a key barrier to realising a Digital Earth system. Chapter 2 described how interoperability of information occurs at several levels and showed how several wide scale initiatives (such as the INSPIRE directive) are being progressed to solve interoperability issues within the geo-spatial domain. Many of these initiatives are aimed at solving the syntactic level of interoperability.

To facilitate semantic interoperability, information needs to be recorded in a way that allows the meaning, context, and lineage of the information to be determined. This level of recording is complex and difficult to achieve in practice, however this is the goal of this work.

Here, several fundamental semantic interoperability concepts are introduced. Also, a review of how data, information & knowledge are represented and persisted in machine readable formats to enable interoperability is provided. The challenges of capturing knowledge in machine readable format are described, initially from a philosophical perspective, but then later from a technical perspective.

This chapter also reviews current state-of-the-art approaches that are employed to achieve semantic interoperability of recorded information and introduces the reader to methods employed in non-Earth system science-based domains (i.e. the Health domain

and the two-level modelling approach). These additional methods may contribute to advancing semantic interoperability and the production of high quality and computable globally shared documentation to support Earth Sciences research.

3.1 Data, Information & Knowledge Representation

As can be seen at this point in the discussion, data, information, and knowledge are fundamental concepts to this work. Data can be defined as the facts regarding the *real world*. Data can be perceived using human senses or indeed man-made sensors. The recording of data happens in countless ways, but databases are used to ensure data are recorded in a safe and accessible way.

Information is different to data; in that it is *structured data* that is usually supported by additional context data. Structuring data into information makes the data more readily actionable. Knowledge is more difficult to define, capture and use. However, it is also highly valuable, consisting of relationships between a conscious subject and a portion of reality (Zagzebski, 2017). Most knowledge is *tacit* and resides in the human brain, such as knowing how to ride a bike which is typically passed on through socialisation and mentoring. Due to the nature of knowledge and the complex relationships therein, the recording of knowledge is hugely difficult. As can be seen above, data, information & knowledge are interconnected.

Thierauf (1999) provides a useful definition of all three:

> "*data is the lowest point, an unstructured collection of facts and figures; information is the next level, and it is regarded as structured data; finally knowledge is defined as "information about information*".

The *knowledge triangle* (Rowley, 2007) (Figure 3.1 below) is often used to visually depict the interconnection of all three. In Figure 3.1 a fourth level can also been seen, referred to as *wisdom*.

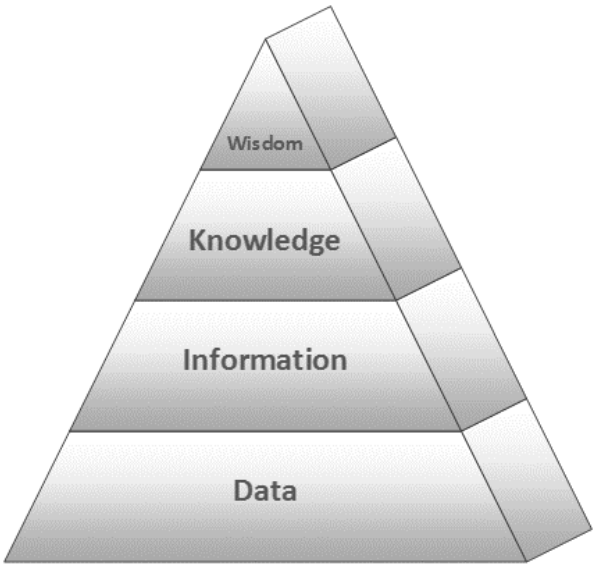


Figure 3.1 Knowledge triangle (Rowley, 2007)

Wisdom is knowing when and how to apply knowledge. Wisdom is a further processing of knowledge, and with this further processing comes the added complexity of recording.

It is perhaps useful to consider a simple example to illustrate the relationship between the four levels.

1. *Data*: rainfall=2cm.
2. *Information*: Rainfall at Leenane weather station [geo tag] today [todays date] = 2.1cm, using xyx rain gauge.
3. *Knowledge*: Killary Harbour is one of the wettest places in Ireland
4. *Wisdom*: If you are visiting Leenane, it is useful to bring, raingear or an umbrella just in case.

In this work, only the first 3 levels within the knowledge triangle (data-information-knowledge) are considered.

To record data and information or indeed knowledge, some form of representation or *formalism* must be used. Traditionally books served this purpose. However, books are not easily understood by machines (computable). Machine-readable representations and supporting management systems require that the information is recorded at a certain level of quality to be useful. *Data quality* is important to ensure that the data, information and knowledge leads to good decision making (see Figure 3.2 below).

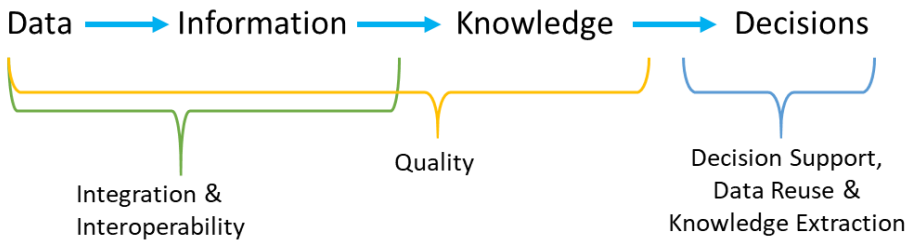


Figure 3.2 Relationship between data quality and decision support

Completeness is often one measure of the quality of data (Ballou and Pazer, 1985). To ensure completeness, one must consider all the attributes of the artefact that need to be recorded. These attributes are not always immediately obvious, as they are often not tangible and lie outside of the physical world. They may for example include feelings or perceptions, i.e. psychological artefacts. Or they may be ideas that do not yet have a physical manifestation. At a basic record level, completeness refers to whether all mandatory fields are present, such as name and address within a patient's health record. However, often data records are not adequately designed, and mandatory or optional fields are ill considered. Fields may not be included at all, or many fields are set as optional.

The prominent 20th century philosopher Karl Popper's seminal work on objective knowledge (Popper, 1948), highlights the need to go beyond the physical and indeed conscious world when considering knowledge representation. When considering what systems should represent to achieve *completeness*, a careful consideration of Popper's

theories is essential. While Popper's work in this area is philosophical in nature, it is core to understanding the main research problem this work addresses (see chapter 1). Also, Popper's work in this area is fundamental to understanding the main hypothesis presented in chapter 1, i.e. that many of the proposed approaches to solving semantic interoperability within the Earth sciences do not go far enough or are inadequate and methods developed for other domains could be adopted within the Earth sciences.

Popper's three worlds is a useful construct to explain why those approaches are inadequate, and additional methods are necessary to address knowledge representation within the Earth science domain. Popper's three worlds are introduced next.

3.1.1 Popper's Three Worlds

Popper proposes a pluralist view of the universe, made up of three different but interacting sub-universes. These are referred to as *Popper's three worlds* and were first described by Popper during the *Tanner lectures on human values* at the University of Michigan in 1978 (Popper, 1978).

Popper's three worlds are made up of the *physical world* (world 1), the *psychological world* (world 2) and the world that is the *product of the human mind* (world 3).

According to Popper, world 1 consists of things that are made up of physical energy, such as plants, animals or radiation etc. He notes we can also subdivide this physical world into the world of living and non-living things.

World 2 is the mental or psychological world, made up of our thoughts, feelings, and decisions. Like world 1, world 2 may also be sub-divided. However, these levels of distinction are not necessary for the current discussion.

At the time of Popper's lecture, many people within the Philosophical fields were supportive of a dualist view. However, Popper's main proposition was a defence of the existence of a third world. A world containing theoretical systems, problems, problem

situations, critical arguments and the contents of books and libraries. These products of the human mind can best be explained by considering the case of a book.

A physical book, or indeed a printed copy of this thesis, even in different forms can be said to belong to world 1, the physical world. However, the thesis itself, which is a manifestation or product of mind of the author, can be said to belong to world 3; whereas the physical world (world 1, the printed copy) *embodies* that which belongs to world 3. As such, world 3 objects are *abstract* objects, and their physical realisations are *concrete* objects.

At the time, Popper had been accused of hypostatization, with many rejecting the idea of the existence of world 3 as misleading. However, Popper's three worlds are fundamental to understanding the nature of information systems, information modelling and in fact semantic interoperability itself. This is because interoperability within information systems can occur at many levels.

There are two types of semantic heterogeneity that can occur, *cognitive heterogeneity* & *naming heterogeneity* (Klien, Lutz & Kuhn, 2009). Both types arise due to different perspectives of real-world facts. Naming heterogeneity arises when the same term is used to describe these different perspectives. Where naming heterogeneity exists within datasets then the interoperability of those datasets is compromised. The naming conflicts must be resolved manually or using some form of mapping or harmonisation algorithm between the datasets (see chapter 2, section 2.5.1 discussion on CMEMS-INSTAC for a real-world example of this issue). Encoding this can be difficult as often the heterogeneity in naming has subtle complexities that only a domain specialist may fully understand.

Basic interoperability therefore occurs at the syntactic level (mentioned previously in chapter 2), where syntax rules must be applied. The syntactic level can be related to world 1 objects, or things which can be named (avoiding naming heterogeneity). This is where

physical embodiments, even in the form of an information instance on disk, are standardised to some agreed terminology and, or syntax.

However, true interoperability at the semantic level, where the true meaning of the *thing*, *entity* or information object must begin at the abstraction of the concrete object, i.e. the abstract object. To achieve full semantic interoperability within concrete information systems and information objects, one must first accept the existence of world 3 and its relationship to world 1; and accept that even with careful rigorous recording of abstract objects true interoperability can be lost, due to insufficient mechanisms to capture world 3 objects.

A central element of the discourse about information systems and relational databases is the concept of an *entity*. An entity is the seed of what will ultimately become a relation (in the relational algebraic sense) or a concrete relational table. An entity is anything that exists. Were we to take a monist or pluralist view (such as Popper's protagonists would have) to entities, the result of any entity-relational modelling process - which is key to successful relational database design - would be wholly inadequate for capturing the problem domain. World 3 recognises the need for a standalone system of agreed concepts, in the form of a terminology or ontology that can be adopted by a community as the basis for communication of agreed semantic content. A consensus-based approach to developing world 3 representations to assist common understanding in the ESS space is a core element of this work.

Relating this discussion back to the knowledge triangle (Figure 3.1), we can see that the human mind spans the knowledge triangle. Therefore, any products of the human mind (world 3 objects) are produced using data, information, knowledge and wisdom. However, embodiments of these objects within world 1 is thus difficult. As in the

discipline of information science, the formalisms used to create concrete objects are typically insufficient to fully capture the complexities of the abstract object.

Having considered some philosophical underpinnings of knowledge and entities, needed to later understand the real problems this thesis seeks to address, the discussion moves to examine information modelling and models which are fundamental to realising concrete information systems.

3.1.2 Information Models

Capturing the complexities of information and knowledge about the world(s) around us requires us to abstract concepts away from certain details. These abstractions allow us to focus on important concepts while hiding their details. These abstractions are called *models*.

Models help in the organisation of knowledge, while also helping to communicate concepts and information in an understandable way. Models allow relationships between primitive and complex phenomena to be captured; this in turn can help us to explain the world around us. Models can also allow different viewpoints to exist and allows for the productive exploration of these differing viewpoints, discovering commonalities and influencing each other by showing new perspectives on the modelled phenomena.

There are numerous advanced modelling techniques, such as entity relationship modelling (ER Modelling) and object-oriented modelling (OO Modelling). These techniques use a visual vocabulary and a standardised methodology to arrive at a final model consensus among informaticians, which seeks to capture their understanding of the inputs and viewpoints of stakeholders.

Both ER modelling and OO modelling have at their core the idea of an *entity* or *class*, and the idea that there can exist relationships between disparate entities or classes. Differences exist in the *expressivity* of the modelling techniques when they are applied to

concrete systems. For example, when ER models are used to realise relational database systems, many-to-many relationships must be *solved*, whereas OO models and resultant OO databases allow for many-to-many relationships. This is because relational databases must adhere to logic of its relational algebraic engine, which in turn supports the application of structured query statements to databases.

One of the many difficulties within information modelling is deciding on which entities to include in the model, especially when concepts are abstract. Information modelling typically follows a structured process which requires informaticians to define several models in a stepwise fashion. For example, firstly a conceptual model may be defined, which may be further refined to a logical model and ultimately a physical model. This process is referred to as *reification* (Friedman and Wand, 1984). Reification turns something that was abstract or implicit into something explicit within a software system. Through reification, the abstract becomes a computable resource that may be manipulated and shared. For example, at an object-oriented coding level, the definition of a class object only becomes reified when the object is instantiated in memory. Reification is also referred to the process of making something a *first-class citizen*.

The concept of first-class citizens was first developed by Christopher Strachey (Strachey, 1966) to describe functions of objects that had certain core properties. First class entities in data systems are data objects that can referenced, passed as parameters etc. It should be noted that not all entities or objects are first class, second class citizens are also common, these are objects that have limited functionality and cannot not necessarily be referenced or manipulated directly. First class citizens are only considered as part of this discussion and are dealt with in more detail later on.

3.1.3 Ontologies & Formal Representation

An ontology is an explicit terminology specification, which formalises a conceptualisation of a body of knowledge, in some area of interest (Gruber, 1983). As a formal specification of the terms within a domain, ontologies enable reuse of domain knowledge. In the context of information systems, ontologies are being used to increase interoperability by structuring and formalising knowledge within a domain.

In recent times ontologies have garnered a broader interest across many domains, including GIScience (Bittner, Donnelly, and Smith, 2009). Previously, ontologies were used in more specialised applications, such as within artificial intelligence. Today, ontologies are used within desktop and Web applications.

Developing an ontology involves the following steps (Noy and McGuiness, 2001):

- Defining classes within the ontology
- Arranging the classes in a taxonomic hierarchy
- Defining slots and describing allowed values for these slots
- Filling in the slots for the instance

There is no one correct way to model a domain, there are always viable alternatives. The best solution always depends on the application that is in mind and the anticipated extensions. Therefore, ontology development is - and should be - an iterative and never-ending process. Ontologies are models of reality (world 3) and chosen concepts during the development process should reflect this.

One area where ontologies have seen a huge level of use is the World Wide Web. Many websites such as Amazon and Netflix are using ontologies to enhance their user experience. The WWW Consortium (W3C) defines the Resource Description Framework (RDF) (Klyne and Graham, 2006) and Web Ontology Language (OWL) (McGuinness and Deborah, 2004). These standards developed by the W3C are the pillars of what is

referred to as the semantic Web. At the same time, these technologies and approaches have been explored within the geospatial community towards developing a geospatial semanticWeb (Egenhofer, 2002).

In Ireland, with the advent of COVID-19, semantic Web enabled geospatial infrastructures are now mainstream. The Irish government's geospatial data portal is driven by semantic Web technologies and has been used by 1000s of citizens daily during the COVID 19 pandemic to gain insight into the progression of the disease[30].

The Semantic Web is an extension to the current Web, in which meaningful relationships between resources are represented in machine readable format. RDF is a language for encoding knowledge in Webpages. OWL is a richer language than RDF for formalising schemas or ontologies. Using these standards, the semantic web is being realised. The aim of the semantic web is to ultimately enable the location and integration of information on demand and without human intervention (Horrocks, 2008). Ontologies enable this by removing the problem of naming heterogeneity using terminologies and improving semantic interoperability by recording rich relationships between standardised named concepts. The main structures of ontologies are described next.

3.1.3.1 Basic Formal Ontology

The basic formal ontology (BFO) is an *upper-level ontology* (Smith, Kumar and Bittner, 2005). Upper level ontologies are special classes of ontologies that are formal and domain neutral. The BFO was designed for supporting information retrieval and the integration of information between domains. Here a domain is a portion of reality that forms the subject matter of a single science or technology area. BFOs are used to support the creation of lower level ontologies and formal (logical) reasoning.

[30] http://data.geohive.ie

Ontologists define BFOs, whereas domain specialists define lower level ontologies, using a BFO, and typically with the support of an ontologist. There are many other examples of upper level ontologies such as DOLCE (Masolo et al., 2003) and SUMO (Pease et al., 2002).

3.1.3.2 BFO Entities

There exist two types of BFO entities (or particulars), *occurrents* and *continuants*, which are the central organising axis of the BFO (Figure 3.3).

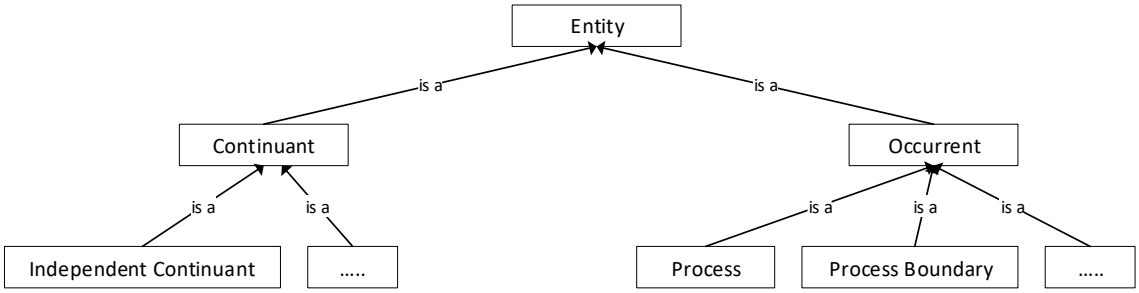


Figure 3.3 Simplified view of BFO entities and relationships.

Continuant entities are defined by the fact they can be sliced into parts only along the spatial dimension (and not the temporal dimension). Occurrents on the other hand can be sliced along any spatial and temporal dimension, again to give parts.

Beale (2002) notes that "in more complex domains, domain concepts fall into identifiable levels of abstraction". Upper level ontologies such as BFO provide a basis to define the principle level concepts within a domain and populate downwards through extensions of BFO.

These ontological levels within domains can be used to further structure information within complex domains such as health (Beale, 2002). This principle is also can also be said to be true for the geo-spatial domain, or indeed any similarly complex domain (Beale 2019, personal communication, August 15th, 2019). The use of ontological levels to structure information within complex domains is discussed in more detail later in section

3.4. The development of lower level ontologies and semantic system implementation technologies are discussed first.

3.1.3.3 Lower Level Ontologies

Upper level or foundational ontologies provide a basic structure for the formation of lower level ontologies. Developing lower level ontologies against pre-existing upper level ontologies increases interoperability against different ontologies. As lower level ontologies share the same high-level parental concepts this enables these ontologies to be merged using a process known as ontological alignment (described in chapter 2, section 2.4.2).

Ontologies are in fact categorised in additional levels such as middle and lowest level ontologies which have increasing specificity of concepts as they move below the upper level to the lower level. For example, an upper-level concept *event* can be further specified towards the geospatial domain as *observation* by adding further specifications or constraint definitions. This increased specificity represents an increased relation to an associated knowledge domain. Here only upper and lower have been considered to illustrate the general concept of levels within ontologies.

Lower-level ontologies tend to exist at the domain level, where upper level ontologies are more conceptual and do not lend themselves well to concrete concept creation (instances) in real world applications.

3.1.3.4 Recording Knowledge Bases

An ontology together with a set of individual instances of classes constitutes a *knowledge base*. Ontologies are used to aid the automatic processing and sharing of knowledge. This implies they need to be machine readable. To be understood and processed by a computer, ontologies need to be formally defined and represented in a machine-readable format. Many languages have been devised to formalise ontologies. OWL has already been

mentioned above. OWL provides a way to formalise knowledge in a machine-readable format. Typically, ontologies use classes to describe concepts in a domain. *Individuals* are the lowest level of granularity represented in a knowledge base.

3.1.4 Modelling Challenges

In any modelling scenario, variability is to be expected. Variability in a model allows differing opinions and viewpoints to be represented. Good models organise commonalities together.

Domain modelling by its nature will never likely to end. However, to realise technical systems the modelling must end before the system can be built. Consequently, most models are in-adequate, and their resultant systems are also inadequate for their particular application domain. To illustrate this let us consider the process of UML modelling.

3.1.4.1 UML

The Unified Modelling Language (UML, 2001) is commonly used in software development. Typically, a concept may be represented as a shape, such as a rectangle, with the concept labelled within the shape. Relationships or linkages between concepts are typically formed with a line drawn between concepts and a label or phrase that captures the nature of the relationship. Discovering and documenting relationships in a visual model requires modellers who are typically themselves non-domain experts, to ask questions of experienced stakeholders and develop a deeper knowledge of the subject which is the focus of the model. The conceptual model over time begins to visually document the knowledge available on the subject matter under investigation.

3.1.4.2 Domains & Idiomatic Expression

An eternal problem within software design and realising usable systems for specific application domains is the communication between the programmer and the customer (Fowler, 2010). It is well recognised that a core reason for failed software projects is the

inability to translate customer requirements into useful software. The reason is that all domain experts use *idioms* and *idiomatic expression*s to *talk* about their specialist area. For example, within a marine context the phrase “at the helm” implies in control of a ship but used outside the marine domain has a much more general meaning. For the most part software systems are written in a generic high-level, non-domain specific language. Therefore, a programmer’s job is to ultimately translate a heavily idiomatic description of some business logic into a generic language such as Java or C++.

If a domain expert can read and understand the code that drives key parts of their domain tasks, they can typically communicate in much more detail exactly what code needs to be written (Fowler, 2010). For that reason, *domain specific languages* (DSL) such as Gradle (Dockter et al., 2017) and OpenGL (Woo et al., 1999) have emerged. DSLs allow idioms to be used to express solutions of the problem domain.

UML and domain specific languages help in minimising this miscommunication, but there are still many challenges. Ultimately the ideal situation would be to allow domain experts themselves to define the systems they need, without having to rely on a translator (informatician or programmer) or without having to have a degree in computing.

3.1.5 Terminologies

Whereas ontologies formalise the concepts and their relationships within a knowledge domain, making domain assumptions explicit; terminologies by themselves represent a controlled vocabulary. Terminologies can be considered as preliminary attempts to model a domain’s knowledge (Zemmouchi-Ghomari and Ghomari, 2012). Within the knowledge engineering community, the distinction between what constitutes a terminology and an ontology remains debateable. The discourse tends to focus on the definition of a *concept*.

The semantic triangle is often used by both terminologists and ontologists to define concepts (Ogden and Richards, 1923). However, in more recent times the literature shows differing views of what constitutes a concept versus a term (Cointet and Chavalaris, 2008) (Gillam et al., 2005). The ISO technical committee 037 “Language and Terminology” publishes numerous standards relating to terminologies[31]. The ISO definition of terminology is a "set of designations belonging to one special language". Designations are further defined to be a "representation of a concept by a sign which denotes it" (ISO/TC037, 2000).

Both terminologies and ontologies require relations of concepts. However, terminologies are more limited in their relationship types than ontologies. Terminologies have a different focus in terms of function to that of ontologies. Terminologies support (among other activities): integration of information, indexing, messaging between systems (Rector, 1999). On the other hand, ontologies support: the retrieval and integration of information from different sources (Staab et al., 2000) as well as providing the prerequisite knowledge for query writing and machine-based reasoning (Bodner and Song, 1996).

3.1.6 Model-of-Reality Versus Model-of-Recording

As noted previously, ontologies are typically models of reality. However, systems require models that will inevitably have different types or categories of semantic meaning. For example, some models may define types that are quantitative in nature, whereas others will define a content model to capture information. For example, to enable the creation of structured-yet-flexible and computable documentation. These models are of different categories and must be developed and maintained separately (Beale et al., 2006). This separation is highlighted in Figure 3.4 below.

[31] https://isotc.iso.org/livelink/livelink?func=ll&objId=8864700&objAction=browse&viewType=1

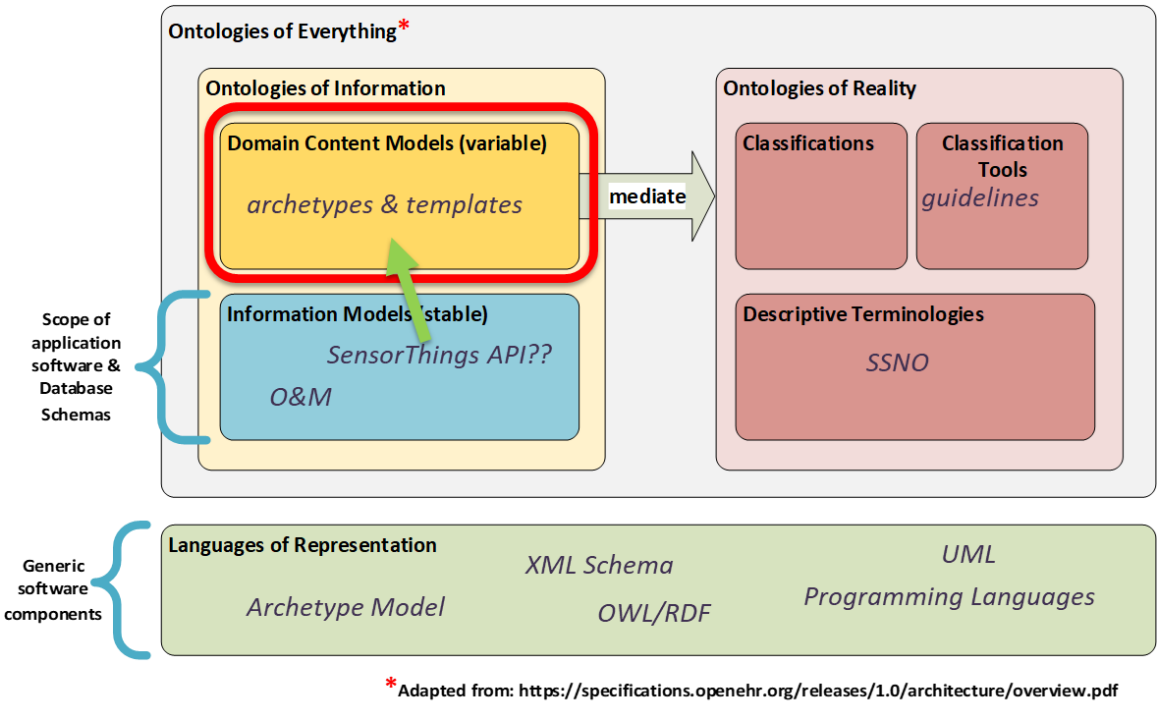


Figure 3.4 The ontological landscape (Beale et al., 2006)

Figure 3.4 highlights the need for not just models of reality but also models of information about *things* or *ideas*, these are also referred to as models of documentation. When developing models of recording of documentation, deciding what entities are valid topics for documentation can be challenging, especially when modelling documentation of ideas, or Popper's world 3 entities.

What is evident from the discussion thus far is that the development of information models that ensure accurate and useable data, information and knowledge formalisms is difficult. For this reason, there are many tools to aid semantic systems development. A brief review of semantic systems and tools is provided next.

3.2 Semantic Systems & Tools

Semantic systems use ontologies to aid integration of heterogeneous datasets. Semantic systems seek to help exploit data and information within systems by enabling semantic search. Semantic search can uncover hidden knowledge. The Semantic Web is an example of a semantic system. In the semantic Web, content is described in a meaningful

way. Meaning is provided by ontologies. Typically, the development of semantic systems is overly complex for casual users, as non-ontological expert users struggle with the formal logic of semantics (Bernstein and Kaufmann, 2006). However, there are many advanced tools to aid the development of semantic systems. For example, for ontology development Protégé is a commonly used tool. At a systems level Apache JENA (Apache, 2010) and Sesame (Broekstra, 2002) provide a rich framework of tools to help realise full semantic Web systems. Many frameworks will include reasoners such as the Pellet OWL-DL (Sirin et al., 2007).

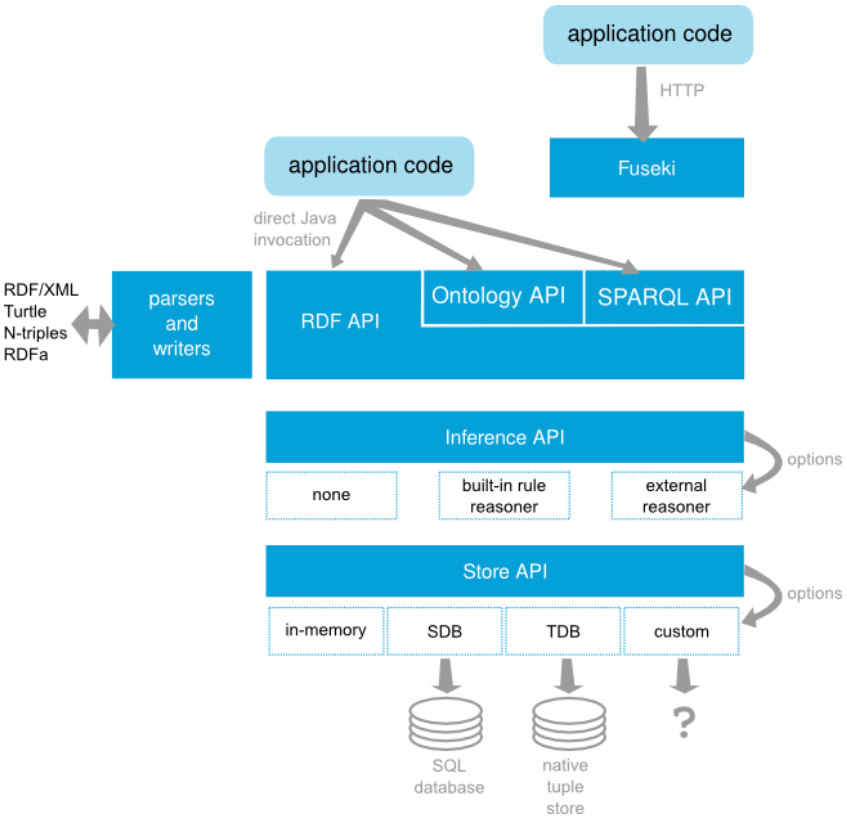


Figure 3.5 Apache Jena Framework Architecture (Apache Jena, 2010)

Apache Jena provides several interfaces for application code, namely: RDF API, Ontology API, SPARQL API and Fuseki (Figure 3.5). The Ontology API supports OWL (Apache Jena, 2010). Where RDF and RDFs are not descriptive enough for the

application area, OWL can be used (Allemang and Hendler, 2011). Ontologies are advantageous over database schemas as they are *explicit* and *first class* (see section 3.1.2).

Jena's framework is primarily for RDF, Ontologies are dealt with in this context and limited to formalisms on top of RDF. Jena takes the view that OWL is RDF centric and treats RDF triples as the core of the OWL formalism. This suits the approach ultimately used within this work. The Jena Ontology API is language neutral so RDFS or OWL could be used to describe an ontology. To represent the differences between the various representations each ontology language has a profile, which lists the permitted constructs and the names of the classes and properties.

Apache Jena provides a Java API to create, append and traverse RDF models. The statement interface provides methods to access subject-predicate-object elements of a statement within an overall model.

While frameworks such as Apache JENA provide rich tools to implement semantic systems the process of developing ontologies and semantic models is separate to application implementation.

There are many tools that can support the development of ontologies such as the popular tool *Protégé* however they are not particular relevant to this discussion. For the interested reader, Noy and McGuinness (2001) provide a very useful and highly cited practical introductory guide to ontology development using Protégé. Although quite old now the guide is still very useful for gaining a good understanding of the basics of ontology development.

Semantic information systems development using ontologies has advanced over the past 20 years. More recently, these advancements have begun to be adopted within data collection systems and specifically geo-observational based systems. Relevant to this work is a relatively new concept, where data can be *born semantic*. Born semantic has

been proposed as a semantic Web analogue to the idea of data being "born digital" (Leadbetter and Fredericks, 2014). Within the born semantic concept, data are captured digitally and at a point close to the time of creation, annotated with markup terms from semantic web resources (controlled vocabularies, thesauri, or ontologies). For example, a born semantic approach to air quality monitoring could require NO_2 measurements to include metadata which links the measured value to a standardised ontological concept definition of nitrogen dioxide. This allows heterogeneous data to be more easily ingested and amalgamated in near real-time due to the standard's compliant annotation of the data. The born semantic concept captures succinctly the requirement of observational systems to mark-up data at the very edge of spatial data infrastructures in order to avoid problems such as conflation that were described in the research problem statement in chapter 1 (section 1.2).

To date, it has been proposed that born semantic systems can be realised using technologies that support linked data approaches (Leadbetter and Fredericks, 2014). The linked data approach and enabling technologies are reviewed in the next section.

3.2.1 Linked Data

Linked Data is an approach for exposing, sharing and connecting structured data using URIs and RDF (Bizer, Heath and Berners-Lee, 2009). Linked data patterns have been used to demonstrate the Linked Data Ocean concept (Leadbetter et al., 2016). Linked data allows data fragments to exist across physical infrastructures while still maintaining their relationships. As will be seen in later chapters, the linked data paradigm has been used in this work to meet one of the core research objectives (research objective 5, see section 1.5.5).

The core principles of Linked Data provide the basic recipe for connecting data using Web technologies. In section 3.1 the concept of structured data was introduced. Structured

data (as opposed to unstructured, discussed) refers to data with a high level of organization, such as information residing within a relational database. Structured data markup is a text-based organization of data that is included in a file served from the Web. Linked Data techniques use the generic graph-data model of RDF to structure and link data within a Linked Data approach. Based on linked open data automated reasoners can be used to infer new information or to check logical data for consistency.

Linked data patterns are typically supported using RDF, which are XML based syntax. XML is a powerful language for defining rules for the encoding of documents with a mature set of development tools and established development communities. However, XML is generally not suited to constrained observational systems, due to its verbosity and the complexity of XML parsers (Castellani et al., 2011), which are key to XML's power and success. Conversely, the JavaScript Object Notation (JSON) is a simple standard for the exchange of hierarchically structured JavaScript objects.

JSON parsing is more efficient than XML and results in smaller exchange and parsing overhead (Nurseitov et al., 2009), which in turn does make it more suitable to constrained systems than XML. JSON has a several extensions such as JSON-LD (W3C, 2014). JSON-LD is a standard designed to serialize RDF using JSON. JSON-LD is a concrete RDF syntax, and so a JSON-LD document is both an RDF document and a JSON document and correspondingly represents an instance of an RDF data model.

3.2.1.1 RDF and OWL

As discussed in section 3.1.2, reification enables something to become a first-class-citizen, by providing a reification vocabulary. RDF is used to make statements about triples. An RDF document is a serialisation of an RDF graph into a concrete syntax, which provides the container for a graph. The RDF data model is composed of atomic data entities referred to as semantic triples (Klyne and Graham, 2006). A triple is composed

of three nodes within the RDF graph and codifies a statement about semantic data. Triples of this type are the basis for representing machine-readable knowledge. An RDF graph can be visualised as a node and directed-arc diagram in which each triple is represented as a node-arc-node link (Subject - Predicate - Object). RDF creates a graph structure to represent data. Serializations of RDF such as JSON-LD allow the markup of data instances using a structured data graph. RDF does not describe how the graph structure should be used.

The RDF model is based on the *node-arc-node pattern*, referred to as a statement (Klyne and Graham, 2006). Within a statement there exists three components, the Subject which refers to the node the structure is about, the predicate which is the label pertaining to the arc between nodes and the object. Statements are also called triples due to the three components that exist. An RDF model then is a set of statements.

RDF schema (RDFs) is a schema language that allows information modellers to express the meaning of the RDF graph data (Klyne and Graham, 2006). RDF and its schema extension RDFs provide support for distributed information and can be used to realize data instance fragmentation described later. However, RDF & RDFs do not provide the same semantic modelling as OWL. The Ontology Web Language provides additional vocabulary and semantic formalisms to RDF/RDFs. For example OWL provides the *owl:Restriction* construct.

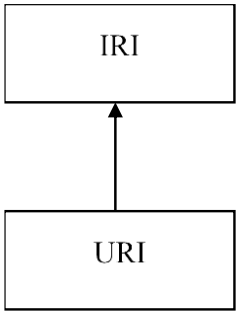


Figure 3.6 UML representation showing the relationship between URIs and IRIs. IRI is a superset of URI. The main difference being is that URIs are limited to using US-ASCII to encode characters, whereas IRIs are extended to use the Universal Coded Character Set[32].

The Web Ontology Language (OWL) builds on RDF and RDFS. OWL provides:

- OWL Lite
- OWL DL (Description Logic used for reasoners)
- OWL Full (has no guarantees on computation because it allows the full syntactic freedom of RDF)

OWL uses both URIs[33] and IRIs[34] (Figure 3.6) for naming and the description framework for the Web provided by RDF to add the following capabilities to ontologies:

- Ability to be distributed across many systems.
- Scalability to Web needs.
- Compatibility with Web standards for accessibility and internationalisation.
- Open and extensibility.

[32] https://fusion.cs.uni-jena.de/fusion/blog/2016/11/18/iri-uri-url-urn-and-their-differences/
[33] https://tools.ietf.org/html/rfc3986
[34] https://tools.ietf.org/html/rfc3987

These attributes of OWL make it a very relevant technology for research presented in this work. An OWL ontology consists of a collection of facts, axioms and annotations defined in terms of RDF graphs and triples.

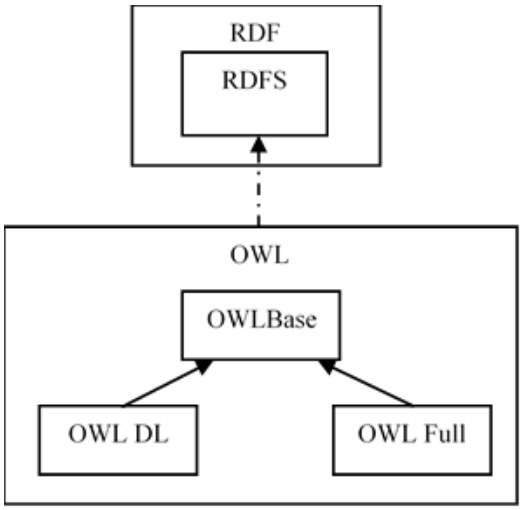


Figure 3.7 UML representation showing the inheritance relationship between RDF, RDFS & OWL

In OWL, classes provide an abstraction mechanism for grouping resources with similar characteristics. Like RDF classes, every OWL class can be associated with a set of individuals or "class extensions".

Boldt et al. (2015) describe a linked data approach built on top of the WiseLib store and show how SPARQL queries can be enabled on wireless sensor networks. Loseto et al. (2016) present a linked data platform to CoAP mapping due to the fact that only a HTTP mapping is provided for within the W3C recommendations. Charpenay, Käbisch, and Kosch (2017) describe a uRDF store for embedded devices as small as 8K that supports basic graph patterns, data are serialised using EXI to reduce data size. Le Phuc et al. (2016) describe the *graph of things* and through experimentation shows the impressive scalability of linked data, graph and semanticWeb approaches to managing connected physical device's datasets.

Le-Tuan et al. (2018) based in the INSIGHT centre in Galway, IRELAND propose the RDF4LED lightweight RDF engine which when compared against Jena's TDB requires 30% memory. Dell'Aglio et al. (2019) note an increased interest in *stream reasoning* research where micro RDF stores are increasingly being pushed to the edge of resource constrained networks.

Horsburgh et al. (2019) describe a 3-layer architecture (storage layer, Web framework layer and interface layer) of a data sharing portal based on the ODM2 (Observations Data Model) standard (ODM2 is discussed in more later in this chapter). The framework uses a *Restful approach* to sensing platform reporting (Fielding and Taylor, 2002). As the framework is based on ODM2 it also inherits the rigidity of ODM2. The framework is also based on HTTP interactions and so is more costly in terms of constrained system deployments.

Zárate et al. (2019) briefly describe the initial research work towards realising *OceanGraph;* highlighting the general trend and acknowledgement of the potential of knowledge graphs within the ocean observing domain.

Kaed and Boujonnier (2017) describe FOrTÉ, a federated ontology query database that uses SPARQL as the basis for federated queries within an IoT environment.

Barik et al. (2018) describe MistGIS, a geospatial data analysis solution enabled by way of a mist computing framework.

Leadbetter, A., Meaney, W., Tray, E. et al. (2020) describe an interoperable modular cataloguing service that employs a "findability" mechanism and improves discoverability of data.

3.2.1.2 Graph Databases

The linked data concept does not mandate a particular storage solution for the data that are linked. However, one of the more common approaches is to use a graph database (De

Abreu, 2013) (Wang and Chen, 2020). Graph databases come in several variants; the most popular variant is the *property graph*. A property graph contains nodes and relationships. Nodes contain properties (key-value pairs) (Robinson, 2013).

Graph database management systems expose graph models and allow CRUD operations to be performed on the graph. Graph databases may store graphs as native graphs, whereas others ultimately store the graph in a traditional format such as a relational-tables. As such graph databases can be categorised into native and non-native systems. Native systems (graph first) tend to perform queries faster and have better scalability.

The choice between native and non-native graph databases ultimately comes down to what the primary focus for optimisation is, this is discussed later in chapter 5.

3.3 Interoperability Challenges

At this point in the discussion it is becoming evident that cross-community sharing of computable information is difficult to achieve in practice. Barriers to interoperability within Earth system science informatics and SDIs means that ESS domain specialists cannot fully exploit the data that may be available. These interoperability challenges are complex, but now more than ever Scientists need to collaborate across conventional disciplinary boundaries. To enable this, they must be able to "first discover and extract data dispersed across many different sources and in many different formats" (Zhao, 2020). Interoperability challenges compound the problem of vast data silos referred to in Gore's vision of a Digital Earth system introduced in Chapter 1.

The challenges of interoperability are well documented and form core elements of many research agendas, including Geographical Information Science (Yuan et al., 2005). Much of the work done to date within the Information Science community has been to

enable interoperability through standardisation, particularly at the syntactic level. However, Goodchild argues that:

> *"Standards have the effect of codifying and constraining, whereas geographic information is evolving rapidly, demanding a much more flexible approach to metadata that reflects changing needs and expanding context."* (Goodchild, 2006)

Goodchild's statement is valid for all complex and evolving domains, where domain concept models also need to reflect that evolution; and traditional metadata modelling techniques are employed. Grossner et al. (2008) refer to this system evolution requirement as extensibility. Extensibility is an essential component for a Digital Earth system. Other essential components listed in the context of a Digital Earth system are semantically and ontological bound data models, and object-level metadata. Object-level metadata refers to the need to distinguish and manage observational data and derived knowledge. For example, this could be in the form of an associated scientific narrative, annotated onto a data object.

Solutions to some of these challenges are beginning to emerge. Standards such as the Open Geospatial Consortium's (OGC) Observations & Measurements (O&M) standard (Cox, 2006) (ISO-TC/211, 2011) enable syntactic interoperability between heterogeneous systems. Semantic interoperability, where the true meaning of the information reported from geo-observational data systems is an active area of research. Semantic integration goes beyond combining associated data points solely based on a syntactic representation. Semantic data approaches record the meaning of data points in some way (typically by refencing an ontology) along with the actual recorded data. This enables enhanced data integration based on meaning, where previously only syntax matching approaches were used. The linking of instance data that adheres to a standard data model (such as O&M) to ontological concepts and terminologies is now enabling

semantic interoperability (Wölger et al., 2011) (Leadbetter et al., 2016). Also, standardised vocabularies such as SeaDataNet (Schaap, Lowery et al., 2010) and NERC vocabulary servers (Leadbetter, Lowry and Clements, 2012) are all helping to realise the Digital Earth vision through semantic data methodologies. However, the extensibility of these approaches is often limited. The problem of unrecorded knowledge still persists as these approaches are ordinarily not flexible enough to be applicable in a large and diverse domain. Typically, domain concepts have been constrained early in the design process, leading to this inflexibility.

3.3.1 Standardisation

Lack of standards within the environmental sciences and information infrastructures is often cited as one of the main challenges to achieving collaborative environmental science information and research infrastructures (de la Hidalga et al., 2020). Mature international standardisation processes and organisations exist at the national (e.g. national standards of Ireland[35] (NSAI)), European level (European Committee for Standardisation[36] (CEN)) and the international level (International Standards Organisation[37] (ISO)). Developing International standards is a slow and complex process. Often technologies advance at a much faster pace than bodies such as the ISO can operate at. For that reason, many domains, such as the geospatial domain have established their own standards bodies to inform international standards development. Often after these more specific standards bodies develops and recommends a standard they may become adopted at the ISO level some time (possibly year) afterwards. One of the main standards bodies within the geospatial domain is the Open Geospatial Consortium (OGC).

[35] https://www.nsai.ie/
[36] https://www.cen.eu/Pages/default.aspx
[37] https://www.iso.org/

The Open Geo-Spatial Consortium is a voluntary standardisation body concerned with defining and implementation open standards for GIS data processing and data sharing. The OGC maintains over 30 standards. The SensorWeb Enablement Framework (SWE) is one of the main suites of standards developed and maintained by the OGC (Botts et al., 2008). Standardisation of interfaces (such as those defined in the SWE) addresses interoperability in sensor systems to a certain degree. However, standardisation of interfaces for the sharing of data does not address the incompatibilities between the actual data and concepts that are being shared. For example, the OGCs SWE (discussed in section 2.4.7) provides a syntactic solution to interoperability between heterogeneous sensor systems. The SWE framework on its own does not allow for semantic annotations. Work is ongoing to address the challenge of semantic interoperability in sensor networks.

> *"A semantic sensor network requires declarative specifications of sensing devices, the network, services, and the domain and its relation to the observations and measurements of the sensors and services."* (Compton, Henson et al., 2009)

The SensorWeb is a framework that allows management & access to real-time heterogeneous datasets. The SensorWeb is a type of Sensor Network. However, SensorWebs are inherently different to sensor networks or a distributed set of communicating sensors. The goal of the SensorWeb is to extract and distribute Knowledge. Nodes or pods operating in a SensorWeb can modify behaviour based on data collected by other SensorWebs.

The geographic information Observations & Measurements (O&M) standard is one of the many standards developed by the OGC as part of the SWE framework. All SWE based standards are aimed at enabling the sensorWeb. More specifically, the O&M standard defines a conceptual schema for observations. Features involved in sampling when making observations are also captured among other elements. The O&M standard was

subsequently adopted as an ISO standard (ISO 19156) and is a good example of how bodies like the OGC contribute to international standards development. But as mentioned above this process can be slow and typically contains many complex stages before final publication of a standard as an ISO standard[38].

Standards are about arriving at a shared view of the world by a diverse set of stakeholders. The ENVIR Community (de la Hidalga et al., 2020) provides a good example of how diverse stakeholders come together to agree standards. ENVIR was established to develop shared environmental research communities. The goal of the ENVIR community is to enable the multidisciplinary Earth system science through the development of standardised and interoperable research infrastructures[39].

The ENVIR community has produced a complex mapping of all their stakeholders; the mapping illustrates the complex interactions that need to take place within their community standardisation process (see Figure 3.8 below). These complex interactions are typical of any large standardisation community. The ultimate output of the community information-based standardisation process is to agree on some shared information model.

[38] https://www.iso.org/stages-and-resources-for-standards-development.html
[39] https://envri.eu/

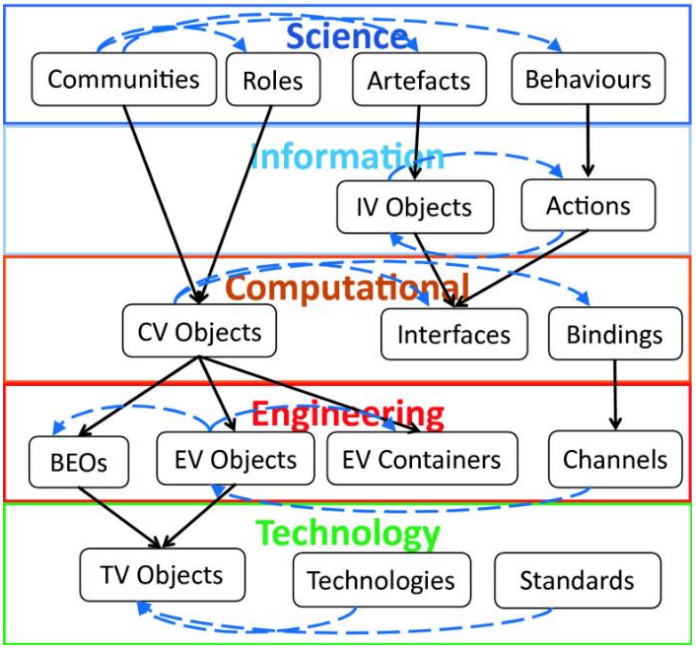


Figure 3.8 Shown are the 5 viewpoints specifications used by the ENVIR RM for stakeholders including correspondences that need to be maintained by all 5 viewpoints to ensure complex environmental systems maintain consistency between viewpoints (de la Hidalga et al., 2020).

Many information-based standards are represented as object-oriented information/data models. The ISO (and OGC) typically publish these standards using UML representations. SDIs and research infrastructures such as the ENVIR Community's infrastructure adopt and implement these standards and also feedback to standards bodies through pilots and submissions updating and evolving the UML based standard. However, UML and object-oriented techniques have been shown to be problematic when applied to complex domains, this is discussed in more detail later in the section 3.4.

3.3.2 Semantics in Resource Constrained Systems

Semantic information at the sensor-data level can have many benefits such as allowing direct interaction between heterogeneous sensor nodes (Hayes et al., 2009) Another reason to push the data processing to the edge of sensor networks is that most work done on the Semantic Sensor Web assumes a centralised approach. Terminology is centralised and inference steps are then carried out on this centralised system too. This approach has scalability issues if the predictions as to the growth of nodes/devices/entities participating

in the semantic sensor web become a reality. De et al. (2014) propose an interesting, federated framework of nodes for the Internet of Things. The framework focuses on two aspects: “inferring automated associations that integrate the nodes digital components with physical entities and a notification algorithm to share knowledge between a determined set of nearby nodes. Larizgoitia et al. (2010) presents an architecture for WSN nodes to integrate to context-aware systems using semantic messages. The expressed goal of the research is that “the information has to be semantically defined from the very moment it leaves the sensor node”.

Semantically annotating captured data at source is problematic. Typically, OWL or RDF is used to add semantics to sensor data. Both of these mechanisms are computationally expensive and, in a resource-constrained environment this may not be possible.

Using XML at node level - up to now - has been for the most part impossible. It has been noted (Chapter 3) that triples are the base of the entire RDF knowledge model. Triples can be represented using many different formats. But none of these formats are suitable for a sensor networks due to computational constraints and limitations on packet size etc.

Again, Larizgoitia et al. (2010) propose a solution to this through an adapted representation of triples that would be suitable for a wireless sensor environment. Compression or codification mechanisms are needed. Each part of the triple will be represented as a URI; however, URI lengths are in general too long for packets in a WSN directly. Codification of the URIs are proposed Code every single term in the ontology.

There are several notable examples of supporting linked data principles on constrained devices that are relevant to this work. For example, Hasemann et al. (2012) developed WiseLib which is a lightweight *tuple* store. Wiselib is part of the SPITFIRE architecture

and provides limited support for RDF on tiny devices. Hasemann et al. (2012) showed that Wiselib incurs some overhead in terms of processing power, memory and bandwidth but overall, the impact was relatively small.

3.4 Representing Complex Domain Knowledge

As with all complex and wide-ranging domains, knowledge construction and persistence are a difficult endeavour, even when it is confined within a specialised sub domain. Earth Science Informatics is an interdisciplinary field and represents a need to share not just data, but interdisciplinary knowledge in a computer process-able way; allowing the information to be trusted by the professional who seeks to use it. A GIS system that is solely based on facts cannot readily share inter-disciplinary knowledge. Examination of the development of Geographic Information into a super-discipline and among cross-sub-disciplines such as ESS illustrates the need for Informaticians to ensure that adequate frameworks are in place to allow domain experts, such as Geographers, to semantically enrich, and document, all generated information and knowledge.

Given how information science has evolved and knowledge engineering techniques and technologies have improved, it is worth examining whether the initial criticisms of GIS (Taylor 1990) have been addressed. As noted previously, the challenge for Earth Science Informaticians is, how to build systems that can represent knowledge within a large and diverse community such as Earth System Science; whilst ensuring that as the knowledge is shared and processed amongst the community, the context and true meaning of the knowledge is preserved.

Firstly, let us examine how information and domain concepts are captured within an information system. Geo-information has traditionally been modelled from a computer science perspective. Traditional relational databases have been the main choice for storing data in many information systems. Schemas of the data and relationships are captured

through the modelling of data. There are many approaches to data modelling. Database design has become strongly influenced by object-oriented techniques. However, Object Oriented techniques are considered too stringent during the early stages of knowledge acquisition (Boegl, Adlassnig et al., 2004).

In a domain such as Earth Systems Science, the representation of knowledge is difficult, as it is ever-changing and evolving (Goodchild, 2006). A means of modelling and thus enabling the recording of uncertainty is not readily possible. Traditional database design and indeed object-oriented approaches assume a static understanding of entities or classes of information. Therefore, these static design methodologies cannot represent the true nature of knowledge within an evolving domain. Over time the model becomes outdated.

Again, we can refer to GIS to understand the limitations of static models such as traditional OO models. Gahegan & Pike (2006) noted that one of the main problems within GIS is "The impoverished descriptions of data and other resources". Also highlighted by Gahegan & Pike (2006) was the problem of unrecorded knowledge, arising from scientific data analysis activities.

> *"Analysts explore complex and voluminous data resources, and combine them in various ways to synthesize new understanding. These activities both utilize and produce knowledge that for the most part remains unrecorded, residing only in the volatile memory of analyst(s)"* (Gahegan and Pike, 2006)

It may well be the case that these problems are symptomatic of the unsuitability of static data models underpinning GIS systems, or any ESS based information system. Four important challenges relating to the representation of geographical meaning were also identified in their work.

- The world is changing, so concepts must either adapt accordingly or become obsolete.
- We as individuals and groups are also constantly changing, so our needs, goals understanding and experience - i.e. our bases for constructing concepts - are also in flux.
- We use words or signs to stand for (encode) concepts, but there is no guarantee that concepts will be understood in the same way by all parties during communication.
- We need to keep track of the conceptual structures we construct and use since they are key to understanding our data and other outcomes.

(Gahegan and Pike, 2006)

These challenges highlight the difficulties associated with the representation of concepts and provides basis for constructing concepts that are constantly in-flux; along with the difficult task of maintaining a consistent understanding of concepts as they are communicated to different parties.

The practice of constraining knowledge at an early acquisition stage is inherent in object-oriented techniques (Boegl, Adlassnig et al., 2004) and leads to impoverished concept descriptions, unrecorded knowledge (Gahegan and Pike 2006) and creeping system obsolescence (Beale, 2002). Knowledge sharing can be maximised across an interdisciplinary super-domain (such as ESS) by empowering suitably-experienced domain specialists to model domain concepts themselves in a computable way, and by allowing for the evolution of the domain concepts within the model.

To date OO based information standards have been defined by large international bodies such as the International Standardisation Organisation (ISO). In the geospatial

domain the Open Geo-Spatial Consortium has been highly influential in the development of geo-informational data models and standards.

More recently, the Earth Science Informatics community has sought solutions to the goal of truly flexible and extensible semantic information systems. Notable projects are the European collaborative project CHARMe (Clifford, 2016) and the SMART-IWRM project (Wolf and Hötzl, 2011) (Kämpgen, 2014). These projects leverage existing standards with the ability to record community generated knowledge. The SMART-IWRM Knowledge Base is a good example of the state-of-the art in systems trying to achieve knowledge sharing between diverse communities of practitioners. Other relevant examples of extending or augmenting object-oriented based standards are the GeoViQua project (Masó et al., 2011), the WMS-Q profile (Blower, 2015) for the WMS OGC standard and ODM2 (Horsburgh et al., 2016). Riepl (2014) proposes a semanticWiki approach for collaborative knowledge generation and sharing. The semanticWiki approach has comparable goals to the approach described developed in this thesis. The systems listed above are reviewed in further detail in section 3.6.

3.4.1 Geospatial Domain

Integration of geo-spatial data requires clear disambiguation of the semantics of the information being consumed. The most basic semantics of temperature observations where the units are expressed within the data are often not included. As was discussed in chapter 2, meaningful geographical representation goes far beyond simply including the unit of measurement within the recorded observations. Ontologies can form part of the solution. However, ontologies are only part of the solution; far from being a “silver bullet”, ontologies by themselves solve only a part of the problem of succinctly representing and communication meaning of resources. Perhaps one could also argue that concentrating solely on ontological knowledge in GIScience might result in a worsening

of the problems described by Pickles and colleagues (Pickles, 1995), in the sense that more objectivity may tend to re-enforce the belief that resources can always be taken at face value.

3.4.2 Health Domain

In recognition of the relatively slow pace of evolution of data standards (section 3.3.1), particularly those that normalise information models, and the problem that coded terms or ontologies alone are not sufficient to achieve semantic interoperability; for over 20 years, health informaticians have been developing a highly sophisticated approach to information modelling, known as two-level modelling (Beale, 2002). Two level models are designed to facilitate large-scale sharing of high quality, multifaceted, flexible and durable documentation.

3.4.2.1 Clinical Information Modelling Initiatives

The Good European Health Record (GEHR) project ran from 1992 to 1994[40]. The aim of the GEHR project was to develop and test a common architecture for digital health records in Europe. The resulting architecture was reported by Ingram et al. (1995). The main results of GEHR were the definition of the requirements for clinical completeness within electronic health records and a first attempt to define a formal data architecture to meet those requirements, which constituted a static domain model for healthcare documentation.

Arising from the work of the GEHR project, two other EHR development projects began in the mid-90s to further the investigations of appropriate clinical information modelling and EHR systems development; Synapses (Grimson et al., 1996, 1998) and the GEHR Australia project (Heard and Beale, 1996).

[40] https://cordis.europa.eu/project/rcn/17093/factsheet/en

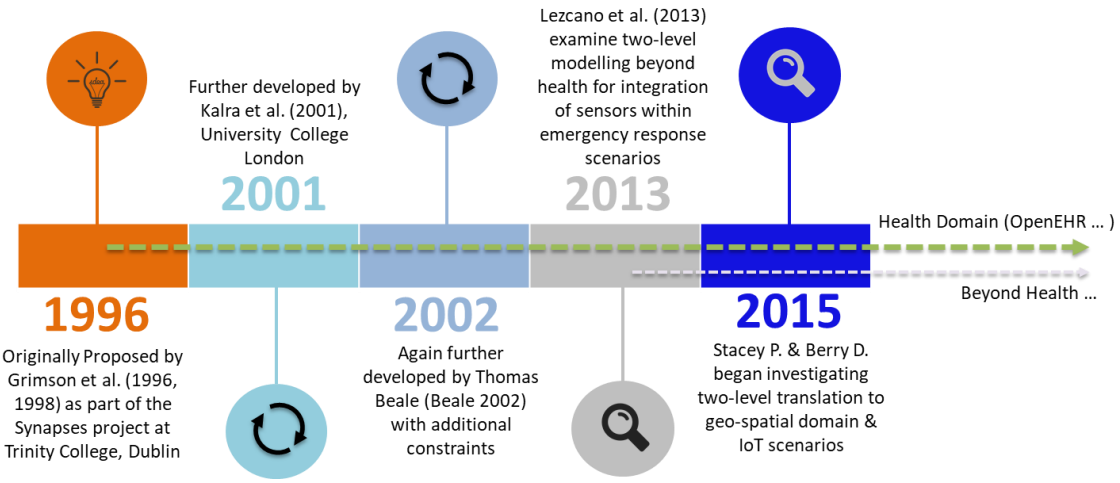


Figure 3.9 A brief blinkered history of two-level modelling relating to this work

Kalra (Kalra, 1997) notes that federation approach of clinical information requires two information formalisms to be specified. In Synapses a *synom* and a *synod* are defined. A synom is an abstract generic model and a Synod is an extensible metadata object dictionary which could be curated by domain experts to produce flexible and updateable definitions of parts of a clinical document. Together they can provide the required dual information formalisms.

Thomas Beale furthered the dual information formalism approach proposed by Synapses, adding additional constraints (Beale, 2002) and feature-rich constraint mechanisms. This more mature approach was described as *two-level modelling* and introduced the concept of *archetypes*.

The GEHR (Australia) was the precursor to what became known as the open EHR foundation (openEHR); whereas the Synapses' project can be credited with the first glimpse of what became known as two-level modelling, the feature introduced in the openEHR approach led to a fully implementable specification. Two-level modelling and archetypes are described in more detail next.

3.5 Two-Level Modelling

Traditional information systems design tightly-couples information and knowledge concepts. This coupling happens early in the design process, at the point where object and data models are developed. Beale (2002) refers to these type of design methodologies as "single-level" models. Beale argues that where the single-model approach is applied to information systems in a constantly changing environment, these systems become expensive and difficult to maintain. Beale also notes that these types of systems need to be replaced after several years. The reason for this is that domain concepts are hard-coded into the software. As the domain evolves and changes, the software becomes outdated and less useful. Single-level systems have also been shown to have limited interoperability, as they may not adhere to a standardised formal model. Beale postulates the core issue for creeping obsolescence in single-level information systems is the constant evolution of the knowledge in a domain (Beale, 2002). Flexible design methodologies are needed to keep up with the non-static nature of the domain.

Two-level modelling systems design approaches arose from the need to avoid the problems with single information architecture-based systems. In the two-level approach, a traditional object model is still developed. This is referred to as the "Reference Model" (RM) or first-level model. The second-level model is where the formalism of the domain knowledge is captured. The separation of domain concepts can be organised as follows (Beale, 2002):

- *1st level*: This is the informational level and contains what are described as the non-volatile concepts required to be modelled for the system. It is a reduced set of classes that have an abstract meaning, but nevertheless, have features to incorporate data types, terminology or ontology bindings. These concepts have

been carefully devised to be used as general but domain specific "building blocks" according to rules described in level 2.

- *2nd level*: This level is the knowledge level where the concepts that will undergo evolution over time are captured and can be bound to ontologies as required. These concepts are specialised from the non-volatile level one concepts, but are themselves volatile in nature and so they can evolve over time as knowledge evolves without “breaking” the system. This level is captured as a knowledge model using “archetypes” and an Archetype Model (AM).

The separation of (recorded or documented) information and (generally applicable) knowledge in information systems design allows a more flexible representation of the domain knowledge (e.g. as part of a separate ontology, section 3.1.3). In a two-level model the reference model contains features to allow individual ontological terms to be “bound” dynamically to any point in the information model, while keeping a rigorous formal definition of the data that are being recorded.

3.5.1 Benefits of Two-Level Modelling

Two-level modelling introduces additional complexity to the modelling of domain information models. However, once adopted within a domain, the benefits can be great. Outside of the perceived technical benefits of semantic search and versioned compositions, additional non-technical benefits occur, that of *domain empowerment* and *community consensus modelling*.

3.5.1.1 Domain Empowerment

One of the core principles of the two-level modelling approach is that it should enable domain practitioners to capture specific domain knowledge concepts and to manage them as they evolve over time. The 1st level, or reference model, is still developed by Informaticians. The 2nd level, or the knowledge level, is developed by a mixed group of

authors, that include the domain practitioners themselves and who now have greater influence on the evolution of models within a community environment (Beale, 2002).

3.5.1.2 Community Consensus Modelling

Community development of Archetypes is a complex task that is performed by domain specialists. Within health a sophisticated framework of tools has evolved over the past number of years to facilitate the development, management and evolution of domain specific Archetypes (Sundvall et al., 2008) (Maldonado, Moner et al., 2009) (Chen and Klein, 2007).

3.5.2 Reference Models

As discussed above reference models are stable structures that include generic information. Reference models in two-level modelling are hierarchical in nature, and typically (as found in openEHR and EN13606) minimally define the following constructs:

- *Folder*: a folder allows for the grouping of different compositions. Grouping is performed based on some common characteristic, usually decided by a clinical team (when used in the health domain).
- *Composition*: a composer creates what is termed of unit of committal for the information system. This may be a patient report or some other record. The composition information structures enable this recoding of the documentation of some clinical encounter within the health domain.
- *Section*: compositions contain sections. Sections are defined by some clinical heading such as family history. Sections can contain additional sections.

- *Entry*: an entry may be a singular clinical observation. It can also be defined in health as a clinical statement about some clinical action, such as reading of a patient's blood pressure.
- *Element*: elements are single data points or values, such as the diastolic pressure value of a patient's blood pressure.
- *Cluster*: a cluster organises individual elements in a nested data structure.

The Folder/Composition/Section/Entry/Cluster/Element multi-level object-oriented structuring evolved is accepted as a core part of the CEN and HL7 standards[41].This structure is an evolution of the original GEHR defined structure of Transaction/Headed_section/Entry/Compound/Item.

Today, OpenEHR defines a mature reference model for the health domain[42]. It is important to note that a reference model is not a singular model, but a collection of object-oriented models that cover the needs of the specific domain. Within openEHR the reference model has a number of formal specifications which each contain several specific models (Figure 3.10).

RM (Reference Model) PRs|CRs	**Demographic**: Party, Party_relationship, Actor, Role, Contact, Address	**EHR**: Composition, Section, Entry, Observation, Evaluation, Instruction, Action, Admin_entry	**EHR Extract**: OpenehrExtract, GenericExtract
	Common: Versioned_object, Version, Party_self, Audit_details	**Integration**: IntegrationEntry	
	Data Structures: History, Event, ItemTree, Cluster, Element	**Data Types**: DvBoolean, DvText, DvCodedText, DvUri, DvQuantity, DvDate/Time types, DvMultimedia	
	Support: Terminology and Measurement service interfaces		

Figure 3.10 Screenshot of current (2019) formal specifications available within the openEHR reference model[43]

[41] https://www.iso.org/obp/ui/#iso:std:iso:13606:-1:ed-1:v1:en
[42] https://specifications.openehr.org/releases/RM/latest/index
[43] https://specifications.openehr.org/

It can be seen in Figure 3.10 that some of the minimal constructs of a reference model listed above are contained within the *EHR formal specification* (composition, section, entry), while others are defined within the *data structures formal specification* (cluster, element).

Data structures within the reference model are defined using object-oriented models. Core to realising the multi-level object-oriented structuring is the adoption of a compound/element pattern within the reference model structures. Figure 3.11 shows a (portion of) object-oriented model depicting the main multi-level structures realised using the compound/element pattern (highlighted in green). Within the model below it can be seen that cluster and element both implement the abstract class ITEM. This modelling requirement within two-level modelling reference models is discussed in more detail in chapter 5. For now, the reference model pragmatics are only considered at a high level to illustrate the overall two-level modelling approach.

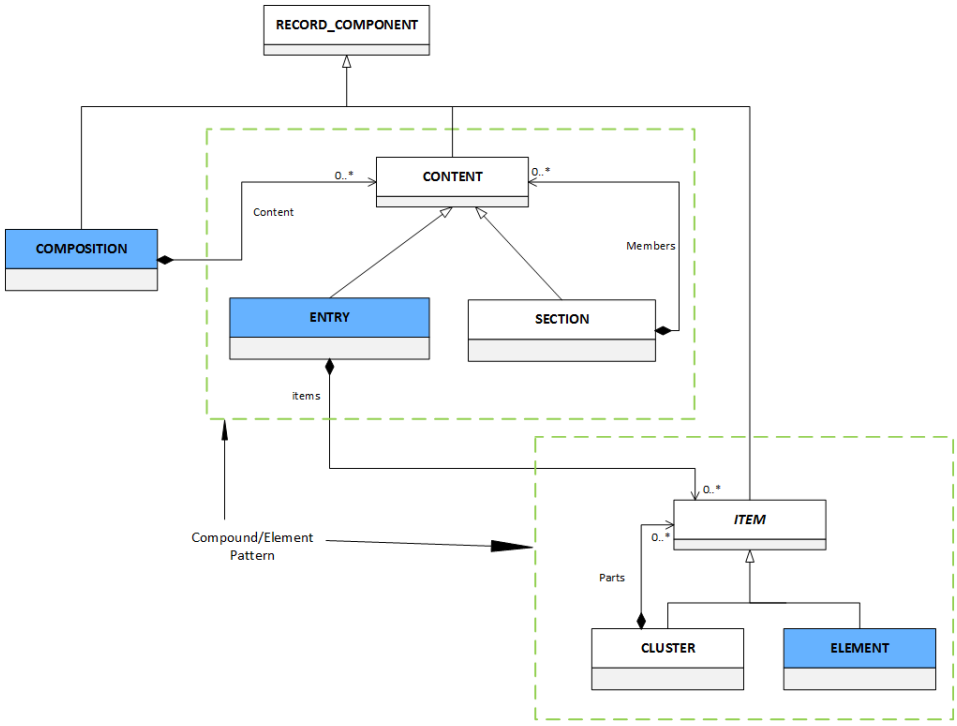


Figure 3.11 A portion of the EN 13606 Reference Model. Compound/element patterns are highlighted in green

The openEHR reference model defines classes beyond the organisational and also provides classes that aid interoperable communication between EHR systems such as:

- Audit information
- Functional roles
- Attestation information
- Related Parties
- Links
- Demographics

These additional constructs further improve the ability of heterogenous systems to communicate and share information in an interoperable way. Many of these classes are not relevant for domains other than health (e.g. attestation), but some may be reusable (e.g. related parties). The core requirement for reference model constructs are that they represent generic informational concepts that will persist and remain constant over time.

The second level within the two-level modelling methodology is not defined using an object-oriented approach. Second level concepts use archetypes, and archetype modelling to define their structures. An archetype is a programmatic definition of a concept, but their definition is normally submitted by domain experts in the form of a mind map. Mind maps represent a type of directed acyclic graph structure, but in a more simple and accessible way. Archetypes are described in more details below.

3.5.3 Archetypes

The capturing of non-volatile or stable concepts in the 1st level, or reference model, can be achieved using traditional conceptual modelling approaches. When a reference model has been developed, the challenge is then: how are the semantics of the reference model, or the knowledge concepts that have not been captured by the reference model to be

defined and implemented? Within the geo-spatial domain knowledge concepts would include scenario specific concepts.

Beale (2002) notes that knowledge level concepts are essentially constraints on the reference level concepts. As such, the knowledge level can be captured as a set of constraint statements. Here a set of constraint statements are referred to as an *archetype*.

The term archetype is generally defined as a universally understood symbol or term. In information systems design an archetype is a set of constraints on a reference model. These constraints provide semantic relationships between elements based on knowledge. Using archetypes, an archetype model can be developed that formalizes the volatile knowledge concepts within the 2nd level of a two-level based information system (Figure 3.12). Archetypes allow for the necessary variability employed by domain practitioners to be managed in an interoperable. In contrast to ontologies, archetypes are models of documentation.

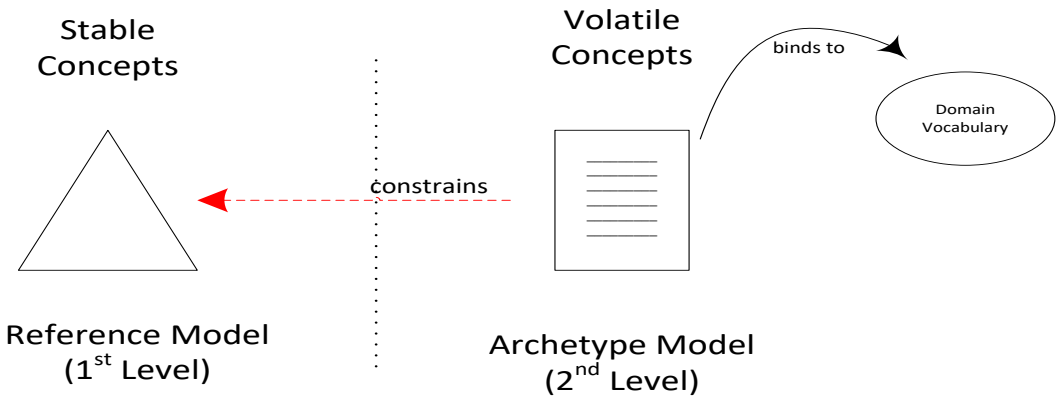


Figure 3.12 Two-Level Model separation of stable concepts from volatile domain concepts

3.5.3.1 Archetype Definition Language

A formal language Archetype Definition Language (ADL) (Beale, 2007) for defining archetypes exists and is maintained by the openEHR foundation (Kalra, Beale and Heard, 2005). ADL is used to constrain information models. ADL is used to constrain information models. It is best suited to information models that are very generic in nature.

As an example, where logical concepts PATIENT, DOCTOR and HOSPITAL would be represented by a smaller more generic number of classes such as PARTY and ADDRESS (Beale and Heard, 2007). ADL can then be used to constrain the instances of these generic classes to represent specific domain concepts. ADL was developed for the clinical domain. However, ADL can be used to define archetypes for any domain where there exists a formal object model (Beale and Heard, 2007).

ADL uses three other syntaxes, cADL, dADL and FOPL (Beale and Heard, 2007).

- cADL captures the Archetype definition
- dADL expresses the data which appears in the language, description, ontology and revision history.
- FOPL is used to describe constraints on data which are instances of an information model.

To illustrate the syntax of ADL an example of a very basic Archetype definition is presented below.

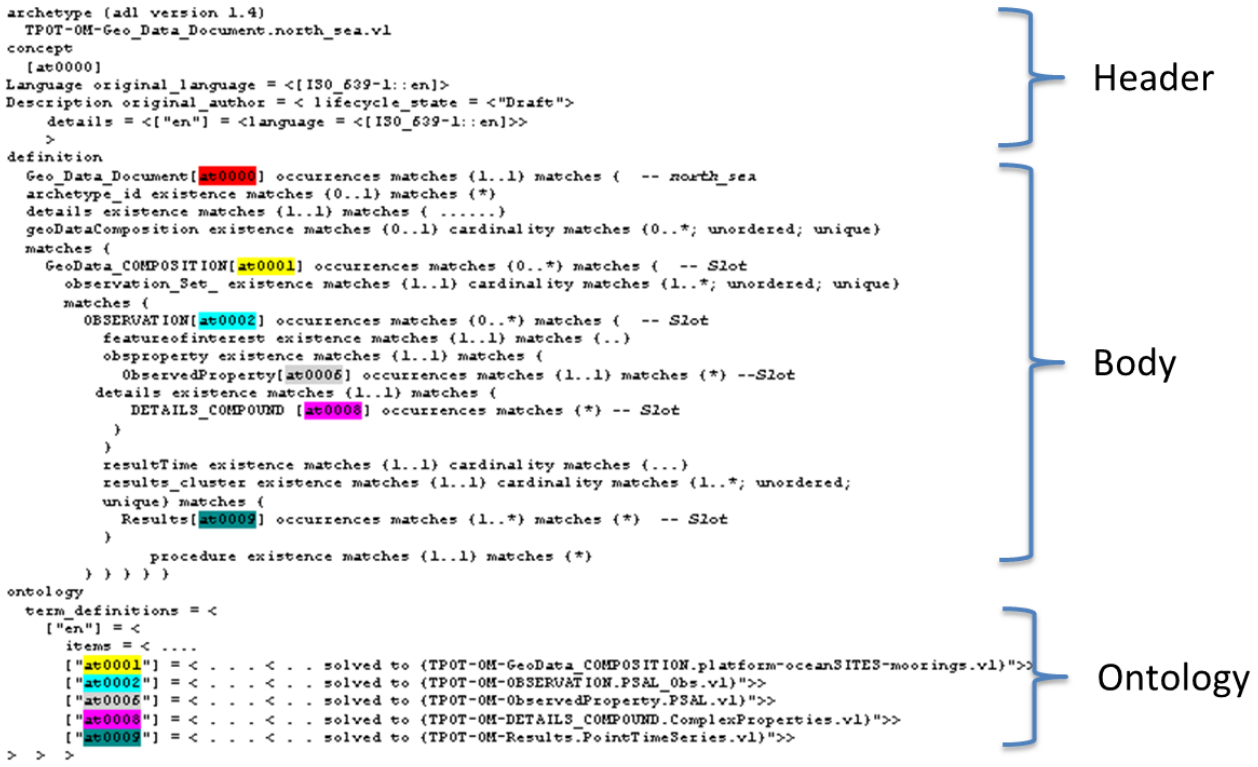


Figure 3.13 ADL example highlighting the three main sections

We can see from the above example an Archetype definition is composed of three main sections:

- Header
- Definition (Body)
- Ontology

The ontology section allows terminologies to be bound to concept definitions. For example, the concept at code `at0006` which provides a constraint definition of practical salinity can be bound to the NERC vocabulary code (Listing 3.1 below)

```
term_bindings = <
      ["NERC"] = <
            items = <
                  ["at0006"] = <[NERC::SDN:A05::EV_SALIN]> -- Salinity
>>>
```

Listing 3.1 Example of ADL term bindings to NERC vocabulary

openEHR provides numerous tools for working with Archetypes. ADL representations of Archetypes can be converted into numerous representation formats such as XML formats. openEHR publishes and maintains an XML-schema corresponding to the ADL Object Model.

ADL is not dependent on the reference information model but is best suited to information models that are very generic in nature, and so in principle can be used for ESS modelling.

3.5.4 Operational Templates

Archetypes are further specialised for use-cases and are combined to produce a set of Operational Templates (OPT). This ability to produce OPTs adhering to a rigorous formalism is a key advantage of two-level models. Operational templates offer additional flexibility outside of the community-agreed archetype model for local uses. This provides for situations where disparate domain expert groups may disagree and can lead to archetype alignment issues as the approach matures within the domain.

3.5.5 Two-level Modelling for Health Applications

There are several parallel attempts at two-level models in healthcare. OpenEHR and CEN/ISO 13606 (ISO/TC 215, 2006, 2008, 2009a,b), Clinical Element Model Specification by Intermountain Health (Oniki, 2014) and the Clinical Information Modelling Initiative (CIMI, 2020). These models embed the following data quality enhancing features.

- A strong recognition that the model is intended for documentation of phenomena, rather than for producing a general model of reality (Beale, 2003). The latter is the role of an ontology (Peirce, 1935). In the healthcare community, this is not considered to be the same as documentation. As mentioned previously, ontological information is “bound” into the information model, which focuses on

documentation.

- Use of commonly agreed identifiers and related mechanisms to uniquely identify phenomena that are the subject of documentation or provide context for the document.
- Use of an evolved temporal model and time-based data types to allow different aspects relating to time to be recorded accurately and satisfactorily.
- Embedded or “bound” ontological codes at appropriate points in the two-level model for referring to commonly agreed concepts and terms.
- Employment of a general and reusable reference model, composed of building block concepts that can be used in many different documentation scenarios. These models are quite similar in intent to the OGC’s O&M model. As previously noted, this reference model corresponds to the first of the "two levels". Figure 3.11 above shows a simplified EN 13606 reference model (Muñoz et al., 2011).
- Development of a consensus-based library of archetypes.
- Recursive aggregation patterns within their reference models.
- Strong data typing.

Two-level models and archetypes go beyond the idea of "recording measurements" to developing community-standardised "documentation" that is designed through consensus of the members within the community itself. The process of developing these archetypes is a slow one, but the benefits are worth the great effort.

Another comparable approach in the health domain is the SHARPn project (Rea et al., 2012). The SHARPn project also decouples use-case knowledge representation from underlying standardised structured electronic healthcare data. Clinical Element Models (CEM) allow for use-case knowledge level formalism, and terminology bindings. CEMs

are analogous to archetypes and are the basis for achieving semantic interoperability between systems.

3.5.5.1 Semantic Web and Clinical Information Models

Sharma et al. (2017) describes how the health informatics community has over the past few years initiated an international collaboration known as Clinical Information Modelling Initiative (CIMI) to provide a shared repository of detailed clinical information models based on a a common formalism. Formalisms such as CIMI allow for the normalisation of patient data for secondary re-use, a perspective that is also a key consideration in the Earth Sciences. Sharma et al. argue that clinical information tools can leverage semantic Web technologies to realise normalised detailed clinical models (DCM). Their paper presents an architecture of four layers. An RDF translation layer. An RDF store-based persistence layer. A semantic services layer and an authoring layer (archetypes). The work initially focused on the first RDF translation layer. The approach adopted was to take an XMI representation of a given reference model and convert it from XMI to RDF using the XML2RDF transformation service. A JAVA program was then created that produced OWL rendering on the CIMI reference model using UML2OWL mappings specified by the OMG ontology definition meta-model (ODM) standard. An OWL based schema for the CIMI reference model was produced.

3.5.6 Two-Level Modelling for non-Health Applications

The main goal of the two-level approach is that it acknowledges the reality of, and thus supports knowledge evolution within a given domain. These characteristics of an information system have wide applicability, especially within ESS informatics. Tavra et al. (2017) highlight the need for further research in how marine spatial data infrastructures (MSDI) design can reflect the “highly dynamic nature of the environment on which it is applied”. Their work to develop a planning support concept (PSC) framework for the

development of MSDIs is interesting in the context of this work. The proposed PSC is broken into phases. Also defined is a *bi-level* goal tree (Tavra et al., 2017), i.e. the goal tree also reflects the need to adhere to the European Spatial Infrastructure INSPIRE directive. However, additional considerations for systems interoperability and secondary information reuse are not explicit within the phases of the PSC. In general, the proposed PSC is reflective of the wider ESS informatics approach to data interoperability. Within the literature it is evident that there has been a greater emphasis on semantic interoperability within health information systems than ESS based SDIs and observational systems. This has largely been driven by public demand for better healthcare (Grimson et al., 2000). That same pressure to do better has arguably not existed to the same extent in ESS informatics. However, the need for systems that support knowledge evolution in multi-disciplinary ESS based domains is increasingly being acknowledged in future research agendas.

3.5.7 Challenges of Two-Level Modelling

Traditionally two-level modelling approaches have been the preserve of health Informaticians. As other domains place a greater emphasis on semantic interoperability and systems that support dynamic information and user needs, two-level modelling approaches are gaining attention outside of health. Lezcano et al. have shown how the semantic integration of sensor data with disaster management systems can be facilitated using a two-level modelling approach (Lezcano, Santos, Garcia-Barriocanl, 2003). Stacey and Berry (2015) and also Diviacco and Leadbetter (2017) have noted the potential benefits of a two-level modelling approach for geo-observational systems. While proposing a translation of the two-level modelling approach from health to other domains, it is necessary to be cognisant of the differences that exist.

Within healthcare informatics, the primary subject of documentation is the patient. The prevailing consensus within health informatics is that the patient should remain the dominant subject of documentation for shared electronic health care documentation. This is a primary difference between healthcare informatics and geomatics. The subjects of documentation in geo-information and documentation in Earth System Science are diverse.

The work of Diviacco et al. (2015) with *boundary objects* highlights this diversity, and further highlights the current efforts within the geo-sciences community to tackle automatic semantic and dynamic knowledge representation. Beaulieu et al. (2016) highlight the growing need and the current state-of-art in cyber-infrastructures to support collaborative processes and semantic communication amongst a diverse set of domain specialists. This automatic recording of information is much more prevalent in the geo-sciences. It could be argued that healthcare documentation is "a matter of life and death" for the subject of documentation. This is usually not the case (at least it is not immediately the case) in geospatial measurement and documentation. Patient safety and quality of care issues impose a strong need for rigour in healthcare, and a certain conservatism about changing processes and systems.

3.6 Discussion & Conclusion

While current emerging solutions such as SMART-IWRM (section 3.4) can bring the necessary flexibility for domain practitioners to share semantically rich heterogeneous ESS datasets, rigorous definitions of the additional use-case knowledge may be compromised. Particularly, use-case specific knowledge and understanding is not being provided for in an evolutionary, interoperable, and computable way. For example, the CHARMe project introduces a flexible approach to structuring geo-data. However, this flexibility makes information consumer applications such as Spatial Data Infrastructures

(SDIs) that automatically aggregate datasets difficult to achieve in practice. The lack of a truly rigorous formalism and flexible definition of the evolving use-case knowledge means that techniques for combining datasets for automatic processing or semantic search are not always optimum. Also, the ability to build useful inference-engines is limited. While the WMS-Q profile discussed in section 3.4 allows the annotation of datasets with quality information, it does not enforce conformance of data to a model and so it limits data validation services. WMS-Q describes the quality of data but does not in itself enhance the quality of the data at the source of capture.

Arguably, ODM2 (discussed in section 3.4) appears to be the most promising of these approaches. ODM2 has adopted its core concepts from O&M and added extension schemas. The extension schemas ensure that it can be applicable to a broad community of practitioners. Also, the extension mechanisms of ODM2 allow for the inclusion of provenance, quality and other metadata. Of note in the development of ODM2 is the collaborative engagement with geoscientists. Although extensibility is very well catered for in ODM2, once extended for a use-case, systems built around the extension do not allow for evolution in an interoperable and efficient manner. Hsu et al. (2017) present several use-cases of ODM2. Arising from their work, several current challenges are highlighted. Adoption of ODM2 enhances extensibility at the expense of reduced optimisation for specific datasets. Also, the generality present in ODM2 makes the schema more ambiguous. Templates for data entry that adhere to the ODM2 information model are therefore difficult to build. Much like the CHARMe project, ODM2 tries to balance flexibility with rigour, an ongoing challenge for interoperability. Also highlighted during use-case implementation of ODM2 was the stark nature of the evolution from ODM1 (Horsburgh et al., 2008). Systems that were originally built on top of ODM1, which now need to evolve to ODM2, require a mapping to be made. This highlights the

problem of creeping system obsolescence and information models in an evolving domain such as ESS. As ODM1 had not been widely adopted, the evolution toward ODM2 was of little consequence. However, when a standardised information model is pervasive in information systems, the evolution of the standard typically slackens as stakeholders are reluctant to re-invest in system migration.

The arguments calling for a more flexible approach to representing geographic information have many similarities to what has been taking place in the health informatics domain of the past decade. In fact, there are many relevant methodologies under development on the fringes of ESS informatics (such as health) that can provide a way forward for interoperable ESS knowledge systems (Diviacco and Leadbetter, 2017).

Archetypes provide an interesting possible solution to the shortcomings of knowledge representation within geospatial information systems. Archetypes have been shown to be flexible, easily scalable and provide a means to handle knowledge evolution. As discussed previously two-level modelling emerged due to the relatively slow pace of evolution of data standards, particularly those that normalise information models, and the problem that coded terms or ontologies alone are not sufficient to achieve semantic interoperability. This issue is also present in the geospatial and Earth system science domain. Two level models are designed to facilitate large-scale sharing of high quality, multifaceted, flexible and durable documentation and with over 20 years of development the two-level modelling community has much to offer the growing area of Earth science informatics.

To help realise the ongoing paradigmatic shift and enable the realisation of the dynamic Digital Earth framework called for by Craglia et al. (2012), this work investigates two-level modelling techniques as a possible solution to manage how information and knowledge concepts are modelled and managed.

Next, chapter 4 considers how two-level modelling can be applied to the geo-spatial domain and presents one of major contributions of this work, a translation methodology of two-level modelling from the health domain to the geo-spatial domain.

Chapter 4

"our knowledge can be only finite, while our ignorance must necessarily be infinite (Karl Popper, 1963)

4. EXTENDING TWO-LEVEL MODELLING BEYOND HEALTH

Chapter Overview: As discussed in Chapter 3 (section 3.5), two-level information modelling has been shown to be a useful tool for tackling interoperability challenges within the health domain. However, to date very little work has been done on applying two-level modelling outside of health (see section 3.5.6). This chapter describes the work done throughout this research project to translate two-level information modelling techniques to the geo-spatial domain; this translation approach is one of the key contributions of this work. This chapter accomplishes the following:

- describes a practical approach for translation of the two-level modelling methodology for the Earth systems science domain.
- examines relevant geographic information-based ISO standards, and assesses their suitability as a basis for a two-level modelling approach.
- identifies key features (e.g. recursive aggregation patterns, ontology bindings) of the two-level modelling approach that need to be embedded in an existing geo-information model to enable it to be repurposed from a model-of-reality to a model-of-documentation while maintaining the core design. As noted previously in section 3.1.6 and Figure 3.4, this is necessary for building real systems.

- proposes a profile of the O&M standard to facilitate flexibility and extensibility in recording observational data, while maintaining interoperability within information systems.

As the novel translation methodology and resultant design concepts are described the text continues to draw reference to other related works within the literature. The structuring of the literature review to continue throughout chapter 4 and also into chapters 5 and 6 has been necessary due to the wide body of work that this research draws from and contributes too. It is also a consequence of the research design approach i.e. the design science methodology used within this work (see section 1.6.1). The *assess and refine* iterative cycle employed as part of the design science methodology causes the text to continually refer back to the literature, beyond what would be expected within a traditional literature review chapters structure.

4.1 Geo Domain Comparison & Analysis

Within healthcare informatics, the primary subject of documentation is the patient. The prevailing consensus within health informatics is that the patient should remain the dominant subject of documentation for shared electronic health care documentation. This is a primary difference between healthcare informatics and geomatics. The subjects of documentation in geo-information and documentation in Earth System Science are diverse. Diviacco et al.'s (2015) work with boundary objects highlights this diversity, and further highlights the current efforts within the geo-sciences community to tackle automatic semantic and dynamic knowledge representation.

Beaulieu et al. (2016) highlight the growing need, and the current state-of-the-art in cyber-infrastructures to support collaborative processes and semantic communication amongst a diverse set of domain specialists. The automatic recording of information is much more prevalent in the geo-sciences as compared to health. As noted in section 3.5,

healthcare documentation can be a matter of life and death for the subject of documentation. Patient safety and quality-of-care issues impose a strong need for rigour in healthcare, and a certain conservatism about changing processes and systems. Of course, the technical tasks required to implement two-level modelling would need to be supported by the same vigorous type of community-wide engagement and dissemination that has characterised the adoption of two-level models in the health domain.

4.2 Domain Translation Methodology

The following technical tasks have been identified as being parts of the process of translating two-level models from the healthcare domain to the ESS domain:

- Develop a generalised identity model that fits the ESS domain.
- Develop functioning binding to coding that is used within the ESS domain.
- Develop a multi-purpose and generic reference model for ESS.
- Development of two-level information representation, communication and processing for resource constrained devices.
- Formation of a suitable community of supporters.
- Development of consensus based ESS archetypes.

4.2.1 Generalised Identity Model

Traditionally in the health domain, subjects of documentation have been restricted to health professionals and patients (Chen, 2016). This has the consequence of limiting two-level modelling to EHRs and the health domain. Chen (2016) also notes that the literature demonstrates that a more flexible definition of the subject of information in the health domain would be beneficial.

In most disciplines, shared identity information is fundamental for interoperability. The geo-spatial domain must adopt a generalised identity model to take into the account the many valid subjects of documentation that may exist. However, this is not an arbitrary

task. The identification of subjects of documentation is highly heterogenous between different systems. For example, documentation about features-of-interest in geomatics could align with the OGCs *general feature model*[44]. However, the particular viewpoint must be taken into account. Observations models typically take a user-centric viewpoint. However other models may take a provider-centric viewpoint.

An identity model must be based on *traits* associated with the subject of documentation. In health, as patient is commonly the subject of documentation, traits may be a patient's name and date-of-birth. Typically, GIS systems are information systems relating to the management of information about geographic objects. In chapter 2 the concept of discrete and continuous geographic objects was presented. Discrete geographic objects have well defined and agreed traits, such as boundary and properties. Continuous geographic objects are different and are typically related to geographic phenomena.

Several identification schemes are employed within geomatics. Object identifiers or OIDs are a standardised mechanism for naming any object, concept, or "thing". OIDs are globally unique and persistent. The global OID reference database maintains a "full-world" record of OIDs[45]. OIDs are also used extensively in two-level information modelling systems within health (such as HL7 and EN13606) (Berry et al., 2010)

Within the geospatial domain, feature identity should relate to moderately persistent real-world objects which are observable as distinct entities such as a lake or an urban area. These entities should exist long enough to be worth naming and talking about (Sargent, 1999). Many of these objects will have fuzzy boundaries. The definition of fuzzy boundaries is an ongoing open research question and is not considered here. Sargent

[44] http://docs.opengeospatial.org/per/16-047r1.html
[45] https://oidref.com/

(1999) notes that both *feature* and *dataset* (feature collection) exist, implying a feature identifier is needed. However, Sargent argues a feature description and a feature handle are needed. Descriptors are used outside any software system, and a handle is an internal tool. Also proposed by Sargent is that geographic objects are "live" objects (as opposed to static) geographical object identifiers should follow a live Web object. Sargent concluded that feature handles are promising, but no unique identifier mechanism was satisfactory.

Today, the problem of unique geographic feature identifiers is still an open question, not least due to the problem of defining the properties of non-discrete geographic features.

4.2.2 Terminology Binding

As already discussed, the two-level information modelling technique relies on archetypes to formalise and define the meaning of health-based data. Archetypes also provide a way to bind data points to recognised terminologies. In health terminologies such as LOINC, SNOMED-CT are typically used for this purpose. OpenEHR for example defines the values of coded attributes within the reference model using its own internal terminology, which defines the meaning of each element. External bindings to terminologies are also supported within the archetype. These archetype bindings connect to external terms that in turn allows querying to be performed using external terminologies.

External binding also allows the specification of value sets from external sources for attributes that may be defined within the archetype. The pre-existence of rich terminologies and ontologies within a domain is beneficial to improving semantic querying where external bindings are possible. Where this is not available (yet), internal definitions should be used in their place. Within the geospatial domain, several rich ontologies and terminologies have been identified that are suitable for this purpose. An example of how bindings can be achieved using these is shown in later in listing 4.2.

4.2.3 Reference Model Selection

Reference models should only contain a small number of concepts or classes. The nature of the classes that appear within the reference model are of key importance. The concepts represented should be non-volatile and also valid for all instances and constant in time (Beale, 2002). Reference model selection is discussed in more detail in section 4.3 below.

4.2.4 Constrained Kernel Development

Within two-level modelling-based systems, at runtime, archetypes are represented in-memory as a set of instances of the classes within the reference model whose characteristics at runtime are constrained against the associated archetype definition. This run-time task is performed by an archetype enabled runtime kernel. The runtime kernel's code base is hardcoded against the reference model, whereas the semantics of the instance data is dynamically retrieved from the archetype definition.

Until now, two-level modelling systems have been developed to be used in a traditional client-server clinical setting. Both client and server have typically been resource rich with "fat" clients dominating real world two-level based systems. In this work, a more *federated* (data are distributed but standardised) approach is required to enable adoption within the geo domain, where tightly constrained and remotely deployed observational platforms are used. Also, a kernel's implementation is tightly coupled to the reference model it supports. As such a new system kernel to support this paradigm is required to be developed.

The implementation details of these new system components are largely dealt within the next chapter (chapter 5), within this chapter the core design principles of the modified constrained two-level modelling kernel are defined and evaluated.

4.2.5 Community of Supporters

Ultimately the success of any information modelling exercise relies on the input of the many stakeholders that may exist. This is the case for two-level modelling. As discussed previously, a core aim of the two-level information modelling approach is to allow domain practitioners to become the key drivers of domain information object definition. Thus, it is important to build a community of supporters within any domain. The work detailed in this thesis is a prelude to this for the geospatial domain. Before a rich community of supporters can be built, the merits of the two-level modelling approach within the geospatial domain must be proven. These benefits must be then communicated to the community and a rich set of tools must be available before widespread adoption and support can be achieved. The building of a community of supporters is a slow process and has been ongoing in health for over 20 years.

4.2.6 Archetype Development

Once the reference model for a domain has been developed, it is then possible to proceed to build a library of archetypes. The quality of archetypes within a domain is dependent on the community of supporters available and the quality of the archetypes, along with the experience of the community members, matures over time. The aim in this work is show by way of example the process and benefits of two-level modelling by providing a base archetype library which can be used to encourage domain practitioners to experiment with the technique. Archetype development is typically done with non-technical actors and as such needs to be supported with a rich set of user-friendly development tools. Archetype development within the geo-spatial domain is discussed in more detail later in this chapter and furthermore in chapters 5 and 6.

4.3 Geospatial Domain Reference Models

From the previous section (section 4.2) it should be clear that the two-level modelling approach is highly dependent on having a valid reference model that is relevant to the application domain. As such, within the translation approach described here the definition of the reference model (level 1 within the two-level model approach) for the geospatial domain is the first task to be performed. Definition of the reference model must be performed first as all other additional translation tasks are dependent on having a valid reference model. Next, the main attributes of a valid reference model are defined and then discussed to explain the rationale for the reference model selection and design within this work.

Beale (2002) comments that one of the main problems with "standard" models is that "they embody no single point of view". Standard models do not deal with change very well and invariably implementations tend to wander to accommodate the peculiarities of any implementation. O&M is part of the OGC's Sensor Web Enablement (SWE) architecture (Botts et al., 2008) and is a semi-structured model. However, as noted by Beale (2002), while semi-structured models are an improvement on standard models, they introduce additional problems; strong typing is typically lost. For example, with loose typing temperature data may not be explicitly defined and left to the system programmer to define. Loose typing leads to differing implementation approaches within real systems, i.e., in a semi-structured model, while the model is still partially concrete, *assumptions* about the information are made and encoded.

As interoperability at the knowledge level is a key requirement for future ESS observational systems and the realisation of a Digital Earth system, it is proposed here that stable domain concepts be captured using the OGC's O&M standard and concepts

with the potential to evolve, or "volatile concepts", be captured using techniques derived from a two-level modelling approach (ultimately archetypes).

Archetypes provide a mechanism for avoiding the pitfalls of over-codification within a singular-model and offer more advantages over semi-structured models. It has been noted that a two-level model is designed to bind to a common terminology to support the creation of archetypes (see section 4.2.2). Within ESS sub-domains there are many advanced domain vocabularies and ontologies, examples include, the Semantic Sensor Network Ontology (SSNO) (Compton, Barnaghi et al., 2012) and Semantic Web for Earth and Environment Technology ontology (SWEET) (Raskin and Pan, 2005). These domain vocabularies can be used to provide semantic support for a two-level approach.

Whereas a reference model is a collection of coherent information models and should capture the stable non-volatile concepts within a domain, ontologies are typically organised into levels. Foundation concepts, which are general across many domains, are captured in an upper-level ontology. Foundation concepts tend to remain stable (i.e. they do not change) over time and are used to produce more specialised domain concepts in sub-ontologies. Reference models therefore should use concepts from upper-level ontologies or knowledge concepts from the foundation/principles level in a multi-level knowledge space. This is to ensure that reference models which form the building blocks for all adoptive systems are stable and generic.

During this work, several geo-spatial foundation level ontologies were investigated, and candidate reference model concepts were identified. Ultimately these were not chosen for this study. The reasons for this are discussed below.

4.3.1 O&M and Principles Concepts

Work done by Probst and Florian (2006) on developing an ontological representation of O&M allows us to assess O&M in terms of an ontological hierarchy. O&M as an ontology

would not be classified an as an upper-level ontology in the strict sense (Cox, 2015a) (Cox, 2015b). Examination of the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) (Masolo, 2003) UltraLite (which is an upper-level ontology) and the Semantic Sensor Network Ontology (SSNO) alignments highlights that O&M concepts are not upper-level concepts. In fact, upon closer examination, the definitions of Observation in both O&M and SSN show that definitions of Observation within these two ontologies are not semantically equivalent, but merely a *close match* (Figure 4.1).

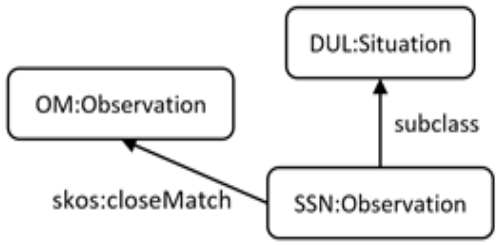


Figure 4.1 DUL, SSN, O&M Alignment

Despite O&M's ontological representation at a sub-ontology level, O&M is considered stable within its given domain. For the purposes of this work and meeting the defined research objectives, O&M concepts are considered as a *principles ontological level* from which content, organisational and storage concepts can be derived within a geo-observational sensor system. Figure 4.2 illustrates this further, where principle level concepts form the core of the knowledge space (centre of the graph). Principle level concepts are true for all scenarios of use within the domain. As we move away from the centre of the graph, additional specificity occurs i.e. a principle level concept is further defined or constrained into more specific concepts, adding meaning and expanding the knowledge space. Concepts can be further defined for particular use cases i.e. a reference model will contain a content, principles level concept that can be constrained to be useful

within an individual scenario. For example, in ocean observing sea_surface may be defined from some content, principles level concept.

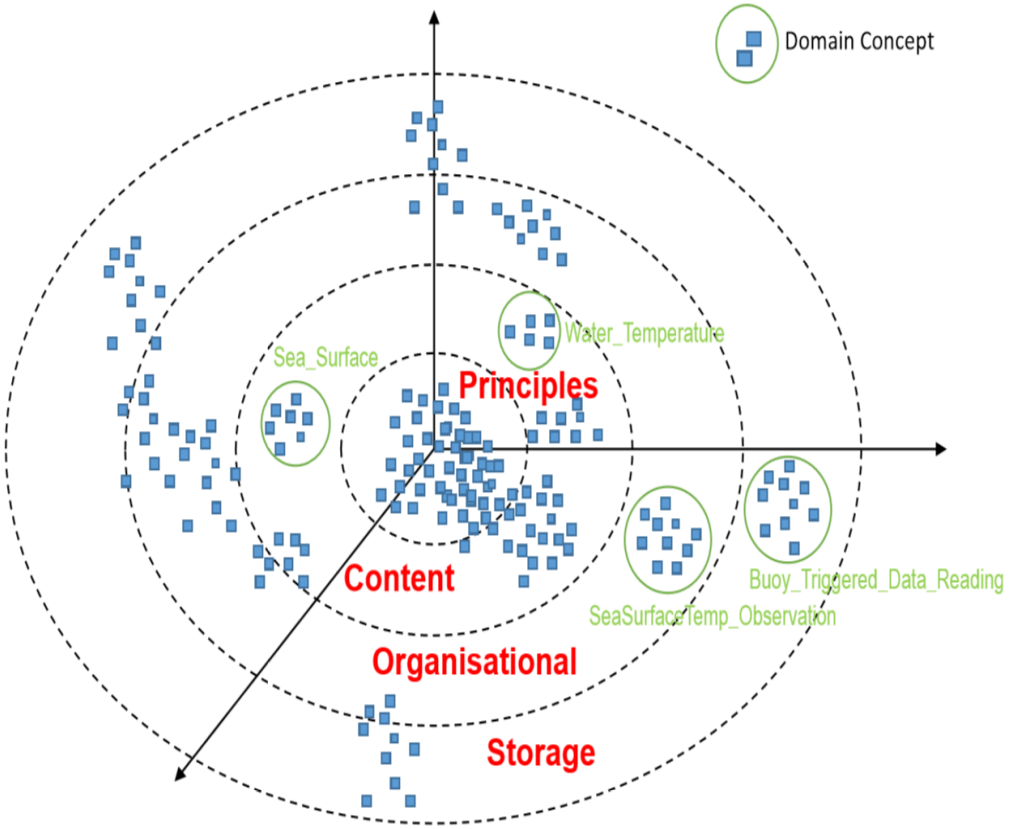


Figure 4.2 Ontological levels. Within a two-level model, O&M as a reference model should only capture stable concepts i.e. at the principles level. They should be true for all instances and all use contexts. Typical of Upper Level Ontologies (Beale, 2002). Here we map higher-level domain concepts derived from O&M principle level concepts onto Beale's (2002) multi-level knowledge space.

In Figure 4.2 it can be seen that the knowledge space is made up of further levels, content, organisational and storage type concepts. Levels form a standardised documentation structure. The documentation structure will be described in more detail below and further illustrated in Figures 4.5 & 4.6 below.

O&M is also chosen as the base reference model to further investigate the applicability of two-level modelling to the geo-spatial domain. Before O&M is used for further investigation, it must first be examined and profiled for the purposes of two-level modelling. This is discussed in more detail in section 4.4.4 below.

4.4 Profiling O&M

The previous section discussed the rationale for adopting O&M as a suitable reference model to investigate to support two-level modelling within the geospatial domain. This section describes the work done on re-profiling the O&M data model to be suitable as a two-level modelling-based and consistent with other two-level reference models within health. Firstly, the topic of recursive aggregation needs to be considered.

4.4.1 Recursive Aggregation Patterns

Careful examination of two-level modelling health-based reference models (such as those within OpenEHR) reveals that two-level modelling reference models are constructed with multiple occurrences of recursive aggregation patterns. This design pattern is essential to enable the main extensibility mechanism provided by archetypes. This pattern is also referred to as the *composite pattern* in software engineering (Bruegge and Dutoit, 2009).

Beale (2002) states that (Beales third principle of knowledge level models):

> *The granularity and composition of a knowledge-level model corresponds to that of domain concepts in the reference model.*

This means that knowledge level (level 2) concepts are derived based on concepts defined within the reference model. Indeed, distinct knowledge level concepts defined using archetypes are *composed* using constraint definitions at the point within the reference model where recursive aggregation is present. A basic representation of the composite pattern is shown below in Figure 4.3. For example, *Leaf* may be a vehicle, which can be further constrained by *composite* into a car.

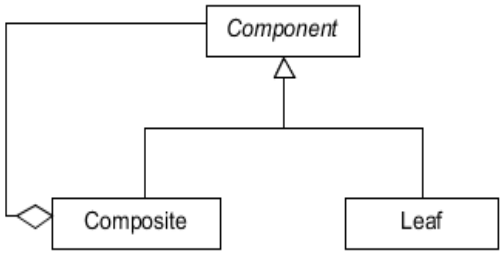


Figure 4.3 Compound/Element Pattern. While not a tree, this structural design pattern enables objects to be composed into (upside down) tree structures and then to handle these structures like individual objects.[46]

4.4.2 Observations and Measurements

Next, a comprehensive overview of the Observations and Measurements data model is presented. It should be noted, at this point, that although it is proposed that O&M has the potential to act as a suitable reference to underpin a two-level modelling approach, it does not contain the requisite design patterns needed to support the proper development of a knowledge model (level 2) using archetypes in its basic form. Therefore, the purpose of this section, is to examine O&M to ascertain suitable points for augmentation and to inform a re-profiling of the data model to make it suitable for two-level modelling.

The geographic information Observations & Measurements (O&M) standard (Figure 4.4) is one of the many standards developed by the OGC as part of the sensor Web enablement framework (see section 3.3.1). All SWE based standards aim to enable the SensorWeb. More specifically, the O&M standard defines a conceptual schema for observations. Features involved in sampling when making observations are also captured among other elements. Before proceeding it is worthwhile defining what is meant by an "observation" and "feature" in the context of O&M. Cox provides the following definition:

[46] https://refactoring.guru/design-patterns/composite (Shvets, 2018) provides a very accessible introduction to this pattern for the unfamiliar reader.

An Observation is an action whose result is an estimate of the value of some property of the feature-of-interest, at a specific point in time, obtained using a specified procedure. (Cox, 2006).

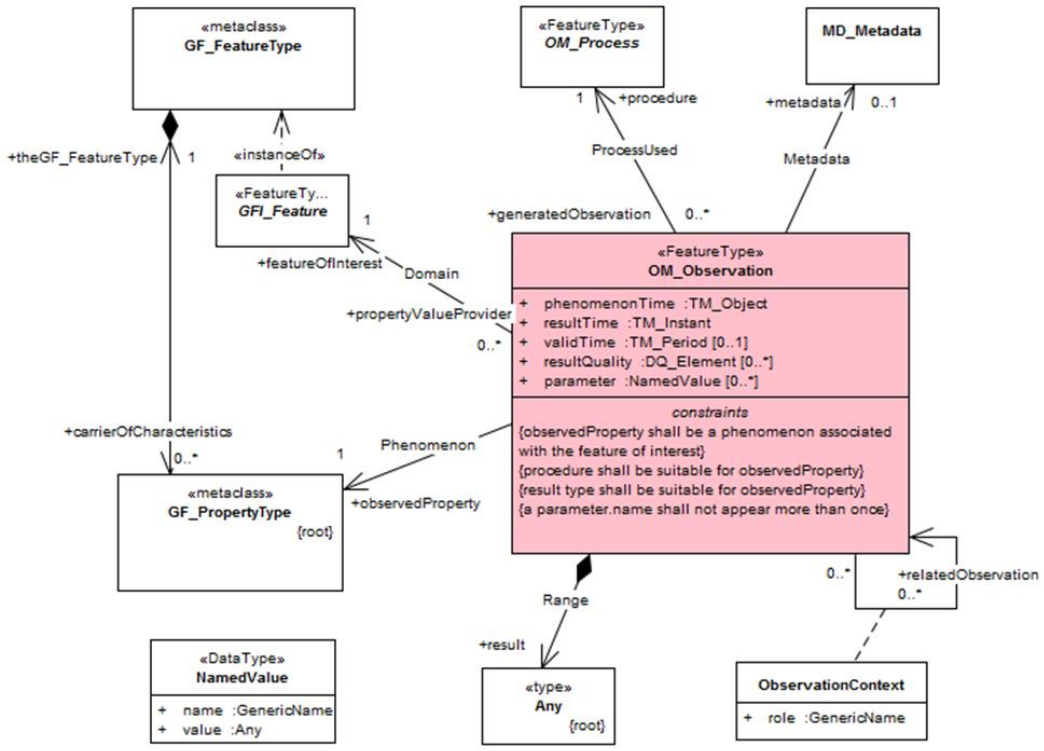


Figure 4.4 Observations & Measurements Standard. (Cox, 2006)

Where a feature-of-interest carries the property that is a representation of a real-world object, or an abstraction of a real-world phenomenon (what Popper refers to as world 1 objects, see section 3.1.1) or for the purposes of this work this could also be considered as a subject of documentation and equivalent to the "patient" in health documentation. Examples could be a domain feature such as the "river Liffey", or a sampling feature such as "tide gauge A" at the north shore light house in Dublin bay. The other elements of O&M that are espoused in Cox's observation definition are captured under the following headings (Figure 4.4):

- Phenomenon time

- Result time
- Procedure
- Observed property
- Result
- Unit of Measure (uom)

Notably the *observed property* is the actual property that is being quantified through sampling. The observed property is a concept description, usually from some controlled vocabulary or ontology, for example “water temperature”. And the result would be the value of this property, for example *18 degrees Celsius*.

As a further example of how O&M concepts relate to the real world, let us take a hypothetical air quality monitoring scenario. Here let us presume an air monitoring station providing air temperature measurements. The feature-of-interest represents the air around the temperature sensor. The property is the air temperature. The observation is the act of measuring the temperature of the air. The result is the value of the property the actual temperature obtained from measurement (single value or time series value) and the procedure represents the sensor or process used to obtain the value.

As mentioned previously, the O&M standard can be classed as a semi-structured model. While Beale (2002) highlights the problems with these types of models in a general sense, Jiang, Li and Guo (2010) highlight this issue specifically with reference to O&M and the issue with standard models within the Ocean observing community:

> “*the design and implementation of the Ocean Sensor Web should maintain a balance between adherence to the GEOSS, OGC-SWE standards, and the concerns of practical and efficient implementation in the ocean observation domain.*”

Jiang li and Guo are referring to the *rigidity* of standards such as O&M while trying to balance requirements from diverse stakeholders. When rigidity exists within a standard,

the concept definitions may not be adapted to a particular use case, and most be adopted. For example, an over specification of the concept temperature to be recorded using Fahrenheit, would be overly rigid for a European context or use case.

As discussed in the section 2.5, O&M is included as an *implementing rule* under the European Union's INSPIRE directive (INSPIRE, 2013). However, even the INSPIRE guidelines for use of O&M notes the issue of variance when using standard models, "O&M is a very generic standard, allowing for very different design patterns depending on the domain as well as the Use Cases to be supported." This type of model genericity and the resulting problems for interoperability was one of the main motivations for the emergence of the precursors to two level models in the health domain 25 years ago.

The application of O&M within a technical community in a new way that enables shared computable resources requires that the community agree on standard content for the key slots in the model, as well as on required extensions to the base classes provided within the standard. In particular, it is necessary to have standard vocabularies.

4.4.3 O&M as a Two-Level Modelling Reference Model

As discussed above, ideally any reference model should be formed using level-0 principles ontological concepts (Figure 4.2). When proposing O&M as a valid reference model, one must question whether O&M represents level-0 principles. For example, it is correct to state that DOLCE UltraLite represents level-0 concepts and could serve as a level-0 principles concepts-based reference model for the purpose of this translation. In that scenario, O&M could act as a basis for discovering content level concepts. Meaning that O&M would be represented in the second level of the two-level model as a set of archetypes. Having considered this core translation decision in detail, the author recommends that O&M should form the core reference model. The stability of O&M concepts is assumed to be sufficient to act as a set of level-0 knowledge concepts, and

hence as a basis for a reference model. This assumption is further strengthened through the adoption of O&M as an ISO standard, and its inclusion within the INSPIRE directive. Next, we must consider which principle (ontological) concepts represented as classes within an O&M aligned reference model can and should be *archetyped*?

An observation consists of: phenomenon, location, value, time, producing sensor; or *OM:ObservedProperty*, *OM:FeatureOfInterest.location*, *OM:Result*, *OM:Observation.phenomenonTime*, *OM:Procedure*. These principle concepts can be used to construct content level domain concepts. ObservedProperty allows for content level domain concepts to be further defined e.g. Temperature. The author has examined O&M's suitability as a reference model and found that O&M does not contain all the necessary base and container types or appropriate aggregation patterns that are required for two-level modelling. Keeping in mind the requirement for the core O&M standard is maintained, while still enabling O&M's use as a reference model in a two-level modelling approach, a recommended augmentation of an O&M aligned reference model design pattern has been developed. The proposed augmented O&M model is presented next.

As discussed above (section 4.4.1) within the two-level modelling approach, the only way in which data instances can be created, is from direct use or specialisation of elements of the reference model. Any two-level modelling reference model must provide a means of representing entities that are not concretely modelled. Using a compound/element pattern within a reference model allows the creation of recursive aggregation of domain specific concept objects from the non-volatile concepts captured within the reference model. The creation of recursive aggregation of objects from the non-volatile concepts is a core requirement for any reference model within a dual-model system.

A proposed augmented O&M model or O&M profile incorporating the necessary compound/element patterns to facilitate domain specific concept creation is shown in

Figure 4.5 below. The compound/element pattern that is needed within any reference model, can be clearly seen in Figure 4.5 (highlighted in green). The insertion of this pattern within the O&M model represents points of extensibility within the model. As highlighted in Figure 4.3 above the compound element pattern allows for the creation of upside down tree structures that can enable an increasing level specificity to be defined at each particular point within the model. For example, the inclusion of the *details* attribute within the Observation class allows for the controlled extension of standardised additional Observational details as details is of type Details_COMPOUND.

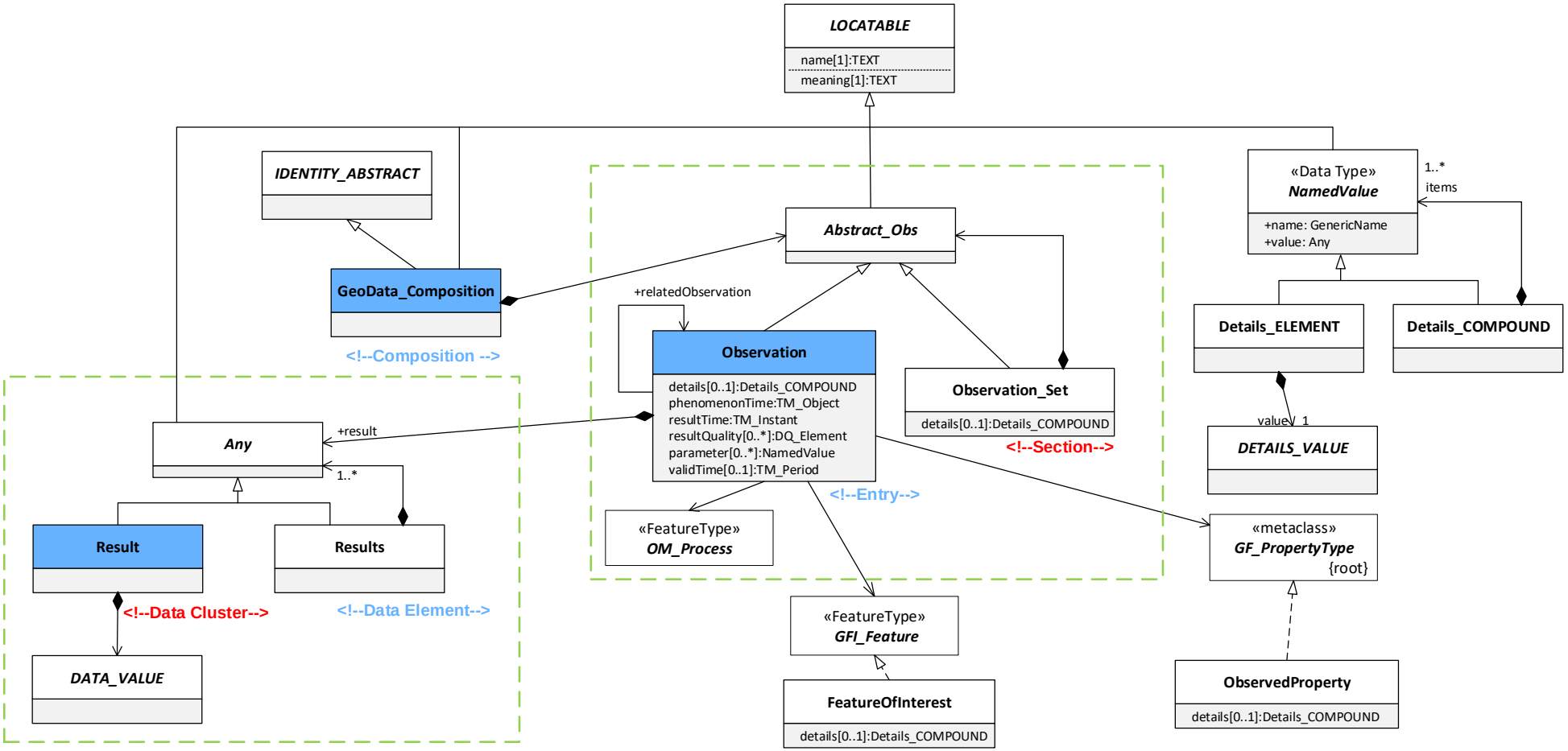


Figure 4.5 Augmented O&M model. Compound/element patterns are highlighted with a green bounding box.

The augmented O&M model is serialised using XSD (Listing 4.1). Appendix A provides a full listing of a serialised XSD representation of the profiled O&M model shown in Figure 4.5.

```
...
<xs:complexType name="GeoData_Composition">
  <xs:complexContent>
    <xs:extension base="IDENTITY_ABSTRACT">
      <xs:sequence>
      <xs:element name="archetype_node_id" ... maxOccurs="1" />
      <xs:element name="name"  .../>
      <xs:element name="details_Compound" ... />
         ...
</xs:complexType>

<xs:element name="GeoObservation_set" type="OBSERVATION_SET" />

<xs:complexType name="OBSERVATION_SET">
  <xs:complexContent>
    <xs:extension base="ABSTRACT_OBS">
      <xs:sequence>
        ...
```

Listing 4.1 XSD snippet of augmented O&M model with compound/element patterns. The augmented O&M model serves as the reference model within the dual-model approach. The 2nd level is captured using an archetype-model which are constraint statements on the reference model.

In Figure 4.5, GeoData_Composition represents a meaningful aggregation level. At this level within the representation, a basic flexible identity model (see section 4.2) is provided for (see IDENTITY_ABSTRACT in Figure 4.5). However, the question of a generalised identity model within ESS information systems remains an open question. Chen's (2016) work on generalised identity models for healthcare may provide a way forward within the ESS domain.

The GeoData_Composition pattern provides a mechanism for domain practitioners to extend the model and create document level knowledge representation of specific use case domain concepts. Note that O&M represents a model of reality, the augmented O&M model provides for a model of recording in addition to the model of reality, this is a well-established design principle in health informatics (Beale, 2003).

The structure of the document provides for three levels of meaningful aggregation from which concepts can be created at the necessary ontological levels i.e. Storage,

Organisational and Content (Figure 4.6). The reference model itself captures the principle ontological level, or the stable concepts within the domain.

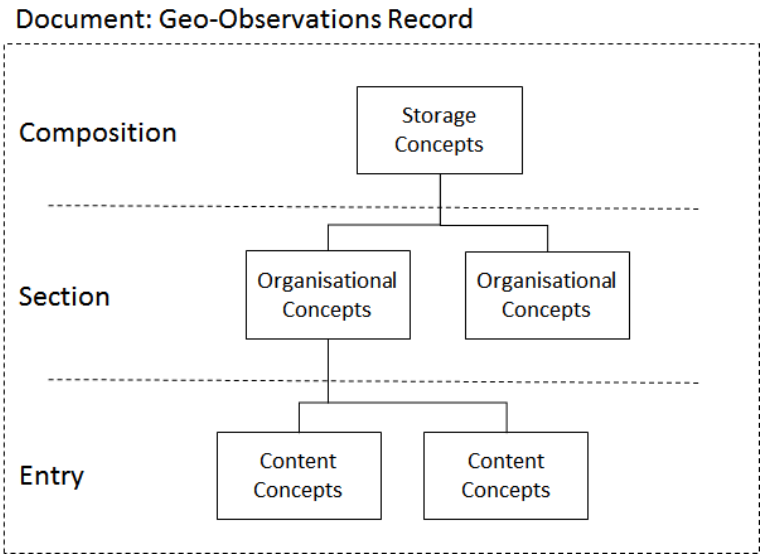


Figure 4.6 Document Structure

4.4.3.1 Archetyping using O&M

One of the core principles of the two-level modelling approach is that it should enable domain practitioners to capture specific domain knowledge concepts and to manage them as they evolve over time. The 1st level, or reference model, is still developed by Informaticians. The 2nd level, or the knowledge level, is developed by a mixed group of authors, that include the domain practitioners themselves who now have greater influence on the evolution of models within a community environment.

Figure 4.7 below shows the separation of the two levels and highlights the mapping of volatile concepts at the Storage, Organisational and Content ontological levels to the stable concepts within the domain at the principles ontological level. As noted previously, all data instances are of the reference model. However, these principle concepts are constrained at runtime using the knowledge model, or archetypes. Archetypes are constraint statements and an archetype model represent a rich knowledge level model of

domain concepts. Information instances are created at runtime from the reference model. These instances also adhere to the constraints defined within the archetype model.

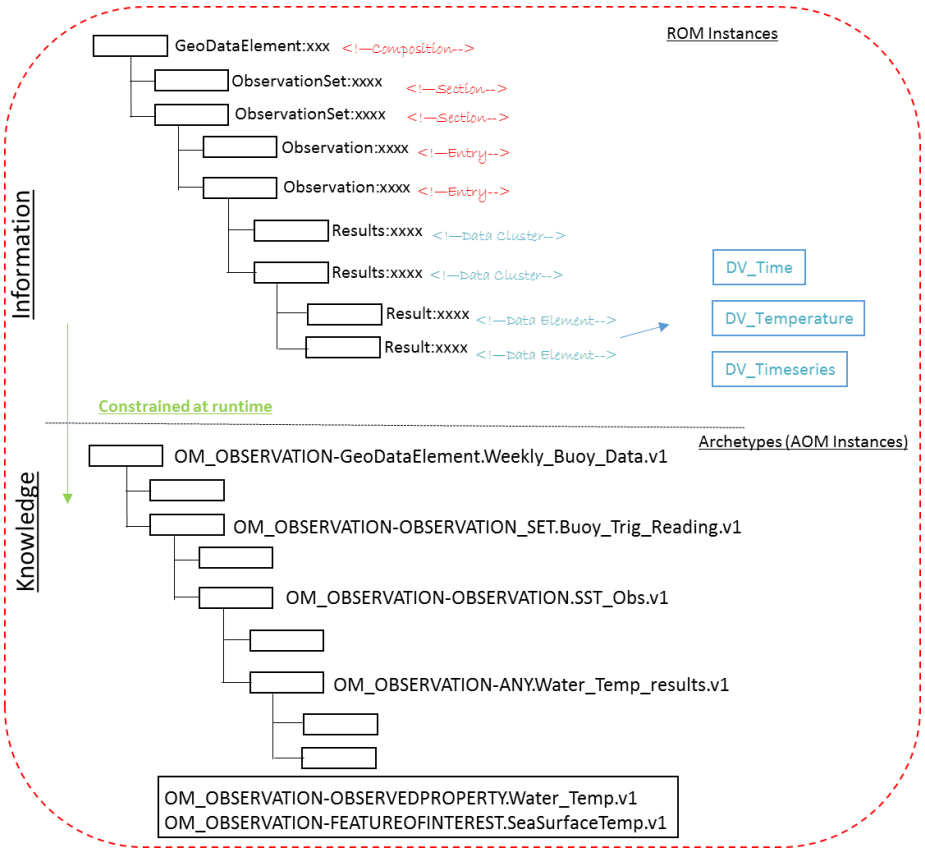


Figure 4.7 Here an Archetype Model (AM) is used to constrain the augmented O&M Reference Object Model (ROM) instances at runtime

The proposed augmented O&M model (Figure 4.5) acts at the reference model level. As this is a novel exercise, as of December 2020, there are no existing tools to fully aid archetype development outside of the health domain. However, there are a number of health informatics-based tools that can be used to aid initial development in other domains such as ESS. The Biomedical Informatics Group at the ITACA Institute at the Universitat Politècnica de València have developed the LinkEHR platform (Maldonado, Moner et al., 2009). The relationship between archetypes and domain expert (health domain expert in this case) is shown in Figure 4.8 below.

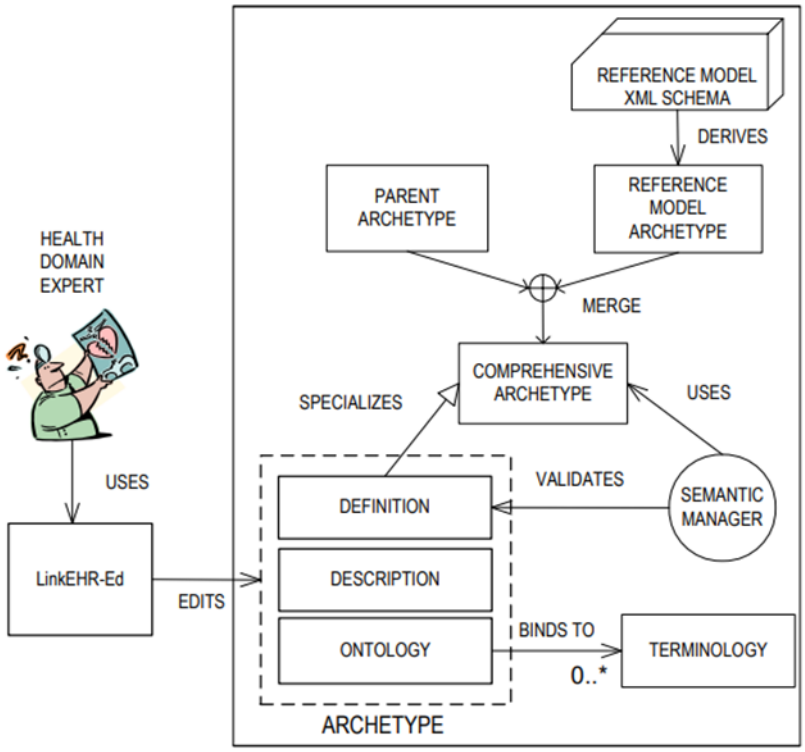


Figure 4.8 LinkEHR archetype editor. Image reproduced from the Doctoral Thesis of Diego Boscá Tomás (Boscá Tomás, 2016)

The LinkEHR platform includes an Archetyping Editor tool that allows for the development of Archetypes from any reference model. EHRFlex (Blobel et al., 2010) provides a flexible tool that may be used to further two-level modelling outside of the health domain. EHRFlex only supports CEN/ISO 13606; support for any archetype-based standard is planned in the future. The OpenEHR Java reference implementation allows for further development of existing tools for non-health domains (Chen and Klein, 2007). Using LinkEHR, the author has demonstrated how a serialised XML form of the augmented O&M model can be used to develop archetypes (see Listing 4.1).

The LinkEHR editor provides a visual development tool for the creation of archetypes or reference model constraint statements. This visual approach to archetype development enables domain specialists, who may be non-technical or expert in information modelling, to produce the required content models. Once the serialised (XML for example) form of

the reference model is available, the visual modelling tool approach can then be used by domain practitioners to meet their needs. Once domain modelling has been agreed upon using the visual interface, LinkEHR will output an ADL representation of the archetype model so that it may be machine readable within a supporting system (as shown in Listing 4.2).

It is important to note that archetypes or constraint statements for particular use-cases are agreed upon by the community, domain experts or practitioners using visual tools. It is this ability to derive community agreed standards through the consensus of empowered domain practitioners that offers the real benefits in terms of knowledge interoperability of systems. Archetype models evolve progressively and "naturally" as the community's knowledge and understanding of the domain advances. The community evolution of archetypes contrasts with the more traditional development of information models and standards; which happens over a longer time cycle, in a more top-down approach. In any use-case, it is acknowledged that there is a need to have a general agreement on the basic structural elements of the information; this is the role of an O&M based reference model. However, using two-level modelling, it is also possible to acknowledge the need of specific practitioners within an ESS community to agree on specific datasets for specific purposes that can be easily changed with evolving requirements and understanding.

Here a simplified use-case is described, which is nevertheless useful for the purposes of illustration. The intended users are a diverse community of ESS-based practitioners. In this example, assumed to be an expert oceanic group or international research project wishing to share information & knowledge from a set of globally deployed data buoys. The adapted form of O&M that is shown in Figure 4.5 is chosen as the common information model to achieve systems integration and process and combine observational data. Through O&M, syntactic interoperability can be achieved. However, since there are

few opportunities for consistent use of constraints in O&M, in a diverse community, variance will occur in implementation of O&M constraints. Also, O&M base concepts can be interpreted in different ways, and so semantic integration must happen manually. With the addition of an ontology, O&M concepts can be linked to common vocabularies to increase semantic interoperability. However, variance in implementation of the underlying information model that covers different ESS observation and documentation use cases has thus far not been agreed by the community, and therefore the implementations may "wander". A framework is needed to allow all parties to agree.

The adoption of a two-level modelling approach provides the community a way to develop a consensus for use-case variance, defined in a rigorous and machine process-able way. Firstly, the community must agree to use the augmented O&M information model. The community then engages in a consensus driven process of agreeing and formalising additional constraints, based on the collective knowledge of the community. The process is supported by a shared set of distributed visual design tools, which enable the rigorous and flexible extension of the O&M model. First the O&M profile from Figure 4.5 is adopted, then a consensus on the domain concepts needed is agreed up, which are constraints on the O&M based concepts. The community agreed knowledge constraints are now captured within a set of archetypes (represented in ADL format).

Listing 4.2 shows an example of an ADL representation of an archetype developed using the LinkEHR editor. For this use-case, the author has performed the modelling exercise. The archetype shown in Listing 4.2 captures a storage level concept Weekly_Buoy_Data, which is a constraint on the principles level concept GeoData_Composition. As GeoData_Composition exists in the reference model, it is assumed that it is stable and generic to the point of being usable as a base concept for the derivation of all other information concepts.

Listing 4.2 serves to illustrate how constraint statements are captured in a machine-readable format. ADL is used in this instance as the representation; however, the archetype may be represented in any serializable format such as XML or JSON. Listing 4.2, (line 10) shows that this archetype provides a metadata description for a record of Weekly_Buoy_Data, which is a specialisation of the reference-model based concept GeoData_Composition. TPOT-OM refers to the augmented O&M reference model developed by the towards People Oriented Technologies (tPOT) research group at the TU Dublin City Campus, of which the GeoData_Composition concept is a member. Weekly_Buoy_Data is a domain specific volatile concept, defined here in the 2nd knowledge level. Weekly_Buoy_Data is assigned the reference [at0000] and is defined further in Listing 4.2 under the Ontology section. Also, of note, is the possibility to bind concept definitions to ontologies (concept at0005 bound to nerc::TEMPR01[47])

[47] https://vocab.nerc.ac.uk/

```
archetype (adl_version=2.0.5)
    TPOT-OM-GeoData_Composition.Weekly_Buoy_Data.v1
concept
    [at0000]
language
    original_language = <[ISO_639-1::en-ie]>
description
    original_author = <["organisation"] = <"tPOT">>
definition
    GeoData_Composition[at0000] matches {                                   -- Weekly_Buoy_Data
        archetype_node_id existence matches {0..1} matches {*}
        ...
        details cardinality matches {1..*} matches{....}
        observation_set cardinality matches {1..*} matches{
            OBSERVATION_SET[at0002] matches {
                ...
                details cardinality matches {1..*} matches{...}
                observation existence matches {1..1} cardinality matches {1..*} matches{
                    OBSERVATION[at0004]occurrences matches {1..*} matches {
                        ....
                        observedproperty existence matches {1..1} matches {....}
                        featureofinterest existence matches {1..1} matches {....}
                        ....
                        results existence matches {1..1} matches {
                            RESULTS[at0007] cardinality matches {1..*} matches{
                                result occurrences {1..1} matches{
                                    RESULT [at0008] matches{
                                        ...
                                        details cardinality matches {1..*} matches{
                                            ...
    }
ontology
    terminologies_available = <....>
    term_definitions = <
        ["en-ie"] = <
            items = <
                ["at0000"] = <
                    text = <"Marine_Data_Buoy_Weekly_Report">
                    description = <"Marine Data Buoy Weekly Report">
                ....
    constraint_definitions = <....>
    term_binding = <
        ["nerc"] = <
            items =<
                ["at0005"] = <[nerc::TEMPPR01]>
                ....
```

Listing 4.2 ADL snippet representation of an archetype developed using the LinkEHR editor

Again, emphasising that an archetype is developed using a community consensus approach, we can see in Listing 4.2 (which is a simplified version) that the archetype model allows a community to agree and document the domain specific use-case implementation specialisms needed on top of the reference model. This ability to document specialisms in this way enables the efficient management of the evolution of any system using archetypes. Grossner et al. (2008) refer to this ability as the extensibility requirement of a Digital Earth system. To-date any variance or specialisation needed during use-case implementation of O&M, have not been managed in a structured way. Also missing from the specialisation of O&M is a well-established general community

consensus mechanism (INSPIRE, 2007) (Klein, 2009). This resultant unmanaged variance in the implementation of standards, although often done for valid reasons, is a barrier to semantic interoperability of systems, especially at the knowledge level.

For a system to capture, store and serve data that adheres to the community knowledge model, the archetype is applied as a set of constraints to guide the production of the information objects at run-time. The archetype describes the structure and detail of instantiated records of information. At run-time, archetypes may be represented in memory in an archetype-enabled kernel. Archetypes are intended to be maintained using a Web based management, review, validation and publishing library system. Communities of domain experts access and contribute to the archetype management system, taking part in the review and validation process.

For the purposes of illustration, let us assume that a community of domain experts have agreed on, and validated the archetype *TPOT-OM-Geo_Data_Element.Weekly_Buoy_Data.v1*. Any geo-observation system serving data to the community implements the reference model, i.e. all data objects are produced from the underlying RM. In this case, a system would use the augmented O&M reference model as the basis for data object instantiation. As the RM would be considered stable and not subject to change over time, the core system software does not need to change over time. At run-time, the system constrains the data objects based on the constraints defined in the archetypes. Constraining of RM based data objects takes place at run-time through the processing of the machine-readable ADL file, using the archetype-enabled kernel. As the needs and understanding of the community of domain experts changes and evolves over time, the two-level based system also adjusts how it creates the data objects based on the evolution of the corresponding ADL file that encodes the archetype constraints.

```
[{
    "id":"identity_model_ref_ID",
    "geoData_Composition":{
        "archetype_node_Id":"TPOT-OM-GeoData_Composition.Weekly_Buoy_Data.v1",
        "name":"Marine_Data_Buoy_Weekly_Report",
        "details_COMPOUND":[{ . . . .}],
        [{"geoObservation_Set" :{
                "archetype_node_Id":"[at0002]",
                "name":"Buoy_Instrument_Readings",
                "meaning":"Interval Triggered Buoy Multi Instrument Read",
                "details_COMPOUND":[{ . . . .  }],
                [{"Observation":{
                        "archetype_node_Id":"[at0004]",
                        "name":"observation_measure",
                        "observedProperty":{
                                "details_COMPOUND":[{
                                        "details_ELEMENT":{
                                                "archetype_node_Id":"[at0005]"
                                                "DETAILS_VALUE" : "temperature"
                                                . . . .
                        "featureOfInterest":{
                                "details_COMPOUND":[{
                                        "details_ELEMENT":{
                                                "archetype_node_Id":"[at0006]"
                                                "DETAILS_VALUE" : "Sea Surface"
                                                . . . .
                        "type":"Measurement",
                        "results": [{
                                "archetype_node_Id":"[at0007]"
                                "result":[{
                                        "archetype_node_Id":"[at0008]"
                                        "DATA_VALUE" : "10.23"
                                        "details_COMPOUND":[{
                                                "details_ELEMENT":{
                                                        "archetype_node_Id":"[at0009]"
                                                        "DETAILS_VALUE" : "celsius"
                                                        . . . .
                "Observation":{ . . . "name":"observation_time_series", .... }]}
        }]}
        [{"geoObservation_Set":{...}}]
 }]
```

Listing 4.3 JSON representation of an information instance. The resulting information instance from the compound/element patterns within the augmented O&M reference model (Figure 4.5) highlighted in green. GeoData_Composition is the realisation of the inclusion of a model of recording/documentation within the O&M based reference model. The archetype_node_id attribute is inherited from the LOCATABLE class in Figure 4.5 and allows bindings to occur between instance data and the AM.

Listing 4.3 above depicts how an information instance may be represented. The O&M JSON encoding OM-JSON (Cox, 2015) is used as the basis for this example. OM-JSON provides several schemas where validation of specialisations of O&M may be performed. In the methodology proposed in Cox's paper, the base O&M schema would still be captured (reference model) but the specialisations (volatile quasi-static concepts) would be captured in the archetype model. Validation is one of the primary run-time uses of archetypes. Archetype-validation tools and frameworks are available (Chen et al., 2008). Using this approach, the full power of the community consensus approach and associated

tools would be available to evolve and manage specialisations. Here it is argued that this approach is in keeping with and helps to realise the vision of the dynamic Digital Earth framework set out by Craglia et al. (2012) (discussed in Chapter 1).

Upon examination of Listing 4.3, it is of note that there is some overhead associated with this approach. It is necessary to record which archetype was used for data construction; this appears as `archetype_node_id`. Archetypes themselves are separate from their data, and need to be stored in an accessible repository (Figure 4.9). It can be seen that there are three levels of information in our example, wholly-static concepts that are captured in the reference model, quasi-static concepts which are agreed in each of the archetypes and dynamic data or instance information. A pragmatic approach to managing the growing volume of the dynamic data instance shown in Listing 4.3, is to identify additional static information from the information instance of dynamic data that may reside in the archetype, and remove this from the information instance.

The current approach produces overly verbose information objects and thus is outside the processing power of many sensor based observational systems. The challenges of constrained in situ remote sensor systems are not considered at this point. At this point it is assumed that the sensor system has the resources necessary to support the archetype-template runtime environment and associated kernel. Figure 4.9 shows several separate supporting systems. Development of a library of community-derived archetypes is supported by the online management system and archetypes are available through an online repository. For any specific use-case, the system builder and associated domain specialists use the necessary subset of archetypes available with the library. These archetypes are further specialised for the use-case and are combined to produce a set of

Operational Templates (OPT)[48] (Leslie, 2008). This ability to produce OPTs adhering to a rigorous formalism is a key advantage of two-level models. Previously it was noted that solving the challenge of data entry templates is an ongoing issue in ESS with ODM2. The hypothetical in situ remote sensor system shown in Figure 4.9 uses OPTs locally to instantiate information instances, such as the one shown in in Listing 4.3. Information instances may then be transmitted to a supporting data-store for persistence.

As information-instances are created using the reference model, which in this case adheres to the O&M specification, the observation system can now conceivably publish semantically rich, interoperable data and information, which evolves as the knowledge community evolves. As the system also adheres to a core standardised information model, such as O&M (in this case), syntactic interoperability is maintained, and it becomes a relatively rudimentary task to make observations available to an OGC standardised Sensor Observation Service (SOS) (Bröring et al., 2012).

As archetypes have a predictable structure, derived from the underlying reference model, enhanced querying can be achieved using the Archetype Query Language (AQL) (AQL, 2015). AQL is a fusion of SQL, and XPath style paths, derived from the archetype. Archetype paths transcend the archetype into the instance data, in the form of archetype node identifiers (Listing 4.3). This ensures conformance of archetype path structures as data nodes are constructed at runtime and allows data nodes to be extracted using complex queries.

4.5 System Deployment Challenges & Solutions

Figure 4.9 below shows a hypothetical deployment scenario for an ESS based observational sensor system. It is assumed here that the in situ remote sensor system is a

[48] As of August 2020, the ADL support required to express templates in ADL has been published, however tool support is still some way off.

marine data buoy. Firstly, the observing platform is characterised, before discussing the challenges in deploying a two-level modelling approach to the given scenario.

Data buoys can be categorized into several different classes such as surface, sub-surface, near-shore or off-shore. The term buoy typically refers to the float of a buoy system. Buoy systems incorporate anchoring, floats and installed instrumentation (Berteaux, 1976). Here, the term buoy system is used for a singularly deployed physical float with anchorage and instrumentation, that is both near-shore and of the type surface, with additional sub surface instrumentation.

Data buoy systems are typically technologically constrained systems, with power and deployment location dictating the buoys available computing, storage and communications ability. These resource restrictions typically prescribe the use of somewhat impoverished methodologies to describe, transport and store resultant observational data.

Several disparate inter-connected supporting systems are shown in the hypothetical system. Development of a library of community derived archetypes is supported by the online management system and archetypes are available through an online repository. For any specific use-case, the system builder and associated domain specialists use the necessary subset of archetypes available within the archetype library.

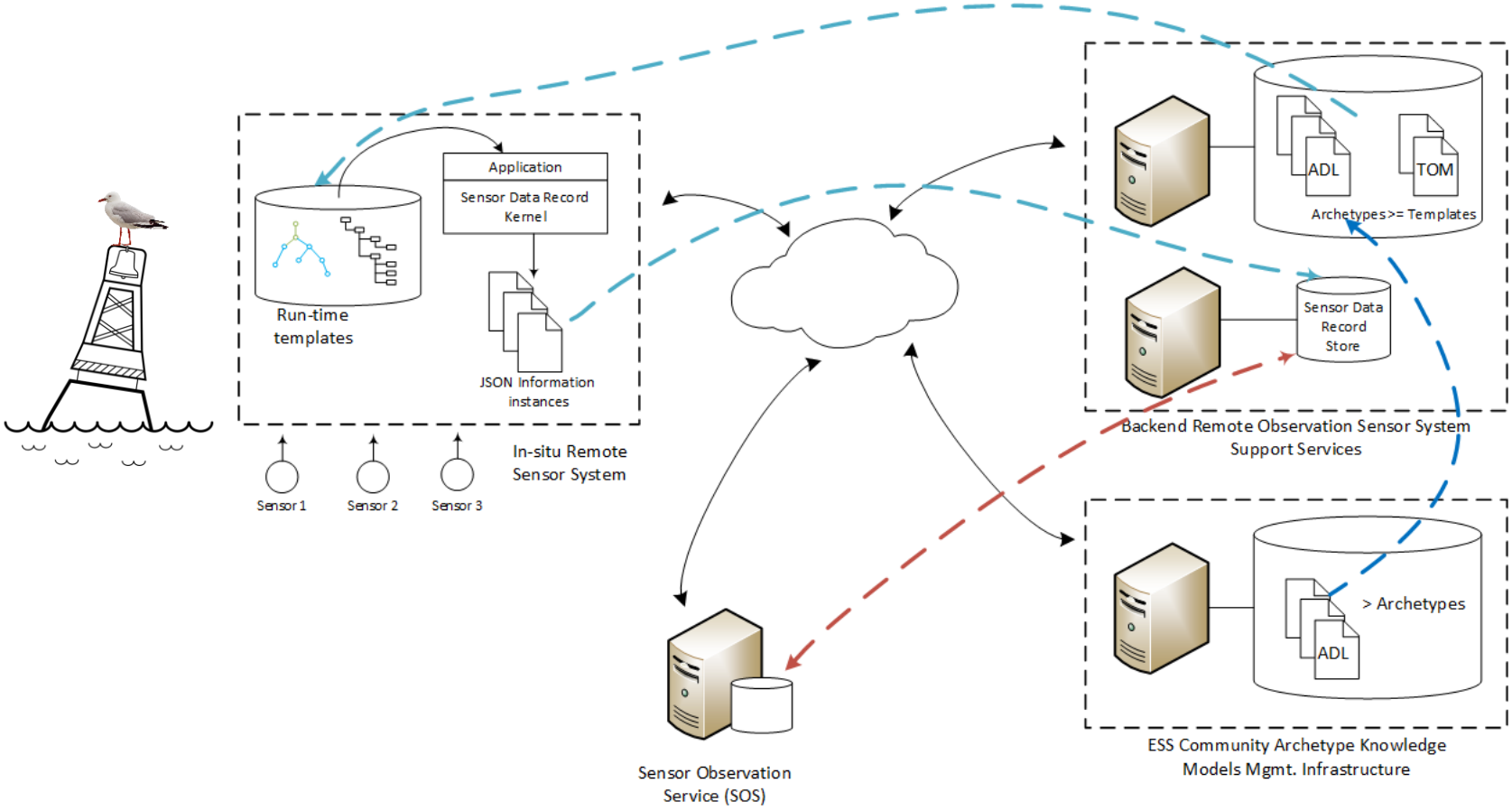


Figure 4.9 Overview of a two-level model support observation sensor system architecture. The additional processing, storage and communication load has been found to be prohibitive for deployment across many data buoy platforms. Run-time templates need a kernel to run on the data buoy platform.

Archetypes may be further specialized per use-case, and location, and are combined to produce a set of operational templates (OPT). The in situ remote sensor system uses these OPTs locally to instantiate information instances, as shown in Listing 4.3.

Once created, information instances are transmitted to a supporting data-store for persistence. Information objects are instantiated from the reference model only. Information instances form a *directed-acyclic graph* that contain labels or bindings at various points. Bindings are in the form of atcodes and relate the information instance concepts to their knowledge domain specific concept, defined within the ADL based archetype or operational template.

4.5.1 Dealing with Technological Constraints

Creating knowledge rich information objects adds significant additional overhead in terms of processing, distributed cross referencing to knowledge resources (terminologies etc.), storage and transportation. Constrained data buoy systems typically do not have the resources needed to implement a typical archetype-based system deployment. Archetype methodologies have been developed for the health domain, where typically constraints on systems are not of major concern. Scaling issues can be solved through vertical and/or horizontal system scaling. This is not possible on a data buoy system.

It has been noted that within an archetype-based approach three levels of information exist, *wholly-static* concepts that are captured in the reference model, *quasi-static* concepts which are agreed in each of the archetypes and dynamic data or instance information. Processing and transportation of static data within a constrained system represents wasted resource usage. By identifying static information residing within an archetyped information instance, and removing this from the information instance, a leaner information object can be realized.

Most observational systems should only need to report timestamped DATA_VALUES, identifiers for identities, associated coded value bindings and the archetype to which the instance is bound; and to which it may be validated against. Presentation layer applications using data instances can later reconstitute the semantically rich information using a knowledge framework that is similar to the one shown in Figure 4.9.

Archetypes follow a tree like structure derived from the compound/element patterns inherent in the underlying reference model. Data instance structures must also follow the same tree structure of the underlying reference model, and associated archetype. As a result, two-level system data instances are a specialized type of graph data. Recently recorded observational data instances may only represent a simple node within the larger complex instance data graph. Fragmentation of an information instance temporarily as a distributed graph, or federated graph, with the bulk (wholly-static and quasi-static nodes) of the information instance residing in a resource rich backend infrastructure, an observational system need only process and transport a minimal data node within the overall data graph. This approach is analogous to the Linked Data (Bizer, Heath and Berners-Lee, 2009) approach developed in recent years to realise the semantic Web. Using Linked Data approaches, a fragmented archetype'd information instance can be hosted across a supporting knowledge eco-system.

Archetype based systems are designed to enable the creation of knowledge rich interoperable documentation of domain use-case knowledge. The creation of information usually results in appending information to a document. Distributing the information instance between the in situ observational system and the back-end knowledge framework reduces the potentially significant overhead on constrained observational systems that is mandated by a dual-model approach.

As described in Chapter 3, Linked Data is an approach for exposing, sharing and connecting structured data using URIs and RDF and JSON-LD, is a more efficient concrete representation of an RDF data model. JSON-LD and the linked data concept has been shown to be useful in managing the overhead of complex information within geospatial data on constrained observing systems. In the next section, RDF and the linked data approach is considered in the context of representing archetypes and archetype models efficiently in technologically constrained systems.

4.5.1.1 Archetypes and RDF

The RDF data model is composed of atomic data entities referred to as semantic triples. A triple is composed of three nodes within the RDF graph and codifies a statement about semantic data. Triples of this type are the basis for representing machine-readable knowledge. An RDF graph can be visualised as a node and directed-arc diagram in which each triple is represented as a node-arc-node link (Subject - Predicate - Object). As described in Chapter 3 (section 3.2.3) RDF creates a graph structure to represent data. Serializations of RDF such as JSON-LD allow the markup of data instances using a structured data graph. RDF does not describe how the graph structure should be used. RDF schema (RDFs) is a schema language that allows information modelers to express the meaning of the RDF graph data. RDF and its schema extension RDFs provide support for distributed information and can be used to realize the data instance fragmentation described above. However, RDF & RDFs do not provide the same semantic modelling capabilities as a reference model with an associated constraining archetype. The Ontology Web Language (OWL) (se section 3.1.2) provides additional vocabulary and semantic formalisms to RDF/RDFs. For example OWL provides the owl:Restriction construct (Antoniou and Van Harmelen, 2004).

OWL provides rich semantics that are useful for a solving heterogeneity within a federated data paradigm such as Linked Data. To enable the power of a two-level information system design approach within a constrained buoy system, the author proposes the fragmentation of archetype'd data instances using a Linked Data approach. Fragmentation can be realized within the dual-model approach by employing Semantic Web technologies and techniques.

Lezcano et al. (2011) have shown how archetypes can be translated automatically into OWL, to enable a reasoning engine based on archetypes. Kilic et al. (2005) provides a succinct introduction to the steps necessary to translate archetypes represented in ADL to OWL. The Artemis project (Dogac et al., 2006) developed a framework to map archetypes between different standards. A syntactic transformation of (ADL-defined) archetypes into OWL format was produced. However, the Artemis framework requires a manual mapping to take place. The Poseacle project (Fernandez-Breis et al., 2007) also provides a semantic transformation of ADL archetypes into OWL.

The Born Semantic approach described in (Buck and Leadbetter, 2015) uses the O&M JSON encoding OM-JSON (Cox and Taylor, 2015) to support a Linked Data approach. The process used is to overlay OM-JSON onto JSON-LD, this allows an RDF inferred graph to be created. In this work, the author proposes that the ADL defined archetype, or operational template serves the function of OM-JSON proposed for Born Semantic systems in a more flexible way, while realizing the greater benefits of two-level modelling.

The archetype approach has been designed to append data to documents rather than replace or delete. This works well for the approach presented here where the JSON instance shown in Listing 4.3 is coerced to the JSON-LD format. The JSON-LD inferred RDF graph is composed of *tripified*-data. Triples serve as the basis upon which

fragmentation of the information object can occur. Figure 4.10 illustrates the approach. The inferred graph in Figure 4.10 is made up of node-arc-node structures. Each node within the graph represents an entity, which can hold any number of attributes. In JSON an attribute is a key-value pair. A triple contextualizes a node, forming a relationship based on a predicate. For example, *Observation – has – Results*; *Results-contain-Result*. Using JSON-LD each set of key-value pairs (node) can be located on a different physical data-store, within a distributed or federated information system, similar to the "shards" concept used in MongoDB (Chodorow, 2013). The distributed graph data approach means that a constrained observational system, such as a data buoy must only serve the necessary key-value pairs of a *Result*, once context for that result is provided, or once the result node is *tripified*.

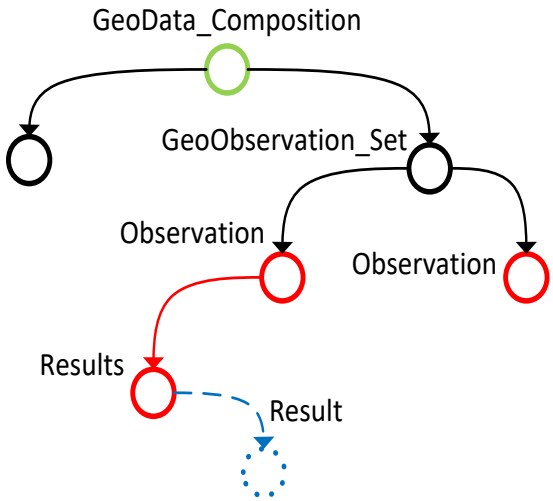


Figure 4.10 Archetype'd information instance graph representation. The Result node contains the dynamic information that is observed from a data buoy system. The graph is formed using a Linked Data approach.

4.5.1.2 Information object fragments & JSON-LD

JSON-LD is a method of transporting linked data using JSON. It has 2 basic types, *Objects* and *Data* type. JSON-LD is designed around the concept of a "context", which

provides additional mapping from JSON to the RDF data model. The context therefore tells how to interpret the JSON document.

JSON-LD introduces the @Context syntax (W3C, 2014), which is used to define the vocabulary binding for the data concepts used in the JSON-LD document. For the purposes of this work, the context is also a set of rules for interpreting a JSON-LD document. Here the author proposes that JSON-LD context serves to enable the binding of a graph node to the information instance graph hosted on the backend supporting infrastructure. A context can be directly embedded within a JSON-LD document, or as in this case put into a separate document and referenced (shown in Listing 4.4 and Listing 4.5). In this work the context is used to link the data instance data to the actual instance hosted on the server.

```
{
  "@Context" : {
    "obj_store" : "coap://tpot.arch-
dev.ie/obj_store/",
    "obj_id" : {
      "@id" : "obj_store:obj_id",
      "@type" : "@id"
    }
    "at0002" : "obj_id:at0002/",
    "at0004" : "at0002:at0004/",
    "at0008" : {
      "@id" : "at0004:at0008",
      "@type" : "@id"
    },
    "DV" : {
      "@id" : "at0008:#at0009",
      "@type" : "@id"
    },
    "resultTime" : {
      "@id" : "at0008:#at0010",
      "@type" : "@id"
    }
  }
}
```

Listing 4.4 Extract from a JSON-LD representation. Information instance fragments are bound to archetypes/OPTs via the @Context. The @id represents the parental information instance of this observation_set. Where at0004 refers to an observation_set with the readings for at0008 (temperature) which is an observation fragment belonging to the sensor_data_record of {object_id} defined by the archetype {opt_id}. The URI fragment, denoted by the # symbol, denotes this the end of of the URI path. This is defined by the last aggregation level within the reference model

To reduce the size of the graph node, key-value pairs are represented using the at-codes defined within the archetype (Listing 4.2). Sundvall et al. (2013) have shown how archetype-based health record systems can be implemented through the application of a REST architecture. In the approach described, a similar methodology is employed to allow the binding of graph nodes to URIs; Listing 4.5 illustrates this.

Table 4.1 triple representing a temperature reading (the coap:// protocol shown in the URL is discussed later in this section)

Subject	Property	Value
`coap://tpot.arch-dev.ie/obj_store/{obj_id}/at0002/at0004/at0008`	`coap://tpot.arch-dev.ie/obj_store/{obj_id}/at0002/at0004/at0008/#DV`	10.23

The context (Listing 4.4) defines keys (of a key-value pair) and their corresponding context within a specific data graph. JSON-LD context definitions are hosted on the backend-support services infrastructure (Figure 4.8). Contexts are created from OWL representations of ADL based operational templates. Contexts are exposed using the RESTful architectural approach via a URI (Fielding and Taylor, 2002). This allows a Result node (Figure 4.10) to maintain its context within the data graph. Listing 4.4 shows the resulting node representation which a data buoy system must adhere too. In this simple example, a data value (DV) key has the value 10.23 (Listing 4.5). This data value is bound the JSON-LD context definition to its meaning using URIs composed of at-codes. At-codes are defined within the archetype (not shown here).

Table 4.1 shows the triple JSON-LD based representation of the value. When the data buoy system transmits the result, the supporting backend infrastructure can process the corresponding JSON object (or information instance fragment) using the JSON-LD context. The result of the backend processing step results in an information instance shown in Listing 4.5.

```
{
  "@Context" : "coap://tpot.arch-
      dev.ie/microcontexts/{microCtxt_id
      }",
  "obj_id" : "{sdr_object_id}",
  "@id":"at0008",
  "DV": "10.23",
  "resultTime": "<time_stamp>"
 }
```

Listing 4.5 Extract from a JSON-LD representation of Result (Figure 4.10).

4.5.1.3 Operational Templates as a Service

A core principle of the approach presented is this work is to enable the fragmentation of archetype-based instance data between a constrained system (data buoy) and backend supporting infrastructure and services. A RESTful architectural style has been adopted to enable the Linked Data paradigm. Fundamental to information instance creation in current archetype enabled system are operational templates, and a runtime template kernel. The federated graph approach described above requires a novel template kernel to support the creation of valid graph data nodes on the data buoy and the backend. The concept of Operational Templates as a Service (OPTaaS) has been developed in this work to support the overall federated approach and facilitate interactions between a micro-kernel and the federated template kernel. Figure 4.11 shows the interactions between the data buoy systems and OPTaaS component. RESTful interactions within a constrained environment require a great deal of overhead and may not be possible using traditional methods.

To support web services running on platforms with very limited resources the IETF formed the Constrained RESTful Environments group (CoRE) (Shelby, 2012). CoRE has been tasked with developing a framework for deploying web services to constrained environments, such as sensor nodes. In the CoRE framework, a network of nodes called devices interact. Devices are responsible for one or more resources, which could be a representation of sensors, actuators, and combinations of values or other information.

Devices in the network can send messages to each other to request, query and publish data. As part of the overall effort to enable constrained RESTful environments, CoAP was defined (Shelby et al., 2012).

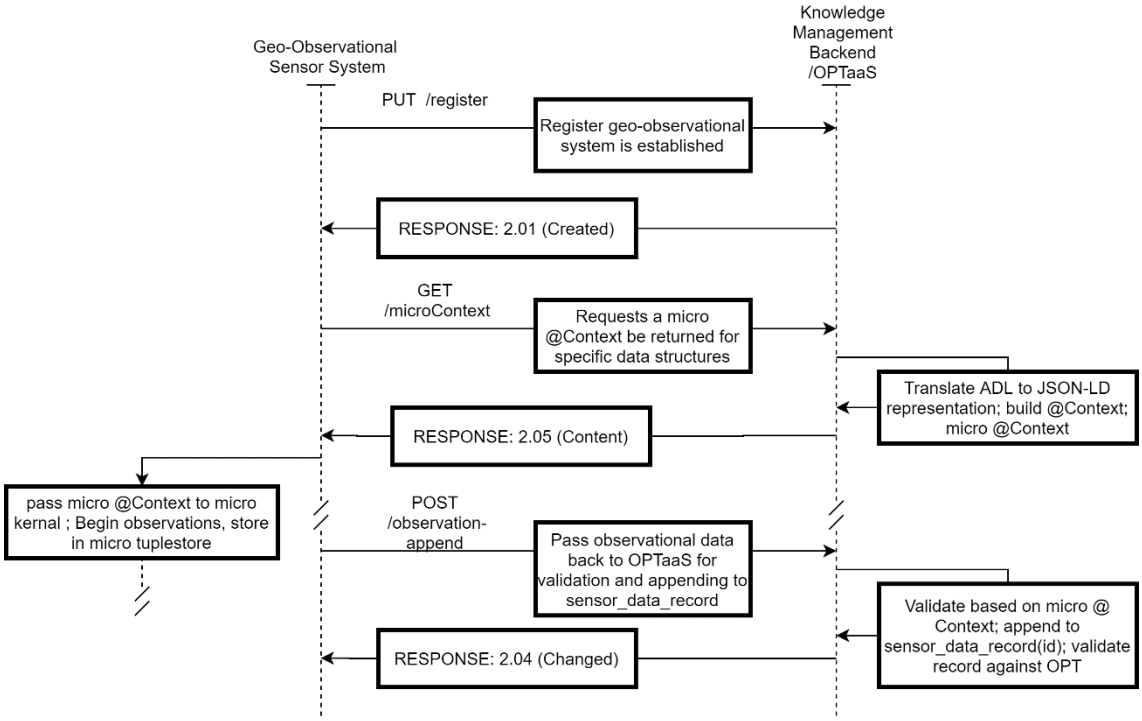


Figure 4.11 RESTful interactions between a data buoy and the operational template as a service (OPTaaS). The CoAP protocol is used for message exchanges. The OPTaaS holds the runtime templates and builds a fragment template as a micro @Context. The micro @Context is cached by the observational sensor system and used to perform prelimary JSON-LD processing prior to posting to the OPTaaS web service. The OPTaasS holds a run-time template for the observational system and performs full validation of the information instance as they are received.

CoAP is a specialised Web transfer protocol for use with constrained nodes and networks. CoAP provides a request/response interaction model between application endpoints. Unlike HTTP based protocols, CoAP uses UDP as its transport layer and employs a simplified re-transmission mechanism. CoAP is designed to easily interface with HTTP for integration with the Web with very low overhead and simplicity for constrained environments (CoAP and associated protocols are described in more detail in later in chapter 5 and Appendix D).

4.6 Limitations

Semantically annotating captured data at source is problematic in constrained systems. Born connected system mechanisms are computationally expensive, and in a resource-constrained environment, this may not be possible. Preliminary evaluation of the described technique has shown that semantically rich data objects can be supported using a Linked Data approach. However, this is a preliminary evaluation of the technique which has not been scaled to include additional reporting platforms or more complex real datasets. As Pottie (2013) observed, every bit transmitted brings a sensor node one moment closer to death.

The use of URIs to semantically enrich data objects can present an unacceptable overhead in some constrained environments. As mentioned previously in Chapter 3 (section 3.3) URI lengths are in general too long for packets in a constrained communication environment directly. The specified message size for a CoAP payload should be less than 1024 bytes to avoid IP fragmentation.

Codification of URIs have been proposed to overcome this limitation. The author is using the experience gained from the described evaluation to further constrain the technique described. The next stage of evaluation is to implement the technique on an constrained test infrastructures, with further constraints on communications and power (this is detailed in chapter 5 and chapter 6).

It is noted that triples are the base of the entire RDF knowledge model. Triples can be represented using many different formats. However, many of these are suitable for constrained systems due to computational constraints and limitations on packet size. JSON-LD has been shown to be an efficient serialisation mechanism for RDF based data. However more efficient approaches exist. Käbisch, Peintenr and Anicic (2015) have developed a promising approach that fulfills the following criteria: Low memory usage,

small message size, type awareness, simple processing, and a standardized solution. Their work uses the EXI format for RDF/XML data representation. XML interchange using EXI has been shown to be more efficient than JSON and binary JSON encodings (Hill, 2015).

4.7 Chapter Discussion & Conclusion

A data buoy software system architecture has been defined (Figure 4.9) to enable evaluation of the described technique. The goal of the initial evaluation is to verify the methodology described, with a core requirement to reduce the size of the data instance required on the constrained system, without comprising the knowledge infrastructure.

A test archetype was developed as part of a proof-of-concept exercise. An XML serialization of the O&M data model was produced. The LinkEHR editor was used to constrain the information model further for the implementation. An ADL representation of the test archetype was produced and stored within a simple archetype store. Community derived archetypes are hosted in the archetype repository in ADL format (Figure 4.9). Operational templates are used to further specialize archetypes for specific use-cases. For this implementation, operational template are assumed to be equivalent to the serialized archetype, i.e. no further specialization or constraining has been performed. The OPTaaS component requests the conversion of an OPT for use within the constrained data buoy test rig system. The Validation & Converter component retrieves the template from the template store in ADL format. This template is then converted to OWL format. The ADL file was translated to OWL using the technique described Lezcano, Sicilia and Rodríguez-Solano (2011). The library Owl2jsonld (Reyes, 2014) was used to produce a JSON-LD context from the OWL translation. The resulting JSON-LD representation was manually fragmented to produce a micro-context for the graph node Result shown in Listing 4.4. The micro-context store is made available via an URI.

A basic data buoy OPTaaS client application was created using node.js. A minimal backend supporting infrastructure was developed. The OPTaaS server was also implemented using node.js. The OPTaaS client and server both use the node.js based COAP library node-coap[49]. A basic runtime kernel and validator was developed, based on the openEHR Java Reference Implementation. The runtime template kernel component is used by the OPTaaS to process JSON-LD observations and resolve the triples to an RDF store. The OPTaaS server interacts with the sensor data record store (SDR Datastore) (Figure 4.9 & Figure 4.11) via a call to localhost. Apache JENA is used as the SDR Datastore.

Node.js is used to implement many of the system components at this point. This has allowed for rapid prototyping to allow evaluation of the proof-of-concept. However, it was found that a lot of manual steps had to be employed within the process. The evaluation has informed the remaining work detailed in Chapters 6 and 7. The openEHR Java Reference Implementation is specifically designed for openEHR archetypes. The LinkEHR editor is a multi-reference model archetype editor. LinkEHR developers have announced that LinkEHR will be made available as an open source project in the near future. For the current work, archetypes are created manually using LinkEHR, an open-source version of LinkEHR would greatly enhance the development described. The proof-of-concept work has also allowed the further specification of the components necessary for further constraining of the system in terms of technical constraints; these are outlined below.

4.7.1 Transformations

Today, the Poeseacle project has evolved from its origins through a number of projects and iterations. Poseacle -> ResearchEHR ->ArchMIS-> Clin-il-Links. The Poseacle

[49] https://github.com/mcollina/node-coap

approach differs from the approach offered by Lezcano. In his Doctoral thesis Lezcano defines two different ways to translate ADL definitions to OWL (Lezcano, 2012).

1. Translating as classes method. ADL definitions can be considered as ontology classes that specialise OWL representations of the reference model.
2. A different approach of translating archetypes as instances. Here archetypes are taken as instances of the archetype object model. In this approach in the clinical setting this leaves no room for patient data.

The Poseacle approach takes approach (1) above, whereas Lezcano takes approach (2) above. Approach (1) takes archetypes and translates them into instances of some classes representing an archetype model. The main objective is to facilitate semantic search at the archetype specification level, as well as other semantic tasks that improve EHR management.

For this work approach (1) aligns better with the application domain, in that enabling the ability to discover related geo-spatial information objects and federate them through semantic search is a key focus.

One possibility as proposed by Sharma et al. (2017) is to represent an OGC based reference model in XMI, as XMI to OWL translations already exist.

ADL is not precise, even though it is designed to be a formal language. ADL's specification present precision difficulties relating to the specialisation semantics of archetypes (Porn et al., 2015). Given the data instances defined by archetypes are also instances of the underlying reference model, understanding the relationship between the reference model and archetypes becomes crucial (Molando, 2009).

ADL is predicated on the existence of object-oriented reference models and the constraints in an ADLarchetypes are in relation to the types and attributes from such a

model. This must be considered while comparing ADL/OO formalisms versus RDF/OWL.

4.7.2 Augmented O&M Open Questions

The way in which the identity of feature-of-interest is modelled within the augmented O&M model needs further exploration. Within the model what is the documentation equivalent within an EHR? Perhaps it could be modelled as presented in Figure 4.12 below (Listing 4.6).

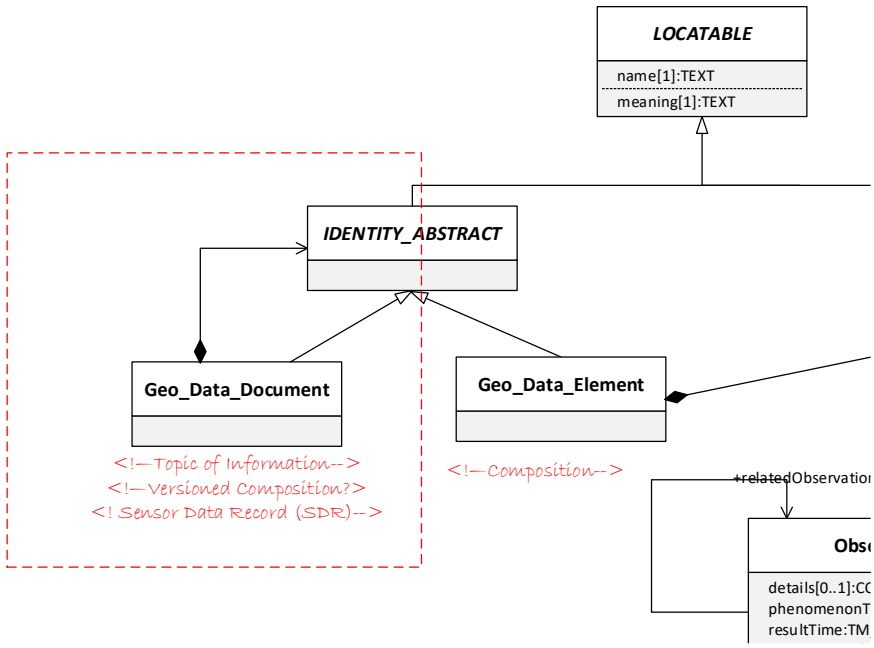


Figure 4.12 Modelling the concept of documentation within an augmented O&M model

```
<?xml version="1.0" encoding="utf-8" ?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema" version="v1.0.0"
targetNamespace="http://tpot.dit.ie"  xmlns="http://tpot.dit.ie">
        <xs:include schemaLocation="OM-dataTypes.xsd" />
        <xs:element name="identity_component" type="IDENTITY_COMPONENT" />
        <xs:complexType name="IDENTITY_COMPONENT" abstract="true">
                <xs:sequence>
                        <xs:element name="name" type="xs:string" />
                        <xs:element name="archetype_id" type="xs:string" minOccurs="0"
                              maxOccurs="1" />
                        <xs:element name="validity_time" type="TS" minOccurs="0" />
                </xs:sequence>
        </xs:complexType>
</xs:schema>
```

Listing 4.6 Identity Component modelled according to the GIRM identity component (Chen, 2016)

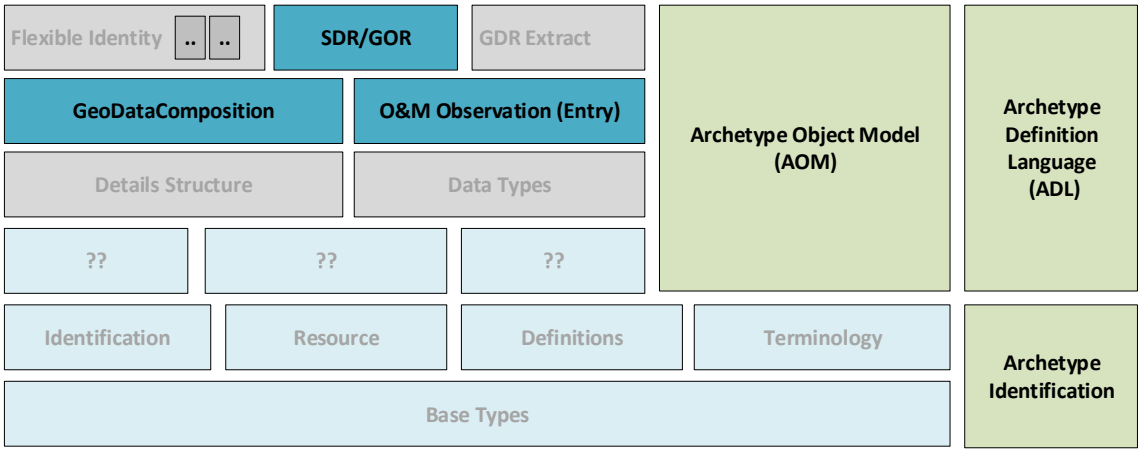


Figure 4.13 Sensor Data Record (SDR)/ Geo Observations Record (GOR) Information Model

Is the *topic of interest* Earth? Earth is the EHR Person/Patient equivalent. Demographics is "celestial body" perhaps? But may be too broad to be useful.

If the Earth is the role i.e. Patient in health domain. This gives further argument to the idea of a flexible identity model. Where we may want to make the heart the topic of interest in the health domain, we may want to make Ireland's weather the topic of interest within the Geo domain. Weather report then becomes the document or Informational unit that is to be used for sharing or communication etc.

Additional considerations are listed below:

- The relatedObservation recursive relationship in O&M, where the role is ObservationContext, is probably now captured with the addition of Observation_Set. Observation instances belonging within the same Observation_Set are related, therefore the context is implied by ObservationSet which is a section. Is this semantically the same meaning as is defined within the O&M model? The observation context may already be defined by the identity model.
- Within INSPIRE (INSPIRE 2016), FeatureOfInterest is defined as:

  "This is a representation of the real world object the property is being estimated on. The following terms are used to refer to the Feature-Of-Interest in other domains: ***Earth Observations****: 2-D swath or scene; 3-D sampling space.* ***Earth science simulations****: Section, swath, volume, grid.* ***Assay/Chemistry****: Sample.* ***Geology field observations****: Location of structure observation; Rock sample"*

  Also:

  "The Observation model takes a user-centric viewpoint, emphasizing the semantics of the feature-of-interest and its properties." I think this view which is in contrast to sensor oriented models gives weight to the argument that O&M is a good candidate as a reference model for a dual-model approach. The user is interested not just in the observation but the higher observational context, or the documentation of the observational context, all the way to the knowledge level i.e. "weather report".
- The question of coverages arises here:

"Many observations are made to detect the variation of some property in the natural environment, expressed as a spatial function or field, also known as a coverage (ISO 19123:2005)"

Does Observation_Set in the proposed model (which is a "Section") allow for the generation of the coverage concept as an Archetype. This needs to be investigated in the context of the GEOSS Architecture.

Having performed a limited proof-of-concept implementation to investigate, validate and refine the translation methodology defined in this chapter, the next chapter describes the further development of the infrastructure required to support the dual-level across technologically resource constrained systems.

Chapter 5

"Every bit transmitted brings a sensor node
one moment closer to death"
(Pottie, 2003)

5. A RESOURCE CONSTRAINED KNOWLEDGE FRAMEWORK

Chapter Overview: The previous chapter presented a novel translation of two-level modelling appropriate for the geo-spatial domain. As highlighted previously, the approach introduces an increased level of processing overhead due to additional metadata requirements. Given many in situ geo-observational platforms are technologically constrained (as discussed in chapter 3) a linked data approach to federate data across geo-observational systems was proposed (chapter 4) as a solution. This chapter presents the reader with an implementable framework design and infrastructure solution that can support the novel methods described in chapter 4. The primary aim of the implementation is to validate the concepts presented in chapter 4.

As in chapter 4, literature review material is again referred to throughout this chapter as part of the assess/refine iterative design science research methodology (see chapter 4: *chapter overview*).

The resource constrained knowledge framework solution described here facilitates the deployment of two-level modelling approaches within constrained geo-observational-systems, to the *edge*. Figure 5.1 below presents a (UML) deployment view of the system described within this chapter. The technical details of the system are shown in Figure 5.1 and will be described throughout this chapter. Figure 5.1 provides the reader with a bird's eye view of the overall infrastructure in deployment view, while the particulars of each node/component and artefact are presented in more detail throughout this chapter.

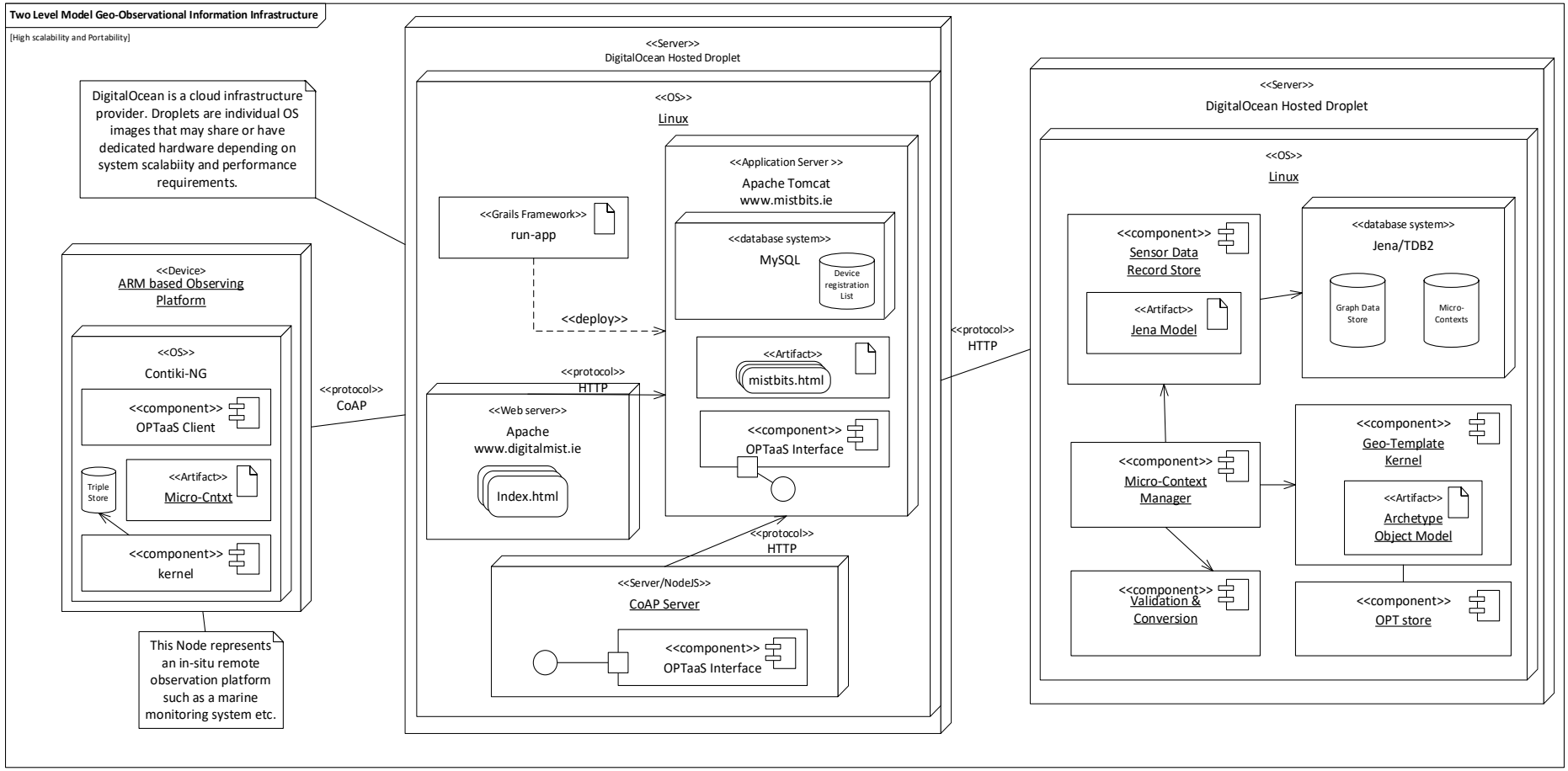


Figure 5.1 System Deployment Diagram.

The approach described here (and supported by the infrastructure shown in Figure 5.1) aims to reduce processing overhead at the edge of the network (i.e. observing platforms), while extending data quality - provenance, completeness, findability[50] and the foundational principles of *FAIR* (Coetzee et al., 2020) - to the edge of the network. The system design, implementation, deployment and validation described throughout this chapter investigates the viability of the concepts and translation methodology presented in chapter 4. The geospatial domain and in particular remote in-situ system deployments present many barriers to the adoption of a two-level modelling approach. In order to validate the ideas developed within this work a build/evaluate cycle is used within an overall design science methodology (see chapter 1, section 1.6). It is not intended that the resultant software components are production ready for real world scenarios. However, the build/evaluate cycle described here provides the basis for realising real systems and confirms the suitability of the approach within the geospatial domain. This is discussed later in this chapter in section 5.5 and in chapter 7.

The core enabling component of the framework - *the knowledge kernel* – (geo-template kernel in Figure 5.1 and Figure 5.2) has been developed and built on top of the Java Reference Implementation open source libraries from the openEHR foundation (Chen and Klein, 2007). To provide the reader with a comprehensive understanding of the knowledge system framework, firstly, the system's initial design considerations are presented before the detailed technical implementation is described.

5.1 System Design Considerations

Taking a user-centric design view, the purpose of the resource constrained knowledge framework that is outlined in this chapter is to facilitate two main actors, *domain experts* (for example geographers, oceanographers, marine scientists etc.) who are not necessarily

[50] https://www.w3.org/TR/sdw-bp/#describing-dataset-structure-and-service-behaviors

specialised in ICT, and any connected *constrained observational platforms* (referred to as *Domain Expert* and *System* in Figure 5.2 below).

The framework facilitates a domain expert's participation in a community consensus approach to fine-grained constraining of general reference models (ontological principles level) into archetype models. Thus, ensuring a highly flexible and structured approach to information modelling can be performed by the domain experts themselves for specific observing scenarios. As discussed in chapter 3, semantic interoperability requires *completeness* within the data or information sets (see section 3.2). Completeness is difficult to define as the attributes required to achieve completeness are not always immediately obvious. As in Popper's 3 worlds, attributes may lie outside the physical world and are often not tangible and therefore completeness is best captured by domain experts. The framework must enable the convenient sharing and consensus-based revision of a managed library of archetypes, with which domain experts can then use to build operational templates (used by real systems), based on the idea of completeness.

Domain experts can participate in the framework using a visual editor that supports at a minimum the O&M based reference model, profiled for a two-level modelling system design methodology (as described in chapter 4). The domain expert uses an archetype editor to propose new archetypes, edit and specialise existing archetypes within the archetype library. The domain expert may also use an editor to build operational templates, which define the required information structures to be used within a given system or scenario of use.

The framework provides a *geo-templating kernel* (see Figure 5.2 below) that supports the profiled O&M model (see section 5.5), parses a given operational template and creates an in-memory O&M based archetype object model (AOM). Using the AOM, the kernel creates *micro-contexts* (proposed previously in chapter 4 and described in more detail

later in this chapter) of the operational templates. Micro-contexts then provide light-weight contexts that constrained observational platforms can use locally to create and link federated (Sheth and Larsson, 1990) sensor data streams that adhere to the O&M reference model and the archetype model.

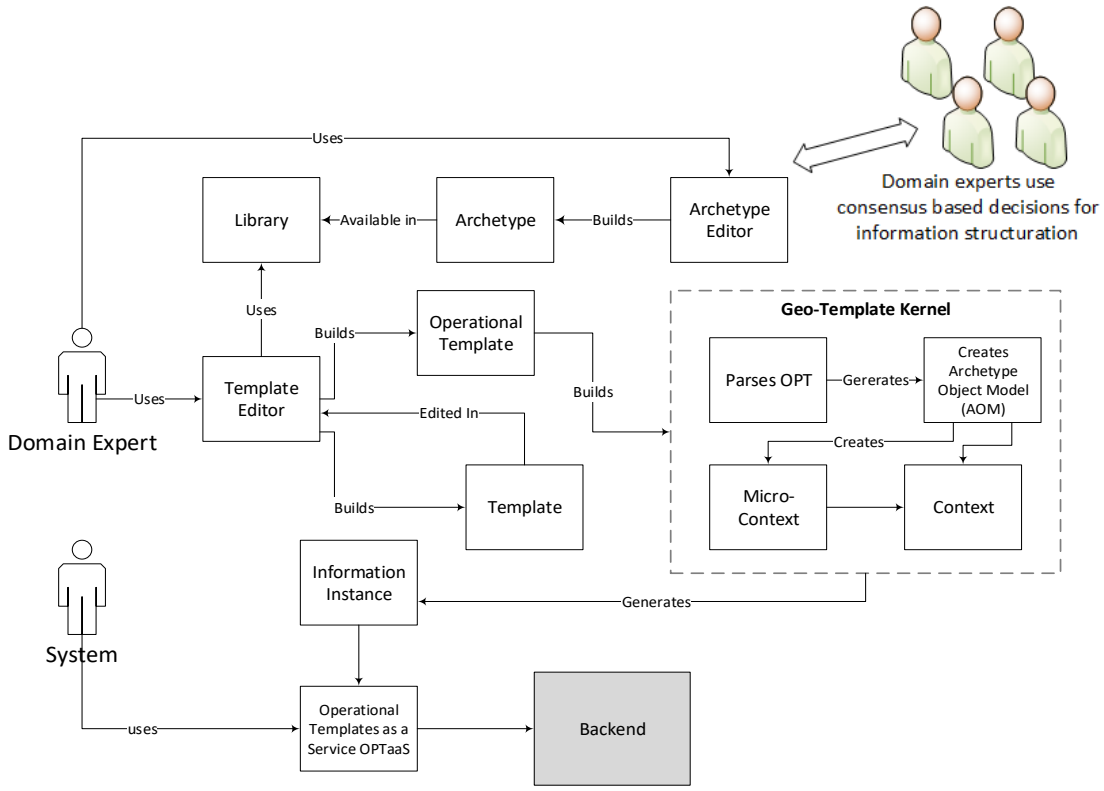


Figure 5.2 Constrained Knowledge Framework System Level View. The System actor represents a participating resource constrained system such as an in situ remote ocean observing platform.

Observational platforms (system actor in Figure 5.2 above) participate in the framework using an operational template as a service (OPTaaS). The OPTaaS is a concrete implementation used to validate the platform-to-backend reporting interface approach described in chapter 4 (Figure 4.11). The OPTaaS acts as a message broker between the observational platform and the backend persistence layer and ensures robust management of the federated data-streams and ensures adherence to both levels of the dual-level model (reference model and archetype model).

The design is a balancing act of maintaining the enhanced data quality and semantic interoperability afforded by the two-level modelling approach with the constrained environment with which it is to be deployed. To that end, a core part of the framework is that of the *geo-templating kernel* (Figure 5.2).

The kernel is shown in Figure 5.2 above in the context of the two main actors, who participate within the knowledge framework. The kernel supports the federating of knowledge artefacts across the information system infrastructure, i.e. between the platform (observational in situ node) to the backend persistence layers (simplified as "Backend" in Figure 5.2, expanded in Figure 5.7 below, and described in more detail in section 5.2 below). Core to this approach is how operational templates are managed within the system. This is also discussed in detail later in this chapter.

Ultimately the framework and platforms should enable pervasive environments to exist within remote deployments, where constrained platforms can cooperate and exchange knowledge and facts directly or between a centralised broker (the server). Within this framework *mist computing* & *fog computing* paradigms are employed (introduced in section 2.5) to enable knowledge exchange. This is discussed in more detail next.

5.1.1 Framework Definition

For this validation work, and to aid clarity of discussion, the framework is made up of two constituent parts, which the author has labelled the *DigitalMist* and *MistBits*. The author defines *mist* in this context as "meaning making in smart things" (Figure 5.3).

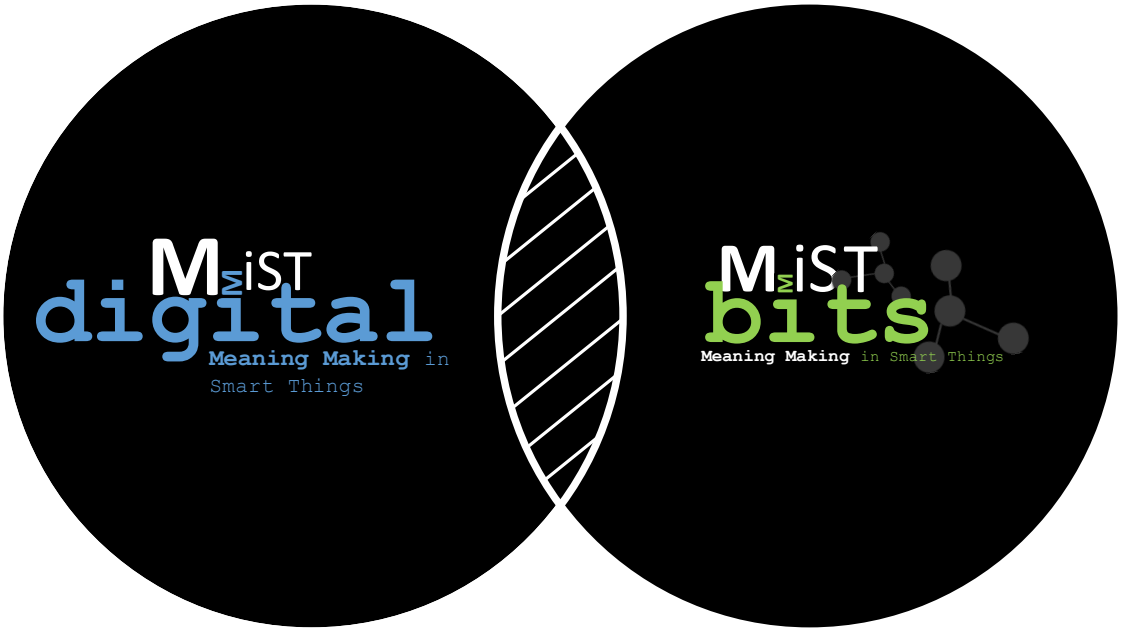


Figure 5.3 DigitalMist and MistBits Framework Components, the intersection between both system components occurs at the OPTaaS.

The *DigitalMist* refers to the backend deployed knowledge engine framework which is hosted within the cloud (see sections 5.2 & 5.4, and Figure 5.7 below). *Mistbits* refers to the individual nodes (or observing platforms) participating within the framework and the backend software that coordinates the individual nodes that register and participate within the DigitalMist. Both the DigitalMist and Mistbits software components overlap at the point of the OPTaaS implementation, where they are both tightly coupled (described in more detail in section 5.2). Core to both DigitalMist and MistBits are the knowledge kernels developed to support federated, semantically interoperable observational data, these are described in more detail throughout this chapter.

5.1.2 Componentisation & Separation of Concerns

Here, the high-level design considerations described above, and the resulting design problem are broken down into the various components needed to realise the overall framework. Each component performs a defined task to realise the overall system functionality and provides an integrated validation approach for the overall translation

approach described in chapter 4. Each system component is briefly described in turn below.

5.1.2.1 Communications

Communication between components is HTTP driven with some additional event-driven messaging via the CoAP protocol and other messaging services. In principal communication can happen in many ways, however as discussed in section 4.5.1, a RESTful design approach is adopted here, which has led to the choice of HTTP and CoAP (see Figure 5.1).

5.1.2.2 User Interfaces

User interfaces are primarily Web based implementations, using HTML, CSS and JavaScript based technologies. These technologies were chosen to ensure cross-platform compatibility and ease of deployment across different platforms (see Figure 5.1).

5.1.2.3 Data Validation & Conversion

Data validation and conversion occurs within the backend using Java based libraries (see section 5.2 below). At the node level C based libraries are used due to the embedded operating systems employed on the technologically constrained platforms (see section 5.3 below).

5.1.2.4 Storage and Querying

Data are persisted within the overall framework in different ways depending on where storage needs to happen and in what format the particular storage and associated needs are (see Figure 5.1). For instance, the relational database MySQL is used to store table-based data and SQL is used for querying device registration details (Figure 5.1). Where the storage of RDF and *Blobs* is required, Jena-Blob[51] is used (see Figure 5.1). Jena-Blob

[51] https://github.com/bluejoe2008/jena-blob

is chosen because Apache Jena is chosen to support a linked data approach to the distributed data approach used. Jena is discussed in more detail below. On observing platforms, an embedded relational database solution is used, again this is discussed in more detail below and shown briefly in Figure 5.1.

The ability to query data is directly related to how the storage of data is defined. There are a number of query-able levels possible using different storage patterns[52]. Much of the implementation decisions around storage and querying are somewhat heuristic, and the real impacts of these decisions would require additional investigation which is outside the scope of this work. However, the enhanced ability to query the data referred to as *semantic querying* is a key objective of two-level modelling and the approaches described here. Semantic querying considerations used within this work are discussed next. However, the reader should be aware that further work is needed in this area in the future to optimise the benefits of the approach investigated here. Within this evaluation and validation work the aim is to show how semantic querying can be supported within the overall translation approach.

5.1.3 Semantic Querying

The use of a common reference model & archetypes ultimately enables better semantic interoperability between systems. However, a method to query the information is essential. A key benefit of semantic interoperability is the ability to perform enhanced semantic querying of datasets (chapter 3).

Within the O&M based reference model, instances of the class *Results* (Figure 4.5) are what is referred to as an *element instance*. Element instances are the lowest level within the data structure to be associated with a universally unique identifier (UUID) (Leach, Mealling, and Salz, 2005). The lowest constraint level is the *Details_COMPOUND* level

[52] https://openehr.atlassian.net/wiki/spaces/dev/pages/6553626/Node+Path+Persistence

which may be identified using a combination of the UUID and a *path* (Figure 5.4 and Listing 5.1 below).

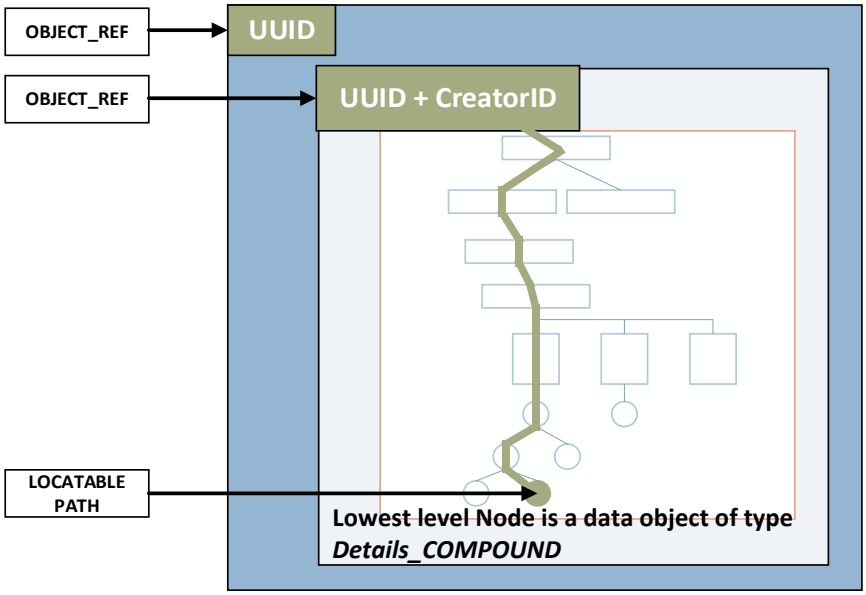


Figure 5.4 Visual map of how a combination of UUID and locatable path can be used to identify and retrieve data instances at the level of Details_COMPOUND.

Within a linked data approach, the element level should be considered as the lowest level for a data node. After that, BLOB data, which is structured using constraints against the lower levels and the reference model structures exist. BLOB data may be further traversed using additional querying such as XQuery/XPath if represented in XML format or for example a JSON parser or JQuery if JSON is used to store BLOBs (Listing 5.1). BLOB storage solutions are discussed in more detail below.

```
"id" : "identity_model_ref_ID",
 "GeoData_Composition" : {
      "archetype_node_Id" : "TPOT-OM-
GeoData_Composition.Weekly_Buoy_Data.v1",
      "name" : "Marine_Data_Buoy_Weekly_Report",
      "details_COMPOUND" : {
            "details_ELEMENT": {
                  "archetype_node_Id":"[at0001]",
                  "name":"buoy_location",
                  "DATA_VALUE":"53.127,-11.200"
            }
      },
      "GeoObservation_Set" : {
         "archetype_node_Id" : "[at0002]",
         "name": "Buoy_Instrument_Readings",
           "meaning": "Interval Trig Buoy Multi Instrument Read",
           "details_COMPOUND": {
            "details_ELEMENT":{
                  "archetype_node_Id":"[at0003]",
                  "name":"triggertime",
                  "DATA_VALUE":"2017-01-11T11:40:00.000Z"
            }
         },
         "Observation" : {
            "archetype_node_Id" : "[at0003]",
            "name" : "observation_measure",
            "observedProperty":{
                  "archetype_node_Id" : "[at0004]",
                  "name" : "temperature"
```

Listing 5.1 Information Instance with UUID used to identify at the BLOB level.

Wang et al. (2005) note that one of the major barriers to widespread adoption of the openEHR two-level information modelling methodology is the lack of practical persistence solutions. In their work, they demonstrate how an archetype to relational database mapping can be achieved. Their results show comparable performance to that of standard relational databases within clinical settings. From their work, it is evident that relational databases can form part of the persistence backbone solution for dual-level modelling approaches. These findings do not map fully to the design scenario presented here. Any persistence solution must take in to account the highly federated nature of the selected system design and deployment (see section 5.2). Several design questions were considered while arriving at the final solution. For example:

- It is necessary to store data in chunks, to improve efficiency. However, what designates a chunk for the current application domain?
- The JENA framework enables the use of BLOBS and the use of *bags*. Can this act as an appropriate solution to the need for chunk storage?
- In what format should blobs be stored? For example, are JSON instances appropriate here?
- Should a blob occur at the ENTRY or COMPOSITION level?
- How are BLOBS identified? Should a BLOB be allocated a UUID if the BLOB is at COMPOSITION level, with paths being used to navigate within the BLOB or sub informational levels?

These design considerations are dealt with in further detail later in this chapter (see section 5.5).

5.1.3.1 Archetype Query Language

Systems that use archetypes may also use the Archetype Query Language (AQL). AQL queries are expressed based on semantics defined within the archetype level. AQL is a declarative language, it is applied to both the reference model and the archetype model.

An AQL query statement may be scoped within a particular record/geo-data-document or all documents based on a particular archetype. A Class expression is used within a FROM clause to achieve scoping. An AQL query snippet is used to discover all ocean observing platforms observing within a certain region in the North Sea (Listing 5.2).

Using AQL a fined grained automatic assessment of newly discovered data-flows relevant to an application can be made. This is enabled by the rich metadata associated with each information object, standardized to meet the community agreed constraints. This is referred to as the findability of the data (W3C, 2017b). The defined framework (shown in Figure 5.1) does not support AQL yet. However, the OPTaaS infrastructure

(section 5.2.2) uses a linked data approach to build information instances. In the OPTaaS backend, archetypes are represented using OWL (converted from ADL). Archetype/OWL governed documents are captured as knowledge graphs and SPARQL endpoints are available. SPARQL endpoints enable the knowledge graph to have a presence on a HTTP network, i.e. they have an associated URL and are capable of receiving SPARQL based requests.

```
SELECT c/…/wmo_platform_code
FROM GDR [include specific scoping here]contains
       GeoData_COMPOSITION c [TPOT-OM-
GeoData_COMPOSITION.platform.oceanobserving.v1
contains OM-Observation_Set […]
contains OM_Observation obs [TPOT-OM-OM_Observation.oceansitesObs.v1]
WHERE  obs/data[at0001]/details_COMPOUND[at0002]../items[at004]/value = “hourly”
```

Listing 5.2 Archetype Query Language (AQL) snippet

Dentler et al. (2012) have shown how archetyped SPARQL queries may be constructed using quality indicators, this will be considered in more detail in chapter 6.

5.1.4 Additional Two-Level Modelling System Components

Within health, there are several vendor specific operational two-level modelling systems. Each system contains many of the same generic core components and are largely based on libraries and standards developed by the openEHR foundation. Each of these generic core components are outlined below, including how they are realised within the final build of the constrained knowledge framework shown in Figure 5.2.

5.1.4.1 Terminology Servers

As described previously (chapter 3), two level modelling requires 3 distinct artefacts: reference models, archetypes and terminologies/ontologies. Terminology servers enable the binding of concepts at the reference and archetype levels to enhance interoperability and standardisation within the wider domain community. In health there are many mature terminology services and servers that can be used for concept binding. For example,

SNOMED-CT[53] is one of the most comprehensive clinical terminologies in the world. SNOMED-CT also provides terminology services.

Term-binding within two-level modelling systems happens primarily within the archetype, where the binding becomes a constraint on some node or data point, or it can happen within a template if there are local term definitions (see section 4.2.2). In ADL this happens by binding an at-code (such as at0006) to a specific term within a terminology such as SNOMED-CT using some coding standard. For example if the data point at0006 had the meaning of *activity of daily living*, the code can be bound to the SNOMED-CT activity for daily living code `SNOMEDCT::129818000`[54].

Within the sub disciplines of Earth system science there also exists many mature and stable terminologies and associated terminology services. For example, within the Oceanography domain the NERC vocabulary service exists (Leadbetter, Lowry and Clements, 2012). As shown in chapter 4, NERC does contain the necessary functionalities to facilitate term binding to ocean observing based archetypes (Listing 4.2).

5.1.4.2 Archetype Library

In order to manage Archetypes & Templates a clear mechanism for publishing & governance is needed. Within the health domain this is referred to as the clinical knowledge management system[55] (CKM). An equivalent tool is required here. To ensure generality, this will be referred to as a "Domain Knowledge Management tool" (DKM). A DKM should act as a benign dictator; consensus is the ideal, but not always realistic. For this work a GIT repository acts as the Archetype Library, hosted on GitHub[56] (Figure 5.5). This is an interim solution to serve as a DKM. GitHub allows the controlled storage,

[53] http://www.snomed.org/
[54] http://bioportal.bioontology.org/ontologies/SNOMEDCT?p=classes&conceptid=129818000
[55] https://ckm.openehr.org/ckm/
[56] https://github.com/pstacey/geo-archetype-library

retrieval and editing of archetypes. Also, governance is managed by the repository collaborators manager.

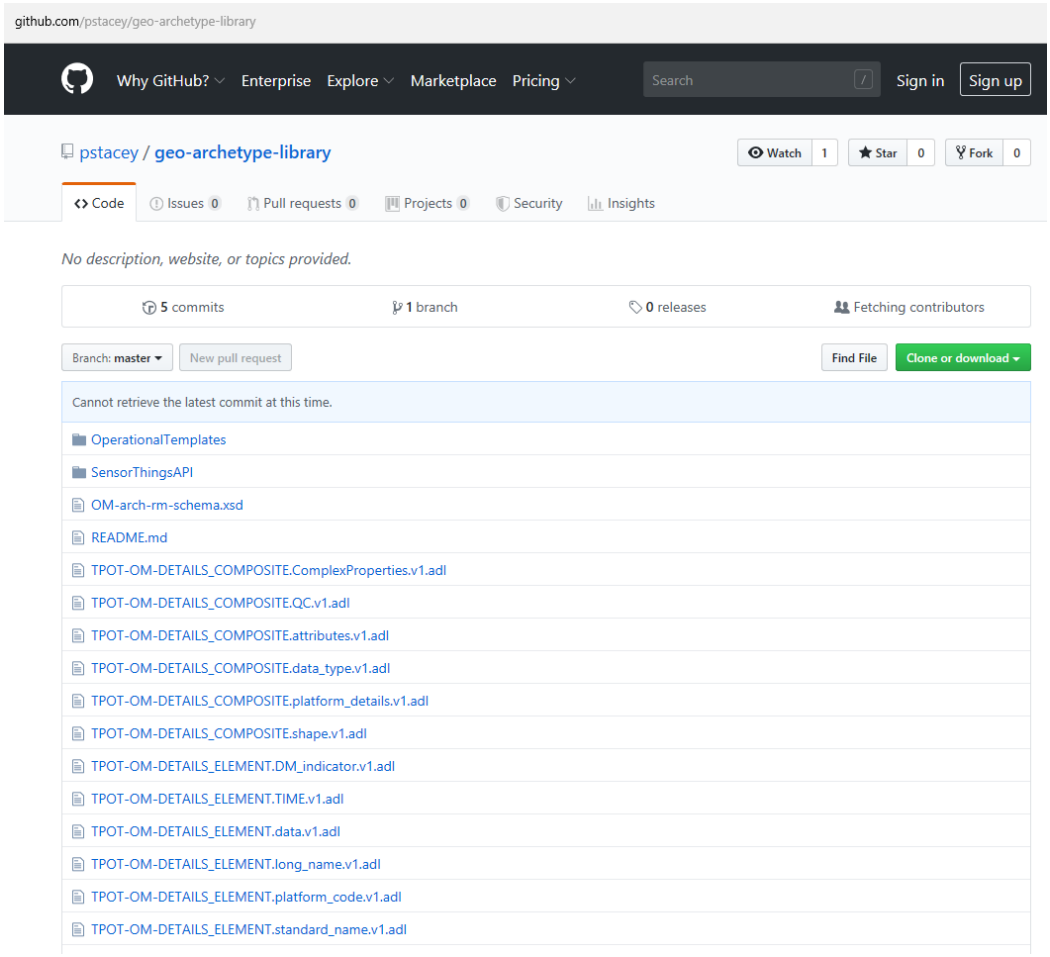


Figure 5.5 Screenshot of GitHub Archetype Repository used for this work.

5.1.4.3 Archetype Editors

The OpenEHR foundation maintains a list of products and tools related to two level health informational modelling[57]. Archetype editors are a key requirement to ensure domain practitioners can visually create, edit and specialise archetypes in a user and non-expert friendly way. The latest version of ADL is ADL 2.0[58] (as of December 2020). Support for ADL 2.0 is limited, with most editors supporting ADL version 1.4. Both ADL

[57] https://www.openehr.org/downloads/modellingtools/
[58] https://specifications.openehr.org/releases/AM/latest/ADL2.html

Workbench and Archetype Editor are developed by the OpenEHR foundation and support the OpenEHR standards. Ocean Health systems has developed a Template Designer. However, as already discussed in section 4.4.3, LinkEHR from VeraTech for Health is the only editor that allows the loading of 3rd party reference models in XSD format upon which archetypes can be defined. ADL workbench by Ocean Health Systems does allow additional reference models to be defined, but these must be defined using BNF notation (Backus–Naur form) (Naur, 1960). Using the serialized version (see Appendix A) of the augmented O&M reference model presented in chapter 4, LinkEHR can be used to perform archetype modelling. While the archetype library or github based DKM used provides change management, LinkEHR is used to create and edit the actual archetypes against the reference model. LinkEHR can also produce operational templates which may be in turn used within the final software solution.

Figure 5.6 below shows the relationship between reference model concepts, archetypes and templates. Archetype definitions typically constrain points within the reference model. As shown below, an archetype may provide a further constraint of GeoData_Document which is in effect defining a new knowledge level, or domain specific concept. The operational template on the other hand may encompass a whole collection of defined archetypes, which is subsequently serialised and consumed by the domain specific application software (built on top of the geo-templating kernel shown in Figure 5.2).

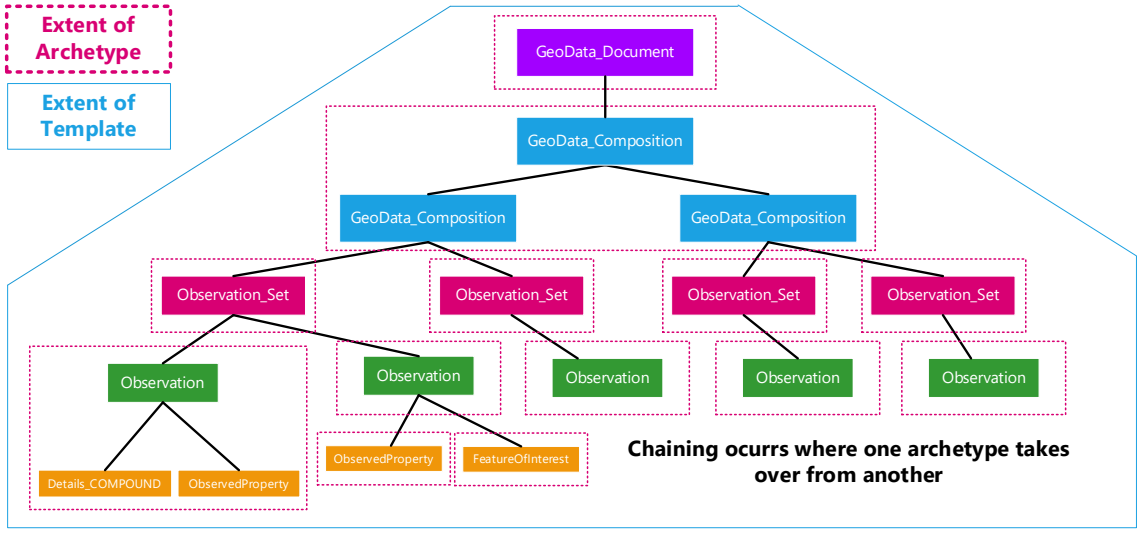


Figure 5.6 Operational Templates Extent.

5.2 System Architecture (Solution)

The goal of the system architecture of the constrained system is to provide a base constrained dual-level knowledge framework and reference architecture that can be used to support the development of two-level information model-based systems and applications within the geo-observational systems domain. Any production ready system, for a given application area will consist of a core knowledge management kernel, domain specific adaptions and associated applications. The main aspects of the system software architecture are described next.

Figure 5.7 below shows a high-level (UML) component diagram showing the software architecture of the constrained knowledge framework that is being evaluated. The architecture supports the activities shown in the system level view (Figure 5.2) by both of the main actors (domain expert and system).

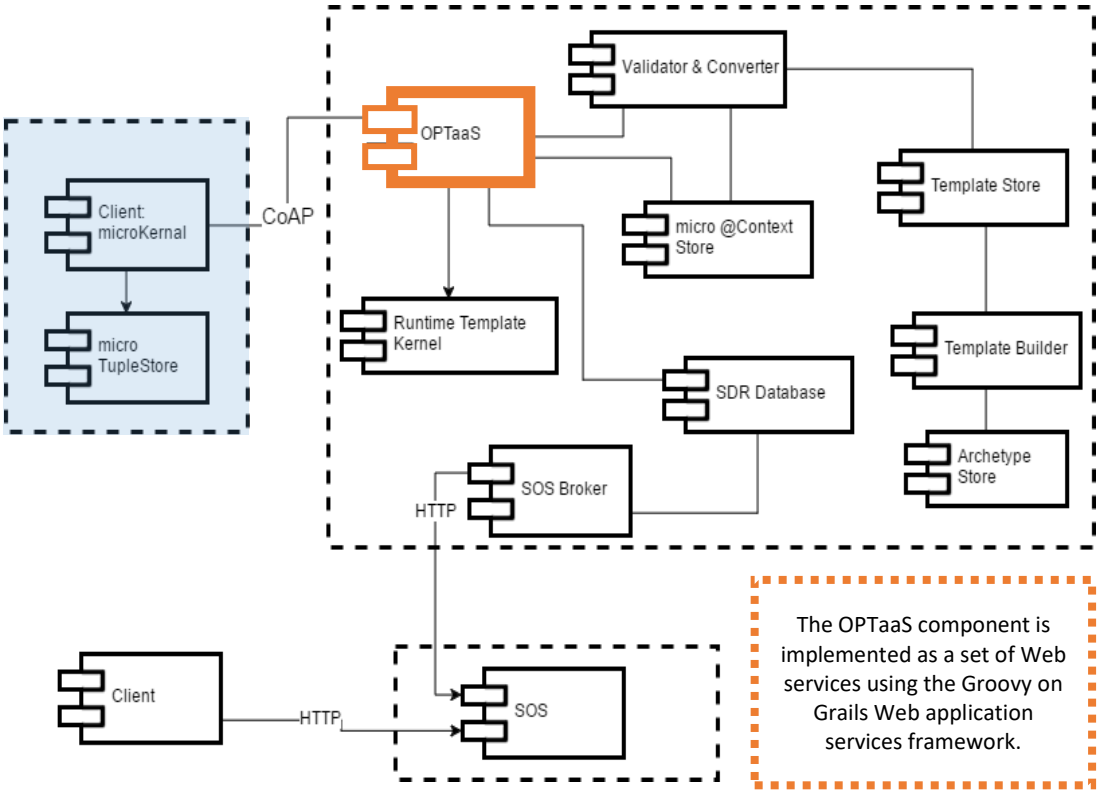


Figure 5.7 Component Diagram of Software Architecture. The OPTaaS component is highlighted in orange. A registered and reporting observing platform is highlighted in blue. The SDR (Sensor Data Record) Database represents the persistence solution for observational data records. The SOS (Sensor Observation Service) broker represents a SWE SOS compliant interface to retrieve observations from.

To implement the linked data approach (micro-context component shown in Figure 5.7) described in chapter 4, Apache JENA is used. An overview of Apache Jena was provided in section 3.2 and Figure 3.5 and its use is discussed in more detail below.

Groovy on Grails is chosen as the Web Application Framework to implement the OPTaaS component described in chapter 4 and shown in Figure 5.6 and Figure 5.1 above. Groovy on Grails is chosen as it is Java based and thus is in keeping with the implementation language and development environment chosen for supporting archetype-based systems etc. discussed next.

The main third-party API used in this implementation is the OpenEHR Java Reference Implementation (Chen and Klein, 2007). The OpenEHR Java reference implementation is an open source collection of Java packages that provide the base classes to build

OpenEHR two-level modelling-based information systems. The reference implementation supports an implemented reference model specification *openehr-rm*, which includes openEHR specific common classes, data structures and supports archetype-based object creation using the *archetype object model* (AOM) package *openehr-aom*.

The AOM is used as the basis for building software that presents archetypes and templates independent of how they are persisted or represented in a data store. The AOM package within the Java Reference Implementation is specific to the OpenEHR standard, and so needs to be adapted or re-written to support other domains (in this case the O&M augmented reference model). As discussed in previous chapters, an augmented O&M reference model has been developed as the reference model of choice to support this work. Therefore, the kernel shown in Figure 5.7 must support the creation of object instances against the O&M re-profiled reference model. This is achieved by adapting the OpenEHR Java reference implementation for the O&M based reference model described in chapter 4. Next, the implementation of this O&M based dual model kernel is described.

5.2.1 O&M Based Dual-Model Kernel Implementation

Although the OpenEHR Java Reference Implementation contains a wealth of reusable components, many of the core software components are tightly-coupled with the OpenEHR reference model implementation. This is to be expected in two-level modelling information systems, as hard coded software elements should rely on the stable reference model, without fear of obsolescence. This however presents difficulties for migrating the approach to other domains, with differing reference models.

For this work, where possible, generic software components have been reused. However, this re-use is quite limited, and many domain specific components have had to be written to support the O&M augmented reference model. For example the core

libraries for working with and producing template objects are specific to the OpenEHR standard, namely the `openehr.v1.template` libraries[59]. These libraries are available as JAR files, but these libraries are closed source and replacement software components have had to be developed here. These resulting packages developed as part of this validation work are detailed below from a static view of the system followed by a selected set of core runtime functionalities.

5.2.1.1 Package Overview[60]

The O&M dual-level model kernel that was developed to validate the application of two-level models on constrained "edge" embedded sensor platforms is described below. For brevity, selected key high-level packages and package structures are described only.

Selected runtime operations of the dual-model kernel implementation are described next. Several important detailed runtime sequence diagrams of the core package operations are presented along with selected code snippets to highlight key implementation detail to the reader to highlight how the approach has been validated through implementation.

The O&M dual model knowledge kernel is invoked with the creation of the DigitalMistMgr & JenaManager classes (Figure 5.8). The method `create()` within the SkeletonGenerator class takes an argument of type OperationalTemplate. The OperationalTemplate object contains a Map and List of all archetypes and attributes resulting from the parsing of an XML based .opt document. `The create()` method in turn calls the `createComplexObject()` method which constructs an RMObject, governed by the underlying reference model data structures, constrained by operational template.

[59] https://github.com/ethercis/ethercis/blob/master/libraries/openEHR.v1.Template-1.0.1.jar

[60] Packages are appended with the namespace *tpot*. tPOT refers to the research group "towards people oriented technologies" which is the research group within TU Dublin where this work has been carried out.

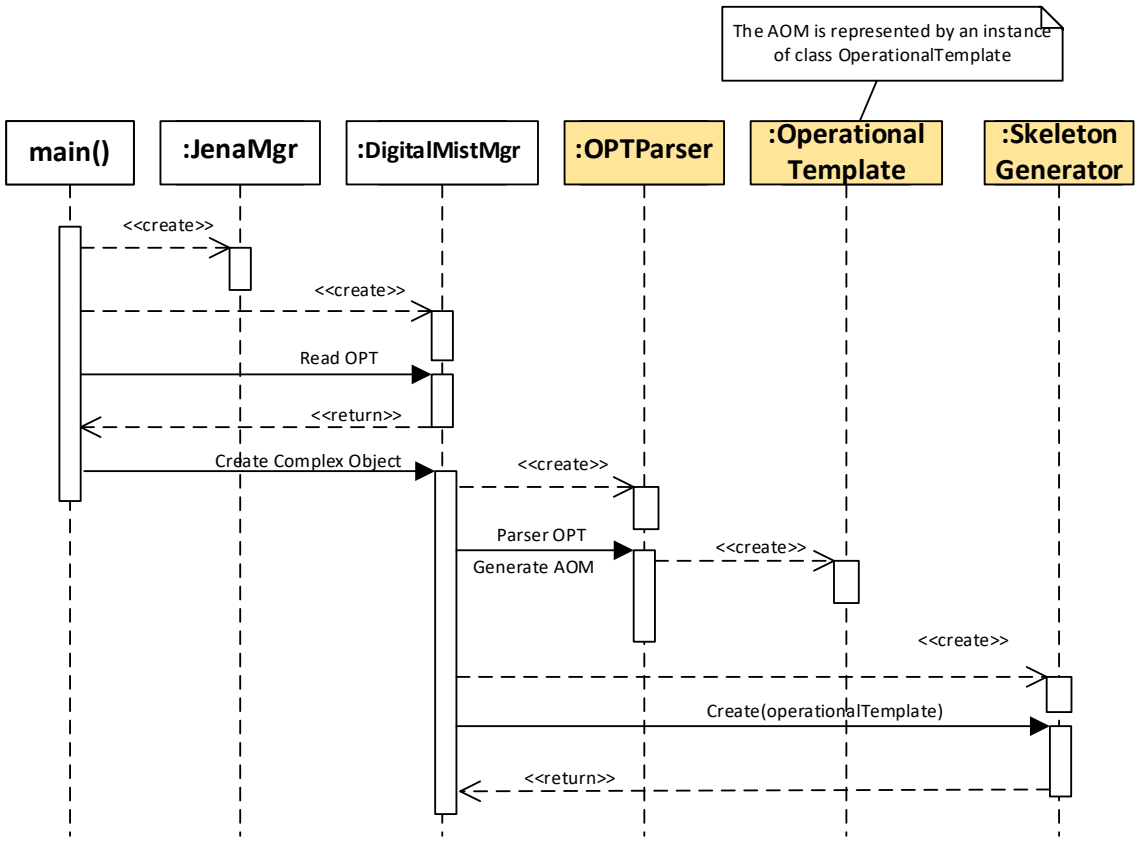


Figure 5.8 UML Sequence Diagram. Kernel initial runtime operation. The presence of the AOM shows the high-level view of how the system creates an in-memory representation of the constrained data. The system is hard coded against the underlying stable reference model. Only instances of the type reference model may be instantiated, however they are constrained at runtime against the AOM, which is runtime representation of the system archetypes that govern information object creation (described in chapter 4). The classes highlighted in orange are part of the *Runtime Template Kernel* component shown in Figure 5.7

5.2.1.2 Package:tpot.archdev.rm.core.util

The *util* package contains the class *SkeletonGenerator*. SkeletonGenerator contains methods to create an object tree of object types adhering to the Reference Model and the Archetype Model. This class uses the flattened operational template described in section 3.5.4 (also see appendix C for an example of a flattened OPT generated as part of further evaluation work presented in chapter 6) and resulting Template Object Model (TOM) to construct the in-memory object tree (Figure 5.9). The object tree allows the system to build complex information objects, independent of the persistence layer. For this work the persistence layer is a linked data graph (as discussed previously in section 4.5.1)

implemented within the cloud based backend server (see package *tpot.archdev.persistence* below). However, according to the edge-inclusive design approach adopted in this work and outlined in section 4.5.1, the linked data graph is distributed across the entire observation infrastructure in a federation of triple stores (Figure 5.1). A complex object is an instance of a reference model object created according to the constraints defined within a given archetype or operational template.

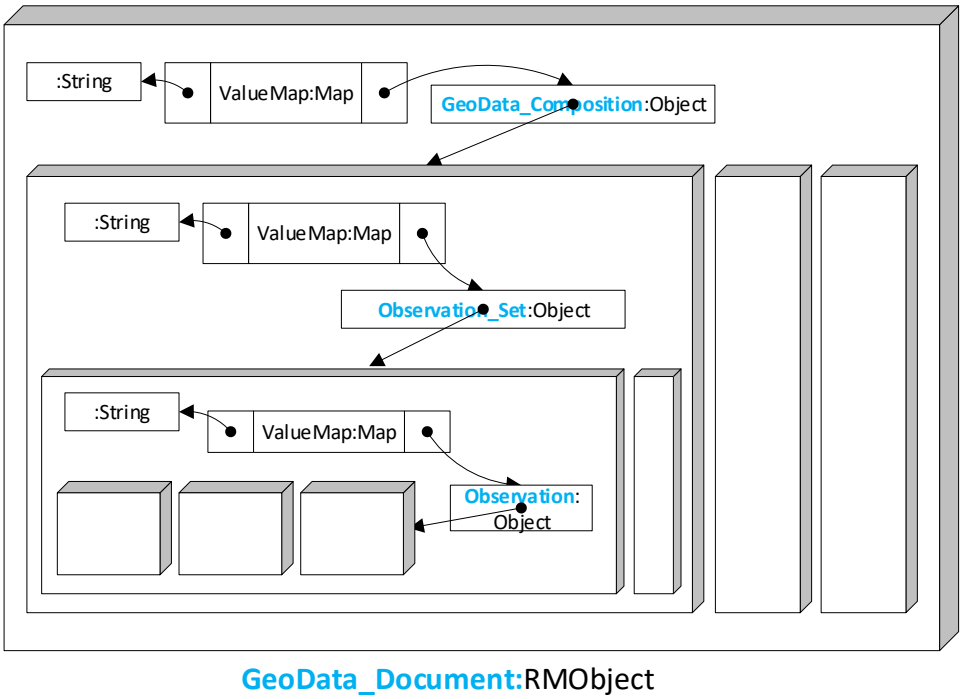


Figure 5.9 In memory representation of object tree representation of complex object tree. The system can only generate instances of objects of types found within the reference model. This validates the future-proofing concept of two-level modelling for the geospatial domain against the augmented O&M model. The system is hard coded against the reference model. Whereas the RMObject shown is further constrained at runtime against the AOM which is generated dynamically against the relevant archetype model (selection of archetypes used to create the template, see figure 5.6 above)

The *SkeletonGenerator* method `createComplexObject()` processes a given OperationTemplate object and builds a nested *Map* of values from the top level object, recursively working through the contained attributes (see Figure 5.9, 5.10 and 5.11). The method create() within the SkeletonGenerator class takes as an argument of type OperationalTemplate:

`create(OperationalTemplate opt, String templateId, Map<String,Archetype> archetypeList).`

The OperationalTemplate object contains a Map and List of all archetypes and attributes resulting from the parsing of an XML based .opt document. The create() method in turn calls the createComplexObject() method which constructs an RMObject, governed by the underlying reference model data structures, constrained by operational template. An object of type RMObject is ultimately returned.

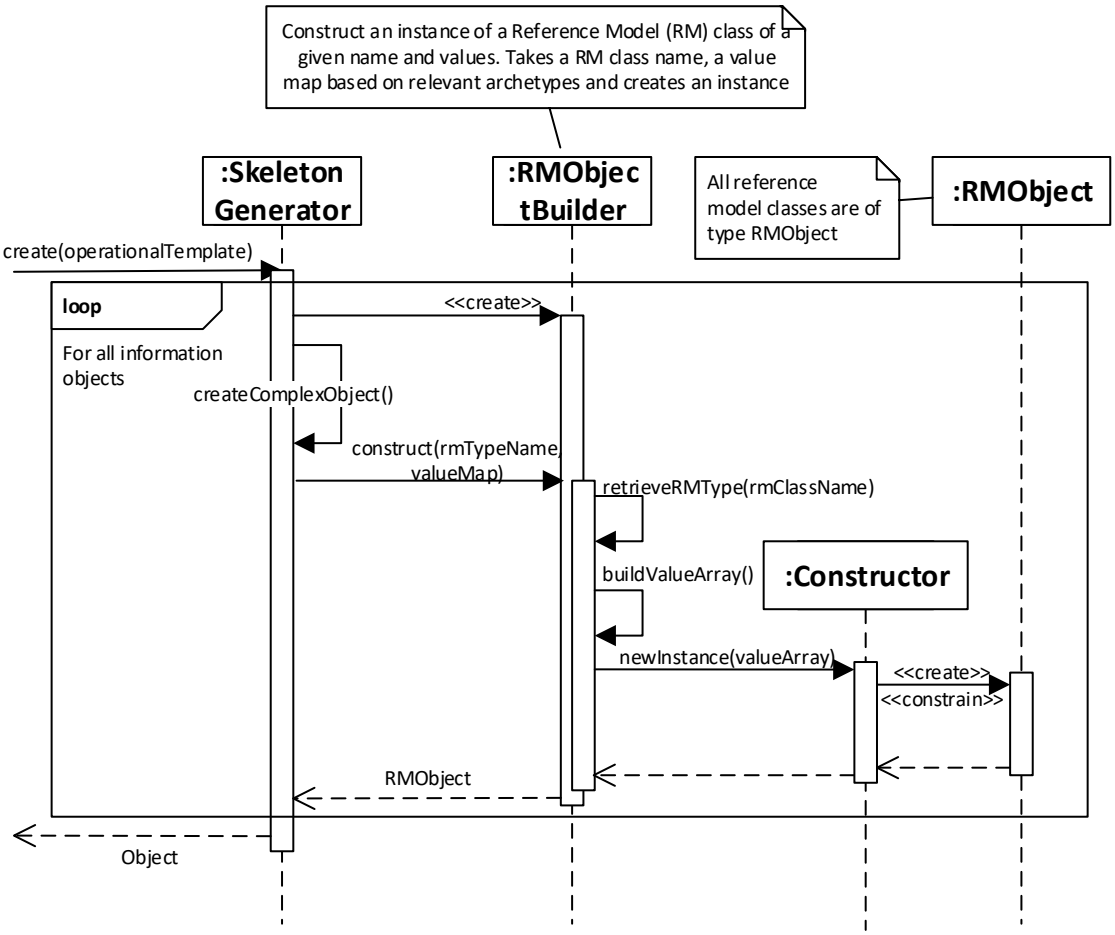


Figure 5.10 UML Sequence Diagram. Constrained reference model builder.

The returned object shown at the end of the operations sequence in Figure 5.10 above is an in-memory information instance, independent of the persistence layer that adheres both to the reference model and the archetype model. The processing steps implemented

here to validate the ability of an O&M based two-level reference model based system to generate information objects while adhering to both the underlying reference model and archetype model (i.e. adhering to a dual model) are shown in Figure 5.11 below.

What is notable is the system's ability to remain stable despite the evolution of domain knowledge which as it evolves is captured within domain specialist's defined archetypes. Once the reference model does not change, the system software remains stable (i.e. not needing to be updated or amended).

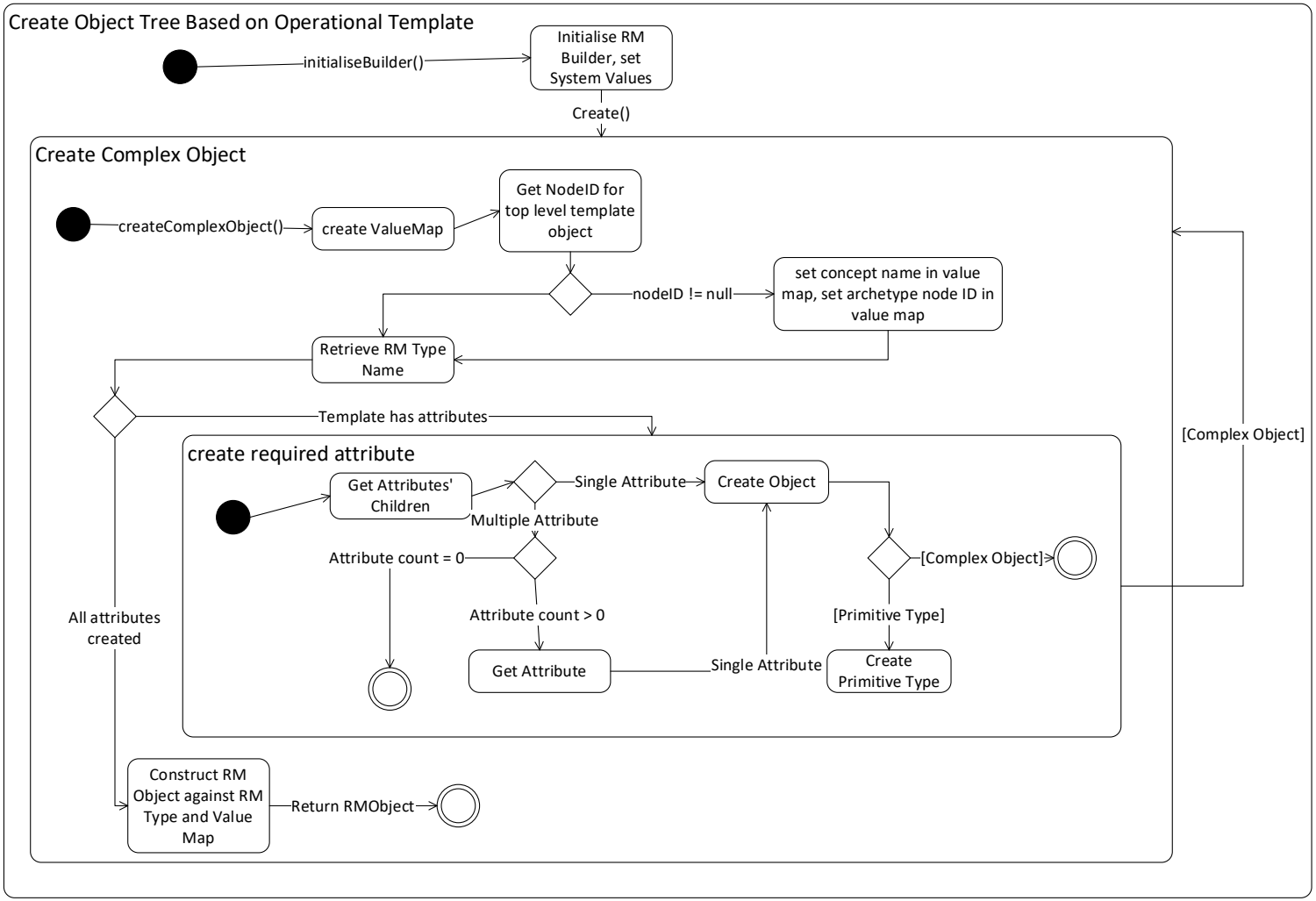


Figure 5.11 UML Activity Diagram. RMObject tree creation, which is constructed against the relevant system archetype model.

5.2.1.3 Package:tpot.archdev.am.template

The core classes within the template package are the OPTParser and the OperationalTemplate classes. OPTParser contains the `parseOPT()` method which takes an InputStream object representation of a textual XML based operation template document and creates an XMLObject from the InputStream. The template document is a flattened operational template, with all archetype constraints resolved and referenced to the originating operational template. XMLBeans (Apache, 2003) are used to parse the XMLObject representation of the flattened OPT. XML cursors navigate the various paths, extracting the required information to create an object of type OperationalTemplate. The class OperationalTemplate represents the TOM, a computable representation of the XML operational template document. See Figure 5.12 below for a step by step transformation approach implemented to aid validation of the structuring of the persistence layer against both the reference model and the archetype model.

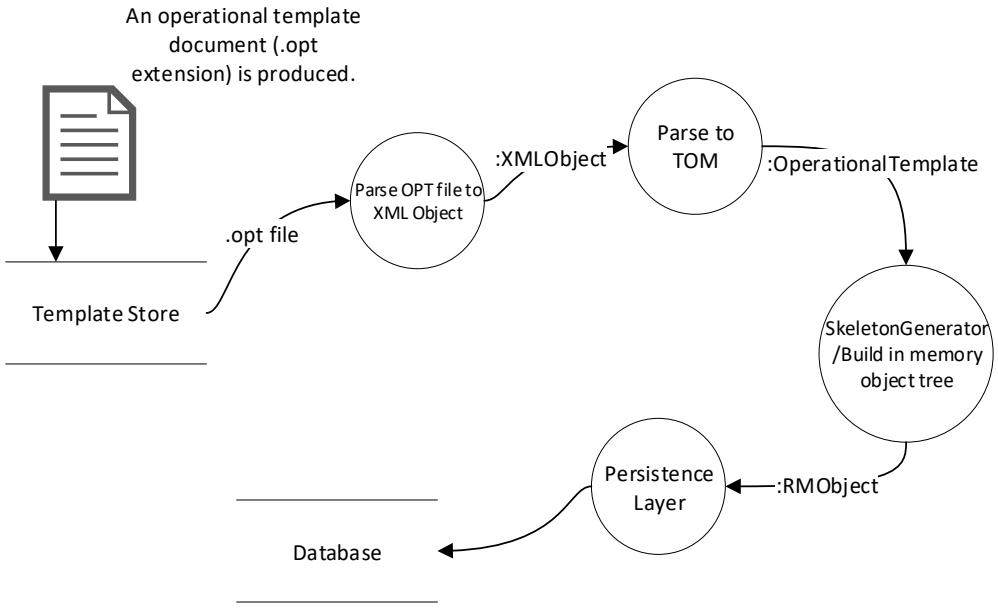


Figure 5.12 Data flow diagram of OPT transformation to persistence layer

5.2.1.4 Package:tpot.archdev.rm.core.om

The OM package contains the core reference model classes to support O&M based two-level model development. A screenshot of the Eclipse development environment (Wiegand, 2004) is shown in Figure 5.13 below. The package structure is shown in Figure 5.13 (the left panel within the Eclipse development environment) including the main hardcoded classes within the augmented O&M model. When the kernel creates data object instances, they are instances of type classes within this om package *tpot.archdev.rm.core.om*. The constructor for GeoData_Composition is also shown in the code view panel in Figure 5.13 below.

The instantiation of objects of class GeoData_Composition is performed with regard to the operational templates in use by the domain specific application. The operational templates (see Figure 5.6) are created from the relevant archetypes (defined or adopted for the specific scenario of use). Although the augmented O&M model is created at compile time, the constraining of these objects against specified archetypes using the operational templates happens at runtime. This concept is fundamental to future proofing systems as the reference model is considered stable while archetype definitions may change and evolve overtime. Instantiations create in-memory data objects that adhere both to the reference model (by way of the class definition shown in Figure 5.13 below) and the current operational templates that are being employed in the evaluation prototype system. These data objects are later serialised and persisted using the classes within the persistence package described next.

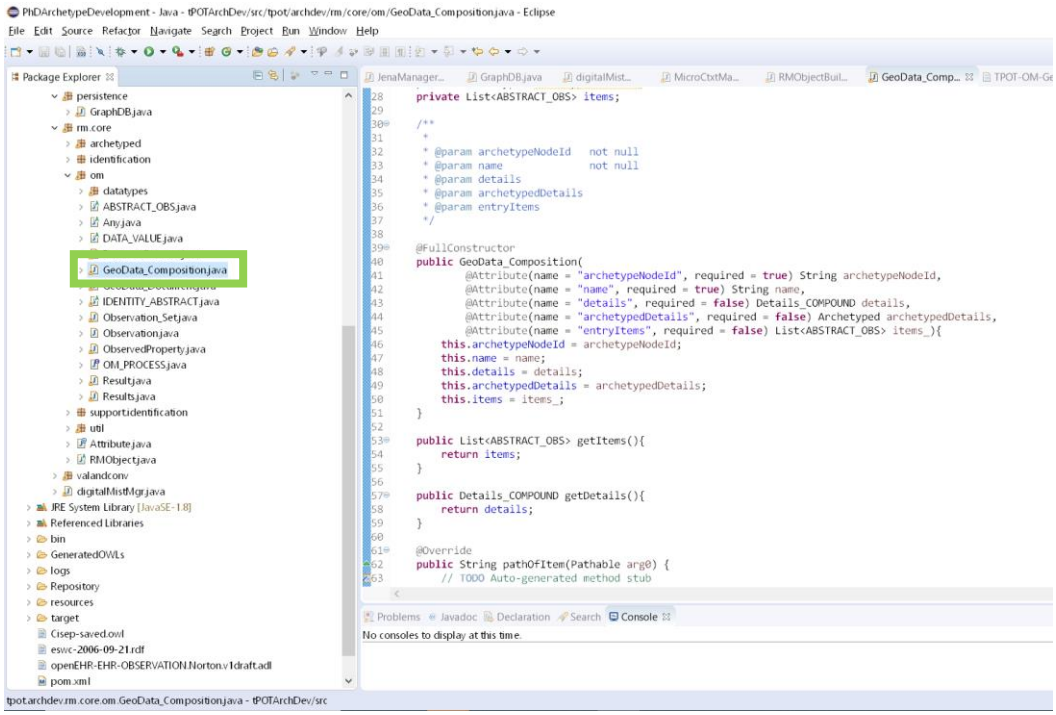


Figure 5.13 Eclipse development environment. The rm.core.om package shown in the package explorer on the left contains the object types of the reference model that are used to build the template object model against the supplied opt and archetype model (or in memory AOM). The class highlighted in green (GeoData_Composition) can also be seen in the context of the RMObject object representation shown in Figure 5.9.

5.2.1.5 Package:tpot.archdev.persistence

It has been noted in the previous sections that for this work, to enable the federated feature of the design, the persistence layer is a data graph, managed through the Apache Jena Linked Data framework (described in chapter 2 and 3). The dataflow diagram shown in Figure 5.12 above shows the final steps for the creation of the required data structure within the chosen database technology. The persistence package contains a number of classes, the main one of interest here is the GraphDB class. The job of the GraphDB class is to provide methods to map the RMObject data tree object (Figure 5.9) to create the required structures within the sensor data record within the TDB graph database. The *createDataGraph()* (Listing 5.3 below) method within the JenaManager class manages this process.

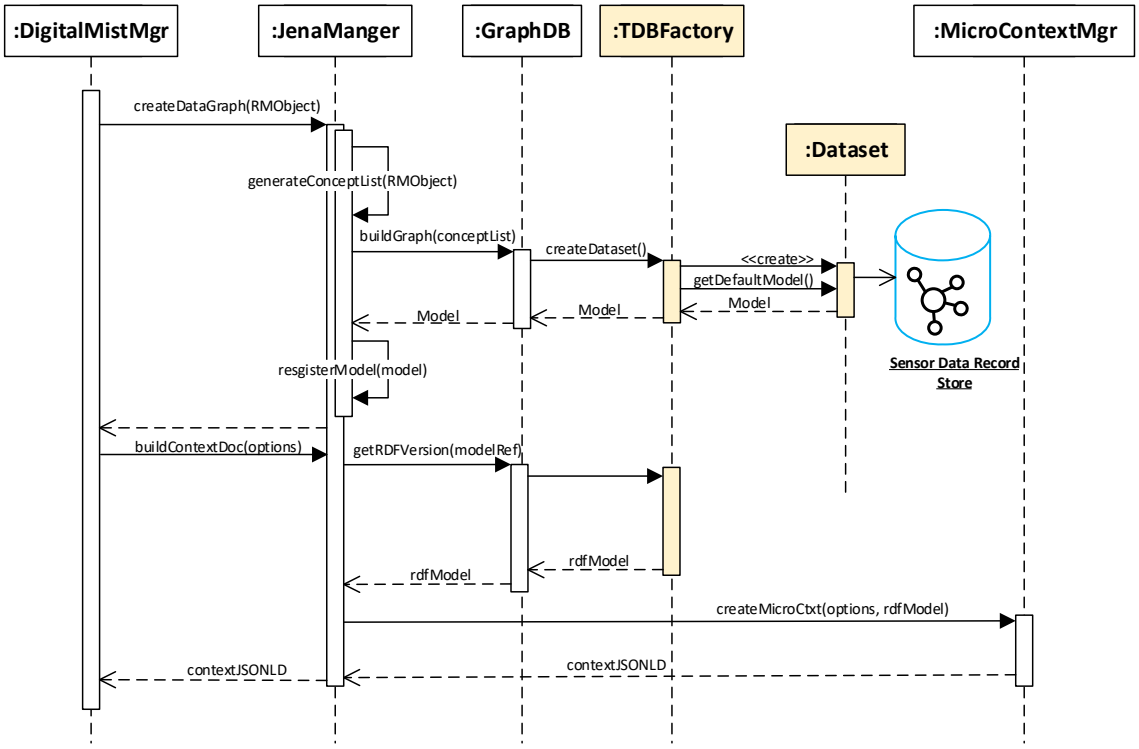


Figure 5.14 UML Sequence Diagram. Data graph builder

```
Public void createGraphDB(Object rmObject){

. . .

//Create Main Node & Compositions
compositionList = ((GeoData_Document) rmObject).getItems();
for(int i = 0; i < compositionList.size(); i++){
        composition = (GeoData_Composition) compositionList.get(i);
        gDB.addStatement(dataSetId, topLevelConcept,
          composition.getClass().getSimpleName(), composition.getUid());

        //Create Details_COMPOUND
        details = composition.getDetails();
        if(details != null){
                //Build details into graph
                buildDetailsNodes(gDB, details, composition);
        }

        //Create Sections
        sectionList = composition.getItems();
        for(int j = 0; j < sectionList.size(); j++){
                section = (Observation_Set) sectionList.get(j);
                gDB.addStatement(dataSetId,
                  composition.getClass().getSimpleName(),
                    section.getClass().getSimpleName(), null);

                //Create Details_COMPOUND
                details = section.getDetails();
                if(details != null){
                        //Build details into graph
                        buildDetailsNodes(gDB, details, section);
                }

                //Create Entries
                entryList = section.getItems();
                for(int x = 0; x < entryList.size(); x++){
                        entry = (Observation) entryList.get(x);
                        gDB.addStatement(dataSetId,
                          section.getClass().getSimpleName(),
                            entry.getClass().getSimpleName(), null);

                //Create Details_COMPOUND
                details = entry.getDetails();
                if(details != null){
                        //Build details into graph
                        buildDetailsNodes(gDB, details, entry);
        . . .
}
```

Listing 5.3 Code snippet of createGraphDB() method

5.2.1.6 Package:tpot.archdev.microctxtstore

It was noted in section 4.5.1 that the concept of micro-contexts has been developed as part of this work to enable fragmenting of archetype governed information instances. Micro-contexts are JSON-LD representations of element level information structures

within the larger data structure for a reporting platform. For each platform, a micro-context is created by the constrained knowledge framework. Micro-contexts contain device-group or device specific quasi-static information relating to the context of data collection for their associated device and are generated from the defined operational templates chosen for the deployment scenario. They are stored in the context store (Figure 5.1) and generally associated with a particular device or group of devices or organisation. The overall translation from archetypes (ADL files) to micro-contexts is shown in Figure 5.15 below.

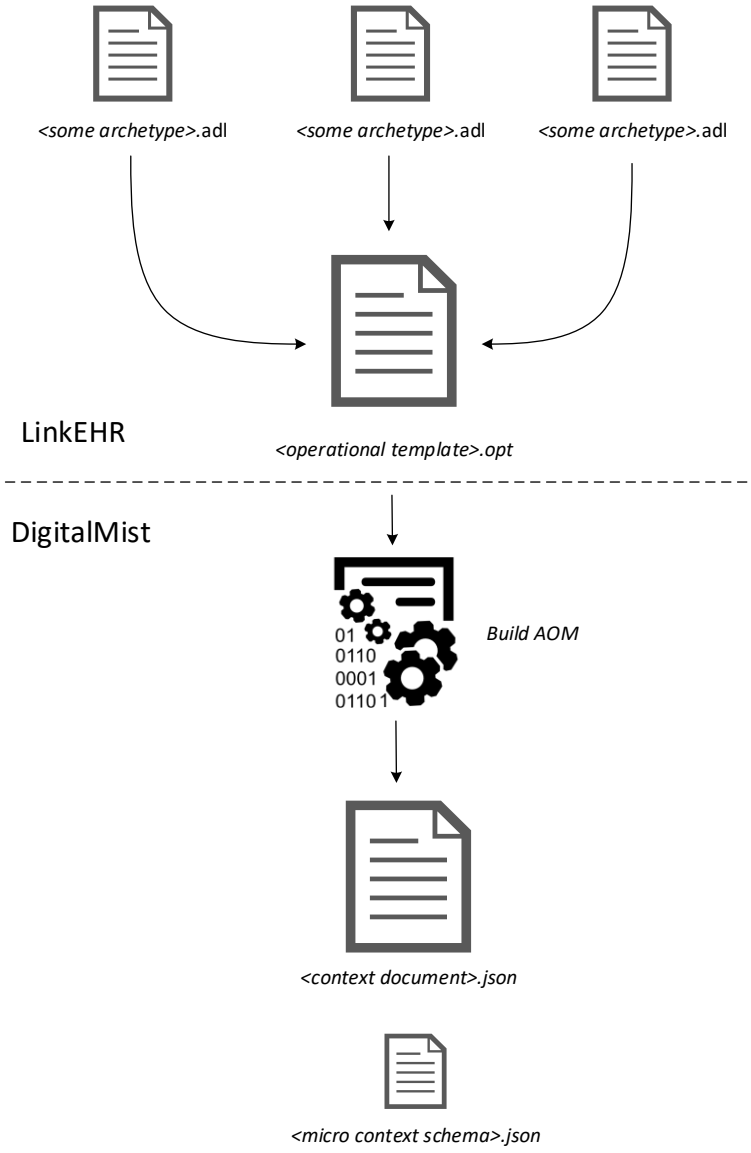


Figure 5.15 ADL to micro-context document transformation

The sequence diagram shown in Figure 5.16 below shows the runtime operation of the kernel creating a particular uContext.

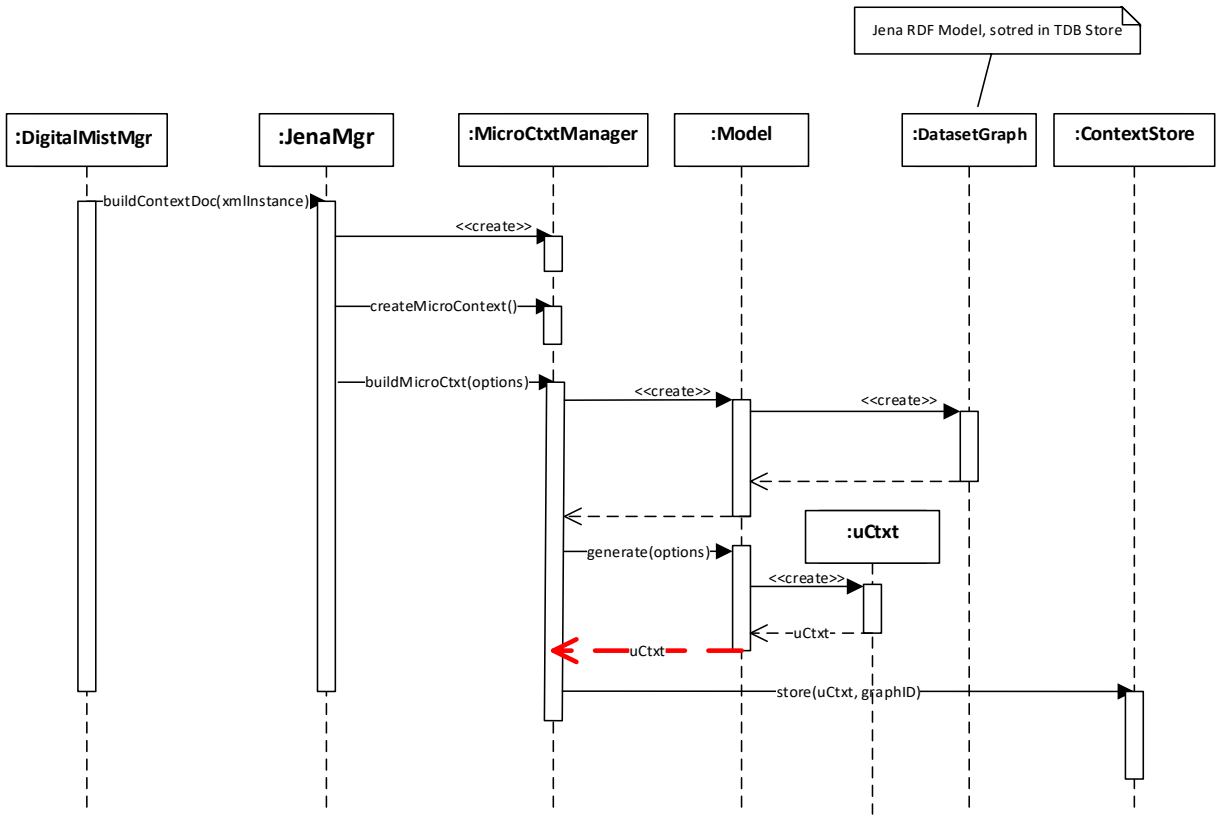


Figure 5.16 UML Sequence Diagram. Building micro-Contexts. The uContext object is returned (shown in red) and then stored within the context store to be later passed to observing platforms using the OPTaaS RESTful interactions shown in Figure 4.11. The Context Store is part of the micro @Context store component shown in Figure 5.7 and Figure 5.1.

Listing 5.4 shows an example of a micro-context document resulting from the processing steps shown in Figure 5.15 and Figure 5.16.

```
"@Context" : {
  "obj_store" : "coap://tpot.arch-dev.ie/obj_store/",
  "obj_id" : {
      "@id" : "obj_store:obj_id",
     "@type" : "@id"
  }
  "at0002" : "obj_id:at0002/",
  "at0004" : "at0002:at0004/",
  "at0008" : {
         "@id" : "at0004:at0008",
         "@type" : "@id"
  },
  "DV" : {
         "@id" : "at0008:#at0009",
         "@type" : "@id"
  },
  "resultTime" : {
         "@id" : "at0008:#at0010",
         "@type" : "@id"
  }
```

Listing 5.4 Sample micro-Context JSON-LD representation

When a device registers with the backend system, a JSON schema representation of the device's micro-context is returned to the device (see chapter 4, Figure 4.11). The device's micro-kernel parses the schema document received and uses it as the template to define and build information instances. The JSON schema document definition of the micro-context shown in Listing 5.4 is shown below in Listing 5.5.

```
{
  "$schema": "http://json-schema.org/draft-07/schema#",
  "$id": "http://example.com/root.json",
  "type": "object",
  "title": "micro context",
  "required": [
    "@context",
    "@id",
    "obj_id",
    "DV",
    "resultTime"
  ],
  "properties": {
    "@context": {
      "type": "string",
      "pattern": "coap://tpot.archdev.ie/microcontexts/{microCtxt_id}"
    },
    "@id": {
      "type": "string",
      "pattern": "at0008"
    },
    "obj_id": {
      "type": "string",
      "pattern": "{sdr_object_id}"
    },
    "DV": {
      "type": "number"
    },
    "resultTime": {
      "type": "number"
    }
  }
}
```

Listing 5.5 Sample Micro-Context JSON Schema Document

Devices essentially enter into a contract with the constrained knowledge framework. The information instances must then adhere to the micro-context, and information instances received by the backend framework will be validated against the context stored against the device's ID within the context store (Figure 5.1). Validation is performed using the JSON schema validation Java libs[61] and is shown as the validation and converter software component in Figure 5.1 and Figure 5.7.

Validation of data instances is also performed locally on the observing platform within the micro-kernel (Figure 5.1 & Figure 5.2). Rigorous information structure definition including multi-step information validation across the framework helps to ensure data quality is maintained right to the edge of the network. The observing platform's micro-

[61] https://github.com/everit-org/json-schema

context schema document (Listing 5.5) enables information object validation at the point of capture. This ability to validate information throughout the data/information value chain is one of the main benefits of two-level modelling.

Listing 5.6 below shows an example of a simple information instance that adheres to the micro-context definition in Listing 5.5 above, and that which an observing platform may report to the backend constrained knowledge framework. The information object is notably small (147 characters in this case), meaning it is well suited to observing platform technological constrains. However, despite its tiny size, by using the methodology developed as part of this work the information object is linked to a wealth of metadata that can support a chain of quality assessment. Further efficiencies can be achieved through the shortening of the URI string.

```
{
 "@context" : "coap://tpot.archdev.ie/microcontexts/{microCtxt_id}",
 "obj_id" : "{sdr_object_id}",
 "@id":"at0008",
 "DV": 10.23,
 "resultTime": 123
}
```

Listing 5.6 Example Micro-Context Constrained Information Instance

Once the context is received and validated by the backend system, the backend system will process the information point against its *@Context* value. The @Context value uses the linked data approach to bind the data point to the platforms relevant data graph within the graph database. The Graph database maintains the information structures against the archetype and reference models.

The interactions between technologically constrained observing platform (client) and the supporting system (shown in Figure 4.11. and Figure 5.1) follow the defined OPTaaS services. This is described in more detail next.

5.2.2 OPTaaS

The operational template as a service (OPTaaS) is implemented using a Hypermedia as the Engine Application Stack[62] (HATEOAS). JAX-RS and Groovy on Grails are used as the implementation technologies. These are largely chosen due to the kernel development being a Java implementation, and the desire to use a Web application framework to support development. Grails uses the Model View Controller (MVC) architectural pattern (Figure 5.15).

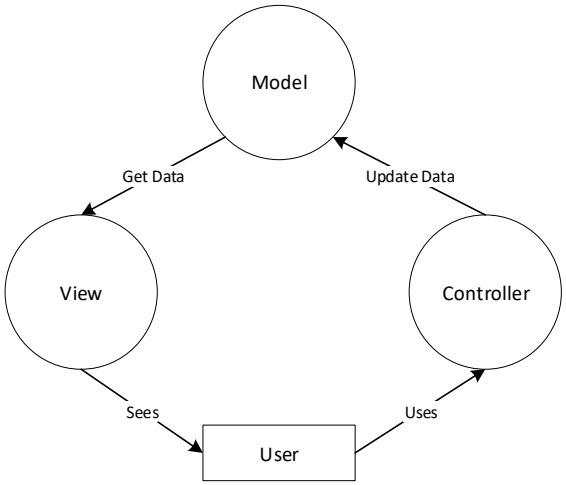


Figure 5.17 Model View Controller Architectural Pattern

In Groovy on Grails the MVC model is defined using Groovy classes. The model is a code level definition of the data model for an associated database. For this work MySQL DBMS is used. For relational databases, class definitions result in table definitions, or entity design. For example, the groovy code listing shown in Listing 5.7 will result in a new table called Platform being created in MySQL.

[62] See Roy Fielding's blog on HATEOAS within REST APIs https://roy.gbiv.com/untangled/2008/rest-apis-must-be-hypertext-driven

```
package tpot.archdev

class Platform {
    String name, type
    Date deploymentDate
    ….
    static constraints = {
          name blank: false, unique: true
    }
}
```

Listing 5.7 Groovy domain model definition example

Controllers are the application logic definitions and are a set of services used by users or clients of the application. Controllers are exposed through URLs. For example, the Groovy listing 5.8 below used for initial system testing shows an example of controller with 3 services defined, *registerdev*, *getuctxt* and *obsappend*.

```
package tpot.archdev.optaas

import tpot.archdev.Device
import tpot.archdev.Registration
import grails.converters.JSON

class OptaaservController {
          def registerdev() {
                def _dev = Device.get(params.devid)
                def _reg = _dev.getRegID()
                _reg.type = "blue"
                _reg.save(flush: true)
                render "2.01 - CREATED"
         }
         def getuctxt() {
                def _dev = Device.get(params.devid)
                def _uctxt = _dev.getUctxt()
                 def microCtxt = _uctxt.getMicroctxt_id()

                ....//code removed

                render responseData as JSON
        } else render "No uctxt found for specified device"
        }
        def obsappend() {
                def _dev = Device.get(params.devid)
                //def parsedReqData = request.
                def JSONrequest_object = JSON.parse(request)
                render JSONrequest_object
        }
        def show() {}
}
```

The register device service receives a registration request for a particular device via a HTTP GET request and sets it's registration status to blue

Listing 1Listing 5.8 Example Controller Definition in Groovy

The main controllers defined for the OPTaaS are `/register; /getMicroContext and /obs-append` (Figure 5.16).

Views are defined using Groovy server pages (GSP) and provide an interface view for the application. Listing 5.9 below shows an example of a simple GSP defined view of a device registration form.

```
<%@ page contentType="text/html;charset=UTF-8" %>
<%@ page import="tpot.archdev.Organisation" %>

<html>
<head>
<meta http-equiv="Content-Type" content="text/html; charset=ISO-8859-1"/>
<meta name="layout" content="main"/>
<title>Device Registration</title>
</head>

<body>
  <div class="body">

  <p>This form allows you to pre-register your organisation's
   Observational Device/Platform. You must select your organisation,
   associated Geo-observational document, Observational Project$
   and appropriate Operational Template and associated micro-
   context.</p>

  <p>Once your device boots up and registers on the system using the
   given ID, a microcontext will be created and returned directly to
   your device.Your device may then begin appending observations to
   the global Geo-Obs Document</p>

  <p>Use the form below to pre-regsiter your device for your
   organisation</p>

  <g:form controller="device" action="save">

        <label>Your DeviceID : </label>
        <g:textField name="devID" /><br/>

        <label>Device Location : </label>
        <g:textField name="location" /><br/>

        <label>Device Latitude : </label>
        <g:textField name="lat" /><br/>

        <label>Device Longitude : </label>
        <g:textField name="lng" /><br/>

        <label>Organisation : </label>
        <g:select name="id"
               from="${Organisation.list()}"  optionKey="id" /><br/>

        <label>Device Description : </label>
        <g:textField name="description" /><br/>

        <g:actionSubmit value="Save"/>
  </g:form>

  </div>
</body>
</html>
```

Listing 5.9 Groovy Server Pages View Definition

The OPTaaS is provided as part of the wider *DigitalMist* backend. The DigitalMist framework allows full management of devices. Devices are maintained in 3 separate states *Grey* (pre-register), *Blue* (registered) and *Green* (observing and reporting). Devices

may pre-register using the url: digitalmist.ie:8080/OPTaaS/pre-reg/. A relational database is maintained in the backend, built using GORM (Rocher and Brown, 2009) and MySQL (domain folder within the Grails/Eclipse project explorer panel in 5.18 below).

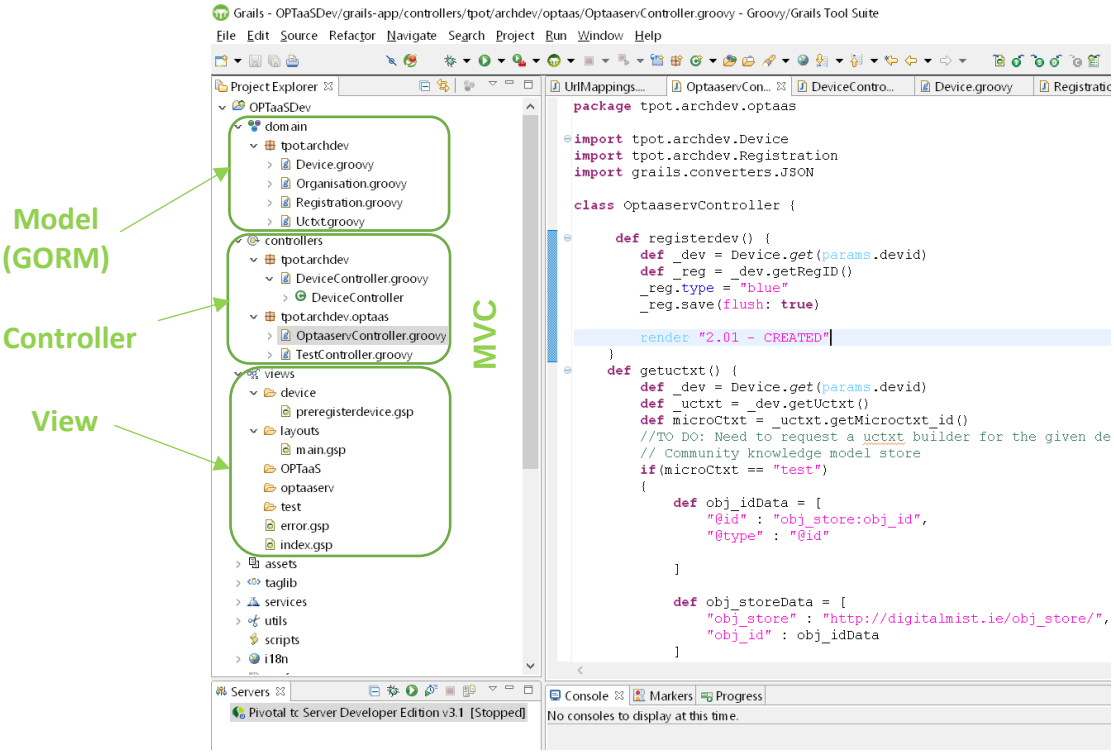


Figure 5.18 Grails Development Environment. Used to build OPTaaS based Web services.

5.3 Device Design & Implementation

A kernel in two-level health-based systems is defined as a constructor and processor of the informational structure of EHRs (Beale, 2000). Given the information objects within the overall system are federated across the observational network, a federation of kernels are needed. The last section described the core centralised knowledge kernel, the requirements and functionality of the constrained knowledge kernel to perform the creation and validation of information fragments on-board devices or observing platforms (shown previously in Figure 5.1). This section describes the design and implementation

approach of the node level to provide validation of the concepts presented in chapter 4 on a technologically constrained observing platform (embedded board).

Reporting devices or observational platforms that participate within the observational network create data nodes as part of a wider linked data graph. Therefore, the device level kernel must also support *tripified* data at the edge of the network (see section 4.5.1). The impact of the solution described here is detailed in the chapter results section (section 5.4.2) below. First, the development of the embedded kernel solution is described.

5.3.1 ContikiMist Kernel

A practical problem while adopting a two-level modelling approach within constrained devices is that of creating a cut-down and lightweight kernel. The kernel among other things generates an instantiation of both the AOM and TOM structure and constraints (section 5.2.1) at the constrained node level. This requires the hardcoding of the reference model within the embedded system implementation. One would be correct in highlighting the additional memory overhead challenges; however, this is not the initial concern here. Most embedded operating systems are programmed natively in C or some other C variant like nesC (as discussed in chapter 2). These languages are structured languages and do not natively support object-orientation. As should be clear to the reader at this point, the dual-model implementation paradigm is inherently object-oriented. This presents a problem. Schreiner (1993) deals extensively with this type of issue in his book *Object oriented programming with ANSI-C*[63] providing practical design patterns for the problem.

[63] A legitimate free copy of Shreiner's book can be obtained at the following link: https://www.cs.rit.edu/~ats/books/ooc.pdf

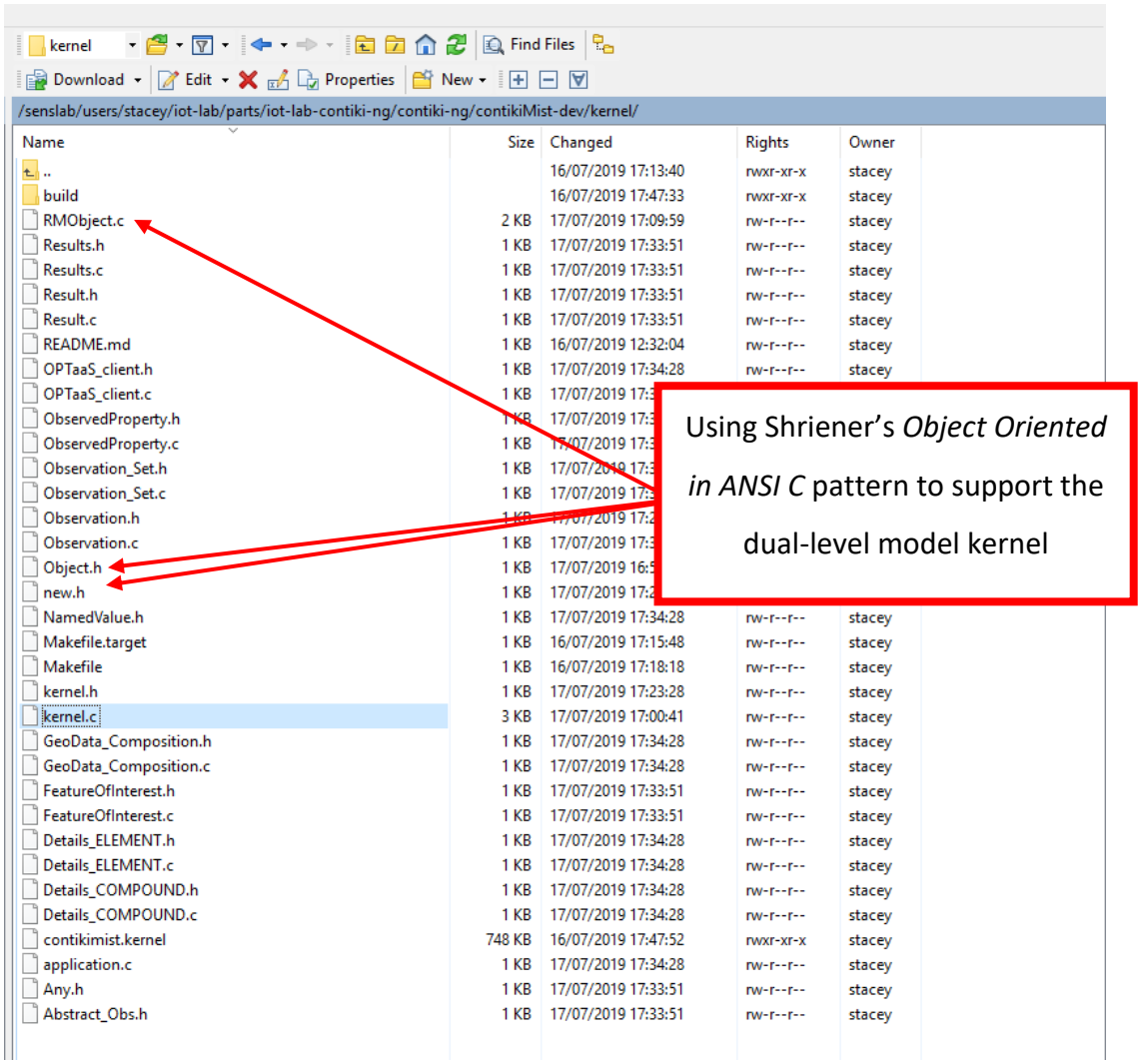


Figure 5.19 ContikiMist File Overview

Shreiner's approach has informed the kernel implementation and the coding of the AOM/TOM software here (Figure 5.19). Specifically, the issue of strong typing within the reference model and result AOM/TOM is addressed using his proposed approach. The kernel must support the parsing of the returned JSON schema based microContext and the resulting *micro-TOM* for the individual node. The JSON schema constrains the production of in memory reference model-based objects, as is the case in the backend implementation. The WJElement[64] library is used to provide JSON schema validation in

[64] https://github.com/netmail-open/wjelement

C. For more general JSON based parsing, Contiki-NG provides JSON parsing support *out-of-the-box* through the `jsonparse.h` api.

Contiki-NG provides the 6LoWPAN (Shelby and Bormann, 2011) network stack via a RPL border router, which in turn acts as a 6LoWPAN router. The Coniki-NG CoAP engine is based on Erbium. The kernel contains a CoAP client, with datagrams sent over UDP.

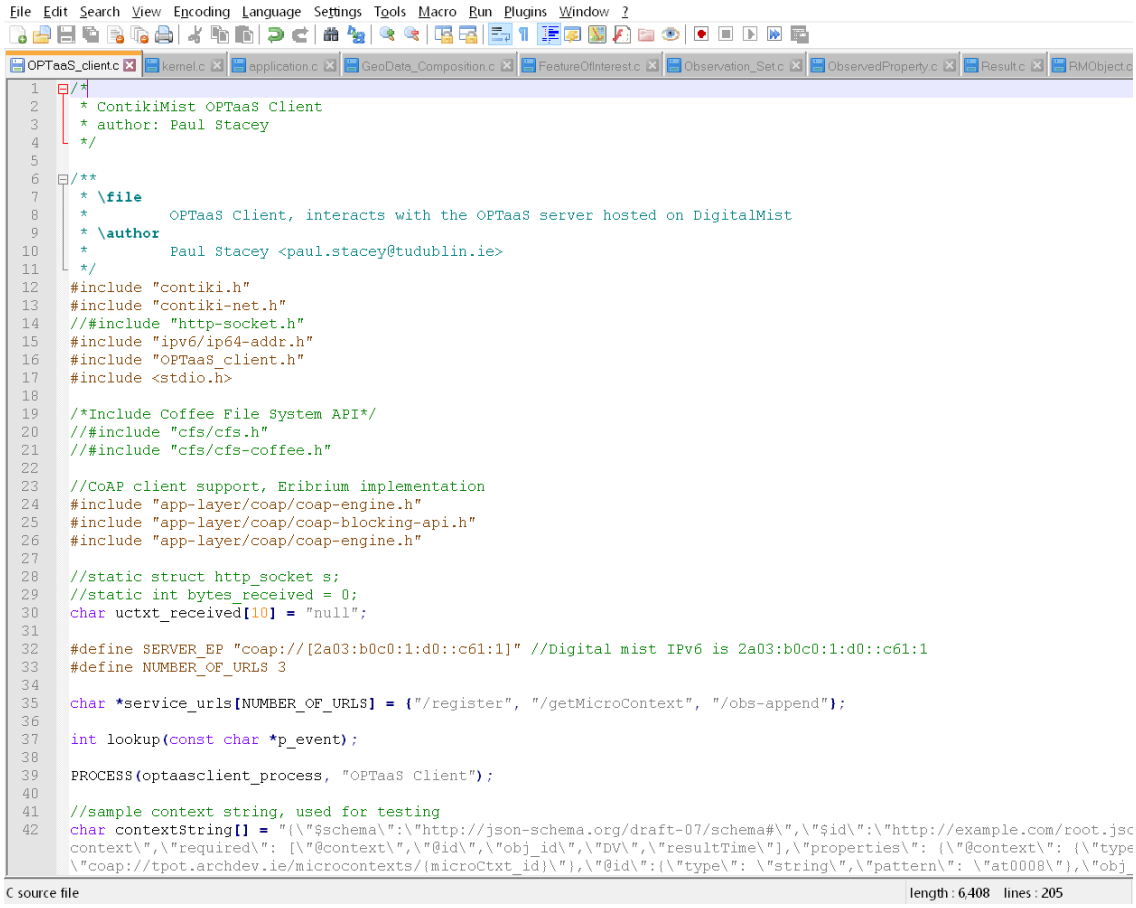


Figure 5.20 ContikiMist Development Environment

To demonstrate how the evaluation prototype would handle CoAP requests, and for ease of development, middleware implemented in NodeJS was developed to handle CoAP based messaging (Figure 5.1). The middleware element of the prototype uses the node-coap library[65] (shown as CoAP Server in Figure 5.1). The DigitalMist server handles

[65] https://github.com/mcollina/node-coap

HTTP requests. However, the middleware, receives CoAP requests and re-routes them via HTTP requests, the OPTaaS handles the HTTP response and relays the response via a CoAP response message (see the code snippet in Listing 5.10 below). The Google Chrome CoAP extension Copper4Cr[66] was used for initial middleware testing before the full system deployment (Figure 5.1) was performed.

```
var coap = require('coap')
var server = coap.createServer({ type: 'udp6' })

server.on('request', function(req, res) {
  var urlReqService = req.url.split('/')[1];
  console.log("CoAP server received a " + req.method + " request with url: "
       + urlReqService + '\n');
  if(req.method == 'PUT'){
    switch(urlReqService){
      case 'register':
         var receivedata = req.payload;
         var deviceID = req.url.split('/')[2];
          console.log('Device ID : ' + deviceID + '\n');
         console.log("Received request to register device " + deviceID);
         res.end(registerDevice(deviceID));
      default:
         res.end('error url not recognised');
    }
  }
  else if(req.method == 'GET'){
    switch(req.url){
      case '/microcontext':
          var receivedata = req.payload;
          console.log("Received request for a micro context \n");
          res.end(getMicroContext(receivedata));
      default:
          res.end('error url not recognised');
      }
  }
  else if(req.method == 'POST'){
    switch(req.url){
      case '/observation-append':
          var receivedata = req.payload;
          console.log("Received the following observation: " + receivedata +
               "from device :\n");
          res.end(appendobs(receivedata, deviceid));
      default:
          res.end('url not recognised');
    }
  }
  else
    res.end('Request method not supported \n');
})
```

Listing 5.10 DigitalMist-CoAP-OPTaaS Middleware Code Snippet. The main CoAP services are highlighted in red.

[66] https://github.com/mkovatsc/Copper4Cr

5.3.1.1 Constrained Storage

At the observational platform *node* level a persistence layer solution is also required to manage the local *data-nodes* within the overall distributed data graph. Contiki-NG contains the Antelope database, which is a relational database that is optimised for flash storage (Tsiftes and Dunkels, 2011). Antelope does not allow for storing variable-length strings, and string size must be configured to be a specified fixed length. The implications of this feature form implementation-based validation of the resource constrained edge-based storage of observations using the two-level modelling approach are discussed later in section 5.5. Antelope has been designed to run on the file-system Coffee (amongst others). Antelope supports its own query language called Antelope Query Language[67] ($AQL_{contiki}$).

The constrained federated knowledge kernel has been developed using linked data principles. As discussed previously, the JENA linked data framework drives the linked data approach on the main backend. At the observational platform level, the linked data approach is coerced onto the relational model (see section 4.5.1) supplied by the antelope database. Listing 5.11 below shows the Contiki-NG based C code used to interact with the antelope database to create the relational table `triple-store` containing the attributes/columns `subject-predicate-object,` which are highlighted in green in the listing.

[67] Unfortunately, in the context of this research Antelope Query Language is normally shortened to "AQL". In order to avoid confusion between Archetype Query Language and Antelope Query Language, here $AQL_{contiki}$ is used to refer to the Antelope Query Language

```
/*
 * ContikiMist Application
 * author: Paul Stacey
 */

.... //code removed

//Create TripeStore Table
db_query(&handle, "REMOVE RELATION triple_store;");
result = db_query(&handle, "CREATE RELATION triple_store;");
printf("result : %i \n", result);

if(DB_ERROR(result)) {
      printf("Query \"%s\" failed: %s\n", "CREATE RELATION
                              triple_store", db_get_result_message(result));
}
else{
  printf("Query \"%s\" : %s\n", "CREATE RELATION triple_store",
                                          db_get_result_message(result));
.... //code removed

result = db_query(&handle, "CREATE ATTRIBUTE _id DOMAIN LONG IN
                                                          triple_store;");
printf("result : %i \n", result);

.... //code removed

db_free(&handle);

result = db_query(&handle, "CREATE ATTRIBUTE subject DOMAIN
                                             STRING(10) IN triple_store;");
.... //code removed

result = db_query(&handle, "CREATE ATTRIBUTE predicate DOMAIN
                                             STRING(10) IN triple_store;");
.... //code removed

result = db_query(&handle, "CREATE ATTRIBUTE object DOMAIN
                                             STRING(10) IN triple_store;");
.... //code removed

result = db_query(&handle, "CREATE ATTRIBUTE context_subj DOMAIN
                                             STRING(10) IN triple_store;");
 .... //code removed
```

Listing 5.11 ContikiMist Application Code – create node level relational table based triple store

Using $AQL_{Contiki}$ the coerced graph can be queried. Antelope allows for the selection of indexing algorithms based on specific use-cases.

5.4 Testing and Deployment

The backend framework is deployed on a DigitalOcean hosted droplet[68] (Figure 5.1 & Figure 5.18). A Droplet represents an OS instance which may or may not have dedicated hardware resources within the DigitalOcean hosted cloud service infrastructure. For this work Linux (Ubuntu) based Droplets are deployed.

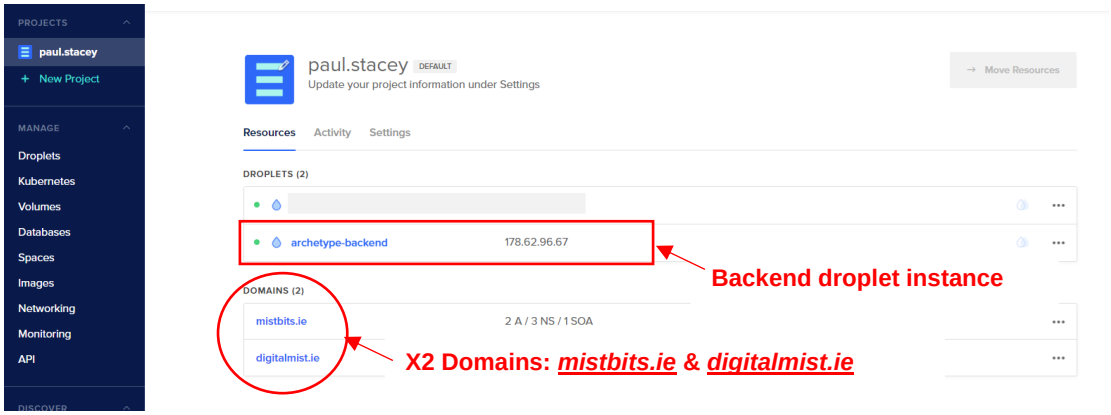


Figure 5.21 DigitalOcean Management Dashboard

The droplet is configured as an Ubuntu NodeJS 6.9.5 distribution on Linux Ubuntu version 16.04 image with a modest 2GB of physical memory and 20GB of hard disk space. The backend also supports 2 website domains digitalmist.ie & mistbits.ie which have been previously described in section 5.1.1 and for the purposes of the evaluation prototype, are Web front ends to support the framework configuration and testing.

As described above, the constrained device implementation kernel (*ContikiMist)* is built on top of Contiki-NG. For testing purposes, the Future Internet of Things (FIT) IoT-Lab infrastructure[69] is used to help scale the testing environment (Adjih et al., 2015).

[68] https://www.digitalocean.com/products/droplets/
[69] https://www.iot-lab.info/

	Wsn430 Node	M3 Node	Linux Node (A8)	SAMR21 Node	ST LoRa Node	Firefly Node	nRF52DK Node	nRF52840DK Node
FreeRTOS	x	x	-	-	-	-	-	-
RIOT	x	x	-	x	x	x	x	x
ContikiNG	x	x	-	-	-	x	-	-
Zephyr	-	-	-	-	-	-	x	x
OpenWSN	x	x	-	-	-	-	-	-
TinyOS	x	-	-	-	-	-	-	-
Linux	-	-	x	-	-	-	-	-

OS & Platform used for testing

Figure 5.22 FIT IoT Lab OS and Node Support

The FIT IoT-LAB provides very large-scale infrastructure for testing small wireless sensor devices and heterogeneous communicating objects. The FIT IoT Lab enables free experimentation on real live devices. The lab provides support for 7 popular embedded operating systems (FreeRTOS, TinyOS, ContikiNG etc.) which have a wide and varying support across 8 popular platforms (ARM M3/A9 nodes etc.) (Figure 5.22). Users may perform remote development through remote SSH to any of site, such as the Grenoble based backend server (Figure 5.23).

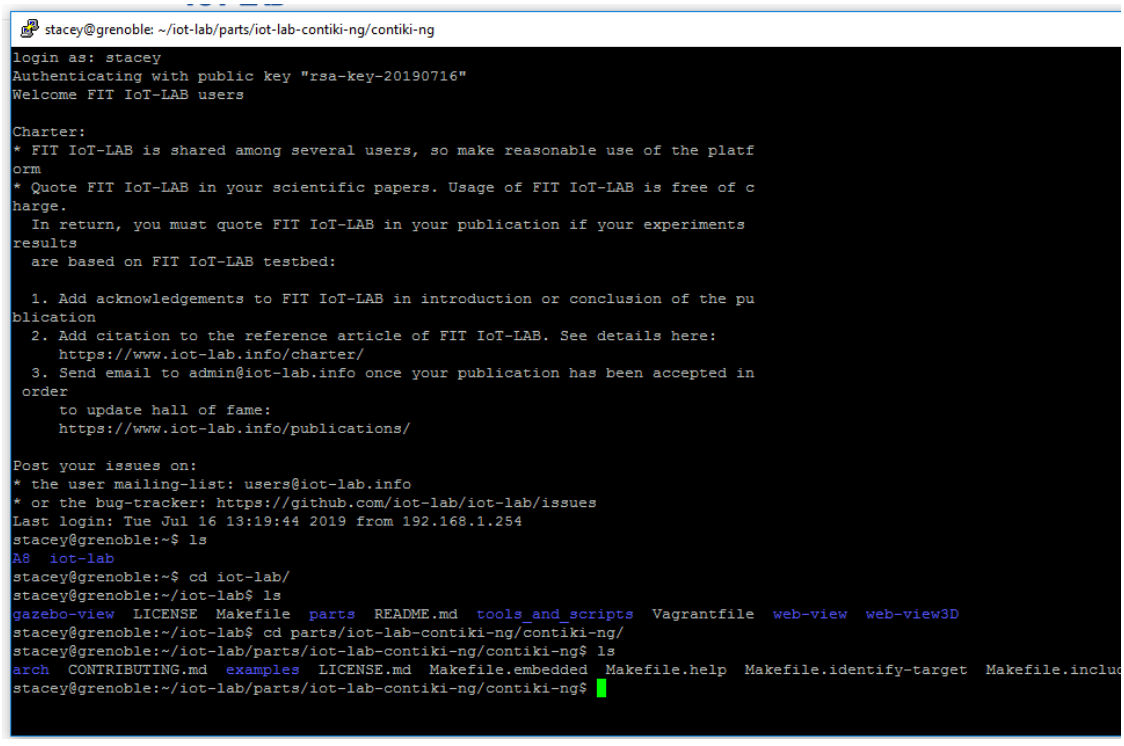


Figure 5.23 Remote SSH into the IoT Fit Lab Grenoble Backend Contiki-NG Dev Environment

5.4.1 Experimental Setup

To implement the system deployment shown in Figure 5.1 an experimental setup using the IoT Fit Lab was performed. The aim of this experiment was to observe and validate the overall constrained system framework for adherence to the system functional requirements of two-level modelling systems described in chapter 4, and the general design considerations described at the beginning of this chapter. The experiment was also run to measure the impact on overhead in terms of battery power, communication load and to measure the server load for the given experimental setup, which in turn would allow a heuristic approach to server requirement sizing for larger scale configuration.

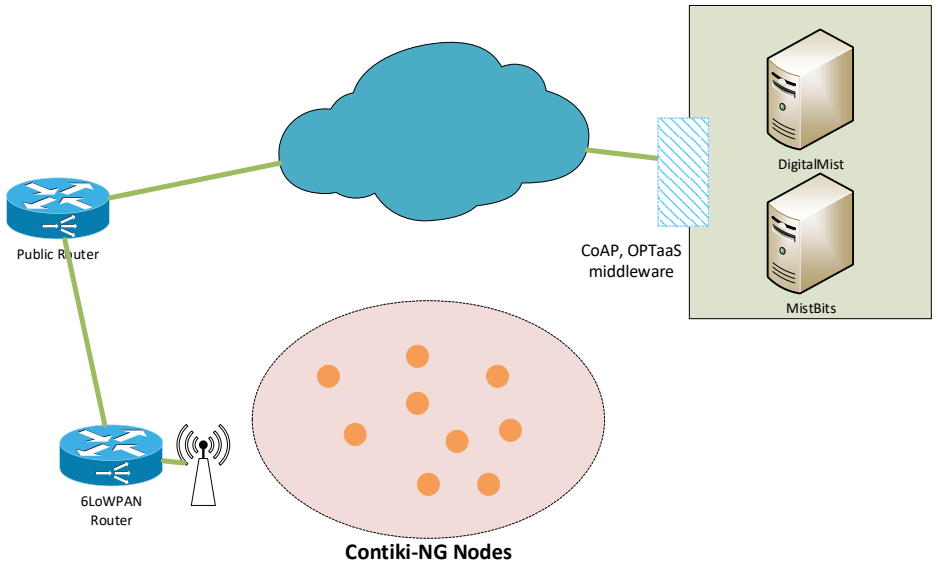


Figure 5.24 Experimental Configuration

In order to evaluate the approach as a large-scale deployment, constrained device development was performed in a cross-compile environment using Eclipse tools on top of Contiki-NG. A ConitkiMist based firmware image was created and deployed across 10 separate ARM M3 platforms. The experiment was configured using the FIT IoT-Lab Experiments interface and was scheduled to run for a 2 hour period on the Grenoble site. The Grenoble site contains 384 physical M3 based nodes and 256 physical A8 nodes deployed and networked for remote access on site (Figure 5.25).

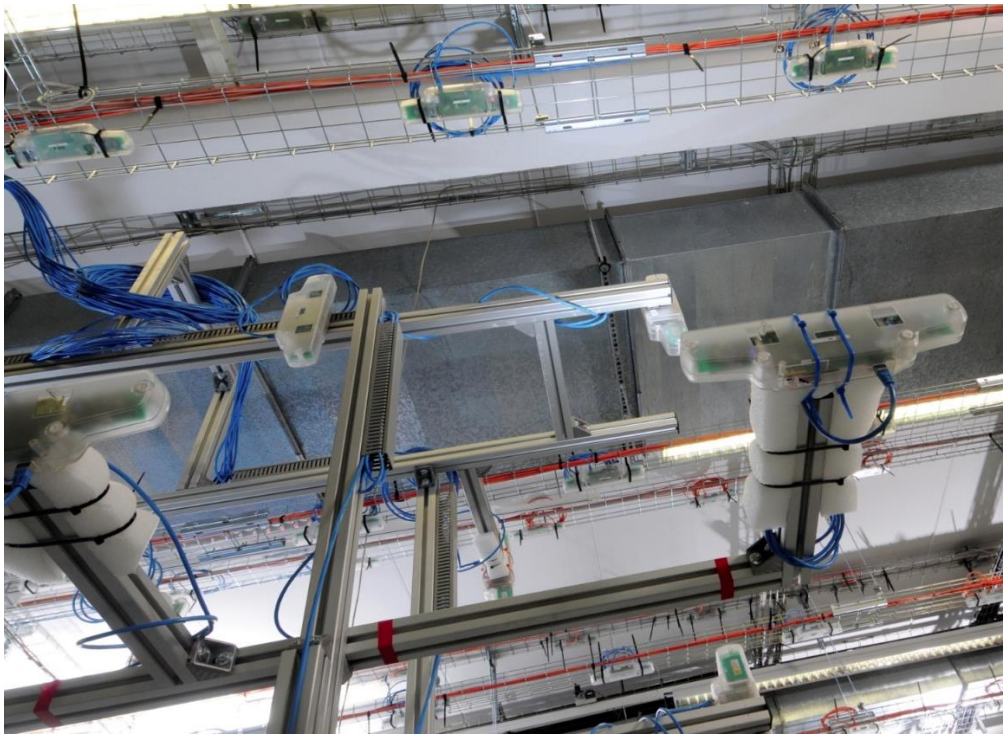


Figure 5.25 M3 and A8 Node at the FIT IoT-Lab Grenoble[70]

The DigitalMist server droplet (Figure 5.1) was configured to support IPv6[71]. The data buoy archetype described in chapter 4 was used for experimentation. Micro-contexts were generated to ensure nodes reported data structures at the Results level of the overall information instances. Devices were pre-registered on the backend system using the *mistbits.ie* Web tool, created specifically for this experiment which is a proof-of-concept implementation validation of the OPTaaS concept presented in chapter 4 (see Figure 4.11). Devices then completed their registration process with the backend adhering to the OPTaaS protocol further detailed in chapter 4, Figure 4.11.

Once registered, nodes received their micro-contexts, which were in turn loaded into their on-board kernel. Each node was configured to "observe" and report simulated *sea_surface_temperature* and *practical_salinity* data every 30 seconds for the 2 hour experimental period.

[70] https://www.iot-lab.info/deployment/grenoble/
[71] https://www.digitalocean.com/docs/networking/ipv6/how-to/enable/

Figure 5.26 below is a screenshot from the FIT IoT Lab experiment management interface showing all 10 observing platforms (ARM based M3 nodes running Coniti-NG and the Contiki-Mist firmware) successfully deployed and running live and reporting on the Grenoble site.

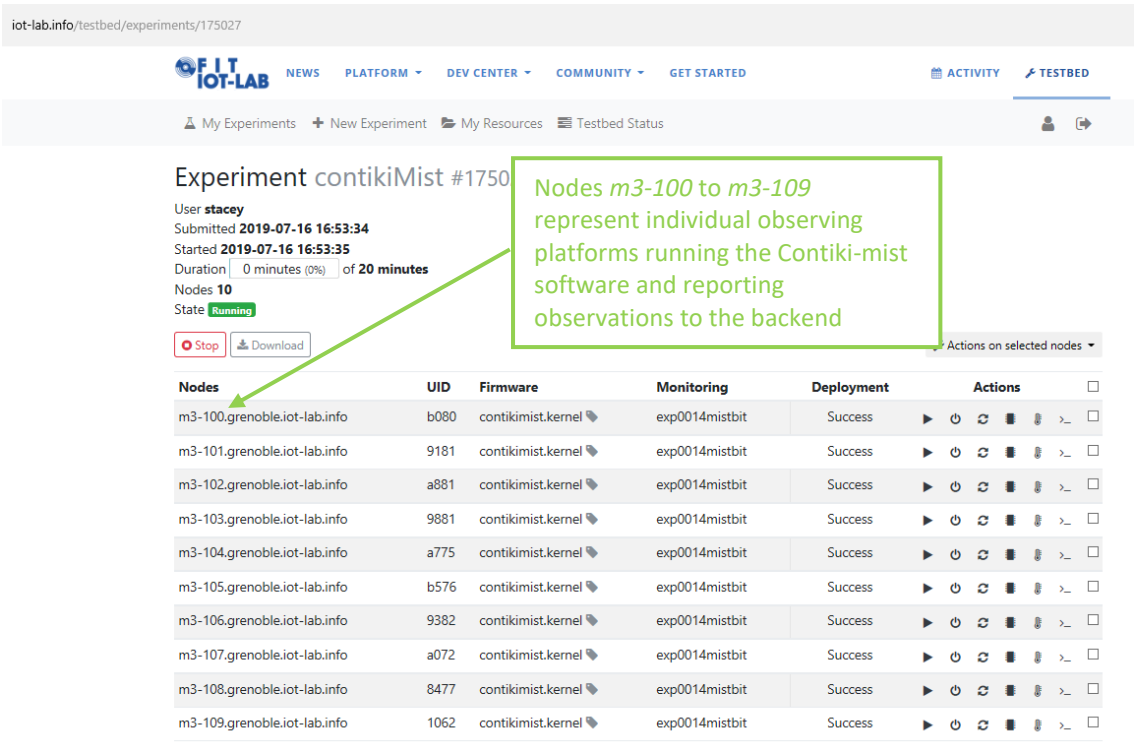


Figure 5.26 Screenshot of Experiment Running on IoT Fit-Lab

Figure 5.27 shows the output of one of the running platforms using the monitor function within the experiment management portal on the FIT IoT Lab Web interface. It can been seen in Figure 5.27 that the platform has successfully registered, and then subsequently connected to the backend two-level modelling based infrastructure, where it received its micro-context schema and is now reporting standardised data elements against the schema to the digital mist backend system. The evaluation validates the ability of the overall infrastructure components (shown in Figure 5.1) to communicate, process and persist multiple incoming observational data-streams.

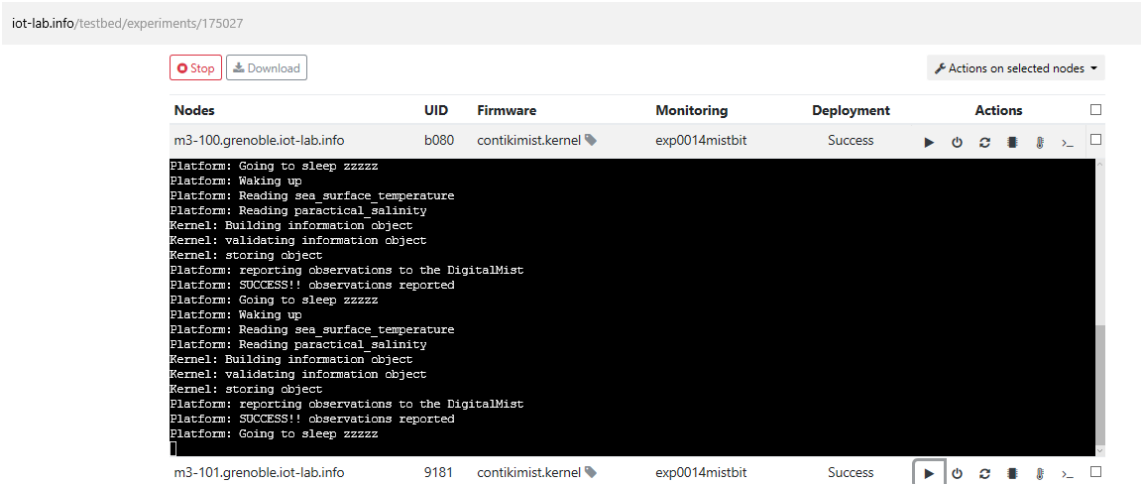


Figure 5.27 Screenshot of Individual Platforms During Experiment.

5.5 Findings and Discussion

Several implementation and deployment investigations were performed before arriving at the approach described throughout this chapter. Of note were attempts to build an ultra-constrained observational node, pushing the two-level modelling to connected nodes running TinyOS (developed in NesC) on MSP40 (g2553[72] + cc2530[73] comms module) based microcontrollers (see Figure 5.28 below). This line of enquiry showed some promise, a prototype system was implemented with communications over an IEEE 802.15.4 wireless link, with client-server RESTful[74] interactions implemented over CoAP protocol. The TinyOS TinyCoAP library was used locally on the observing node (client) and the Java based CoAP implementation Californium (Kovatsch, 2014) provided the basis for the sever side implementation. However, the platform was found to be too constrained for the ESS application environment that is under consideration in this work. Notable limitations observed were the communications latency, basic observations took approximately 5 seconds to report over the CoAP/ IEEE 802.15.4 wireless interface. The

[72] https://www.ti.com/product/MSP430G2553
[73] https://www.ti.com/product/CC2530
[74] Appendix D provides an overview of RESTful approaches on constrained devices using CoAP and IEEE 802.15.4

MSP430's flash memory (at 16K bytes) was completely taken up just loading the base TinyOS operating system, minimal libraries and TinyOS based application code. While running, the generation and processing of a small amounts of observational data would cause the node to fall over due to small 512B SRAM memory issues. The platform technological constraints proved too much of a limiting factor and it proved impossible to implement the OPTaaS protocol defined in chapter 4 within such an environment.

Also, of note was that the TinyOS operating system was a challenging environment to work with. The use of NesC as the application development language gave little room for code portability. Today, TinyOS is very much a niche developmental platform and Contiki-NG proved to be a richer and more flexible developmental environment to work in. Ultimately this line of investigation was consigned to future work as it is a significant task that was not directly linked to addressing the project's research objectives of assessing whether two-level modelling can be translated to the geo-spatial domain. After several other attempts to develop a deployment environment, the FIT IoT Lab infrastructure and the Digital Ocean hosted Linux cloud platforms proved to be much more conducive to meeting the core research objectives.

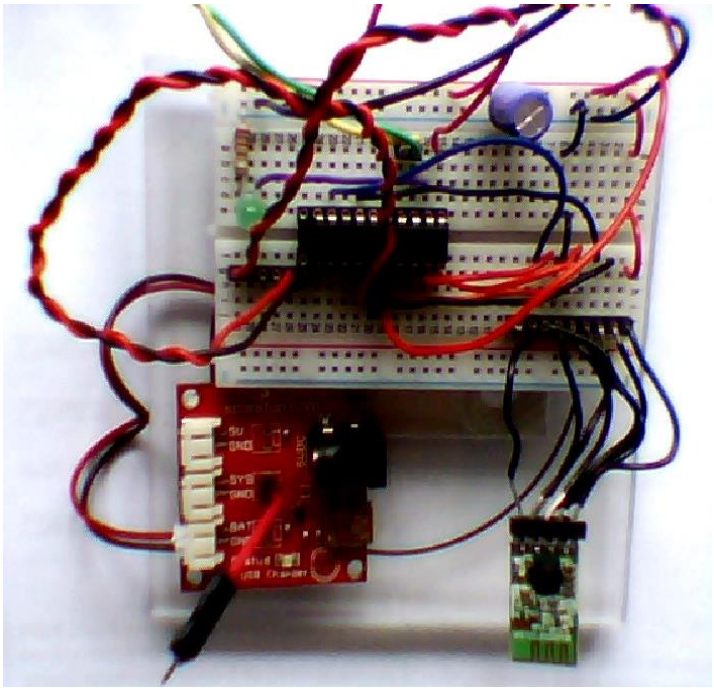


Figure 5.28 MSP430g2553 based constrained node prototype. Photo credit: author

Here the overall findings from the development and deployment of the proposed proof-of-concept two-level modelling based geo-observational sensor system described in the previous sections are discussed. The approach used was informed by the design science methodology adopted within this work (chapter 1, Figure 1.1). This chapter described work done within the design science develop/build cycle, where the theories presented in chapter 4 (and informed from review chapters 2 and 3) informed the definition of system artefacts. Through development and deployment, the hypothesised system and associated artefacts have been realised. This in turn has allowed the research theories to be assessed, justified, and refined further as part of the design science methodology to information system's research and development (Figure 1.1).

The intended outcome of the work described in this chapter, is to ultimately have a well-developed reference architecture that validates the concepts presented in chapter 4 and to promote further investigation and development of the two-level modelling approach with Earth system science informatics (this is discussed further in chapter 7)

It was found that realising a practical implementation of a two-level modelling system is not a trivial task and requires a considerable developmental effort. What has been developed here only constitutes a proof-of-concept system for concept validatation purposes, and further development of the framework should form the part of a future research agenda.

Working with the OpenEHR Java Reference Implementation is complex, and the coding detail is complex and has a high learning curve. This is not least down to the complexity of understanding the concepts within the two-level modelling approach. Nevertheless, a successful base implementation has been realised and the augmented O&M information was successfully implemented and the implemented framework was shown to successfully support fine grained constraining of data objects against the O&M based reference model and archetypes through the processed operational templates. From this it can be concluded that the two-level modelling approach can be applied to geo-observational system scenarios of use while leveraging existing data models re-profiled to enable archetypes to be defined against extension points within the model. Thus it can be said that the system developed here has shown that once domain-based information models are stable within the domain, flexible and future-proofed systems could ultimately be derived using the approach here as hard-coding need only occur against the stable reference model and the system may remain flexible to additional application specific constraints needs once those requirements are defined within archetypes.

To the author's knowledge this level of flexibility has not been shown to be possible with other approaches being proposed within the literature such as ODM2 (Horsburgh et al., 2016) and the CHARMe project (Clifford, 2016).

Scalability and performance issues exist within the current prototype implementation. While this is somewhat manageable when using cloud infrastructures such as

DigitalOcean, adding additional processing power does have an additional cost and there are opportunities for refinement within the software solution without having to rely on scaling the system's hardware. For example, while generic triple stores are conveniently general, they do force a join for each term in a complex query this results in slow processing of queries. This was not overly evident within the deployment here as the datasets used were small, scaling to global scale data portals with the inclusion of historical datasets will prove problematic. This requires additional investigation.

At the observational platform level, this evaluation has validated the linked data approach presented in section 4.5.1 and was successful in reducing the amount of additional metadata and associated storage and processing implications. The ARM M3 based nodes[75] used for testing include a Cortex M3 32bit CPU, 64KB of RAM and 256KB of ROM. This level of processing power and memory was sufficient small observations. Similarly to the TinyOS node shown in Figure 5.28 the FIT IoT M3 use a IEEE 802.15.4 wireless radio for communications. Again, CoAP was used to implement the OPTaaS/RESTful based client-server interactions. The latency issues described above with the TinyOS/MSP430 node were not observed and so it was concluded that IEEE 802.15.4 based radios and CoAP are appropriate technologies to support small and limited observations reporting. Contiki-NG provides a useful platform to build a federated kernel, however the development environment does not provide many of the libraries required to develop a production ready system and many of the components had to be "hacked" which has led to poor software implementation. Therefore, to fully realise an embedded federated two-level modelling kernel these libraries need to be developed which was outside the scope of this timeframe of this work. Again, it is recommended that this should form part of a future research agenda. Therefore, a refinement to the development here is

[75] https://www.iot-lab.info/docs/boards/iot-lab-m3/

to recommend that ARM A profile[76] based board that supports Linux in the interim would form a more suitable observational platform development environment (this is demonstrated within the next chapter).

The work described in this chapter has contributed to meeting research objectives 1,2,3 and 5 (section 1.5.1, 1.5.2, 1.5.3 and 1.5.5) this is synthesised later in chapter 7 with the outcomes of the next chapter (chapter 6) and the overall research question (section 1.4).

[76] https://developer.arm.com/architectures/cpu-architecture/a-profile

Chapter 6

"Occurrences in this domain are beyond the reach of exact prediction because of the variety of factors in operation, not because of any lack of order in nature."
(Einstein,1941)

6. DOMAIN EVALUATION

Chapter Overview: Chapter 5 presented a concrete implementation of the theories described in chapter 4. Theories described in chapter 4 were constructed through the literature review presented in chapters 2 & 3. Several artefacts were developed and deployed to *justify and refine* their theoretical underpinning (part of the Design Science based build/evaluate cycle methodology, see chapter 1, Figure 1.1).

The purpose of this chapter is to further evaluate the constructed theories described in chapter 4, and to ensure that the outcomes of this work meets the research objectives. While chapter 5 presented findings and a discussion resulting from experiences of building the software components and framework arising from the theories presented in chapter 4, this chapter describes the additional domain specific evaluations performed to evaluate the approach. Where chapter 5 validated the conceptual framework and approach from a solely technical perspective (see chapter 1, Figure 1.2), i.e. a system actor user-centric view; this chapter presents evaluations of the theories from a domain expert and user-centric view. As in chapters 4 and 5, literature review material is again referred to throughout this chapter as part of the assess/refine iterative design science research methodology (see chapter 4 and 5: *chapter overview*).

Two evaluation scenarios are described below (evaluations 3 & 4, chapter 1, Figure 1.2: Research Canvas):

1) An air quality (smart city) monitoring scenario for an Internet of Things use-case. The aim of this study was to show how the modelling approach has wide applicability to the IoT domain using related, emerging IoT standards (discussed in section 6.2 below) and existing geospatial standards.
2) An ocean observing scenario, which shows how the approach can improve the harmonisation of ocean monitoring datasets and as a result improve data assimilation techniques to increase the quality of ocean based estimation models, in this case *cholorophyll-a* estimation models in the North Sea (described in section 6.3 below).

Before the evaluations listed in (1) & (2) above are presented, the information modelling methodology used for both evaluations are described.

6.1 Geo-Archetype Modelling Methodology

Thus far, a structured process for developing archetypes has not been presented within this work. As the development of archetypes outside of health has been very limited to date (see chapter 3), there is no literature describing an appropriate two-level information modelling process to produce high quality non-health-based archetype definitions. Even within health the process has been somewhat ad-hoc to date[77].

More recently, Moner et al. (2018) have investigated the various clinical archetype modelling approaches that have emerged within the health domain over the past number of years. The broad archetype modelling experiences examined by Moner et al. (2018) were used to define a structured clinical based archetype modelling methodology (AMM) which is shown in Figure 6.1 below.

[77] A blog maintained by Dr. Heather Leslie documents a wealth of hands on and practical guidance on experiences of archetyping. Although its focus is on clinical archetypes, the information is also applicable here: https://omowizard.wordpress.com/author/omowizard/

Although the AMM shown in Figure 6.1 is specific to the definition of clinical based archetype models, the author has reviewed Moner et al.'s AMM, and the main activities useful to the development of archetypes for the geo-spatial domain evaluations within this work have been identified (section 6.1.1 below). These activities have been highlighted by the author in Figure 6.1 (in red) and represent the basic steps that have been adopted here to produce archetypes for the use-case based scenario evaluations described within this chapter.

The selected activities from the AMM are described in more detail next (section 6.1.1 below) and within the scenarios listed in (1) & (2) above (sections 6.2 and 6.3) thereafter.

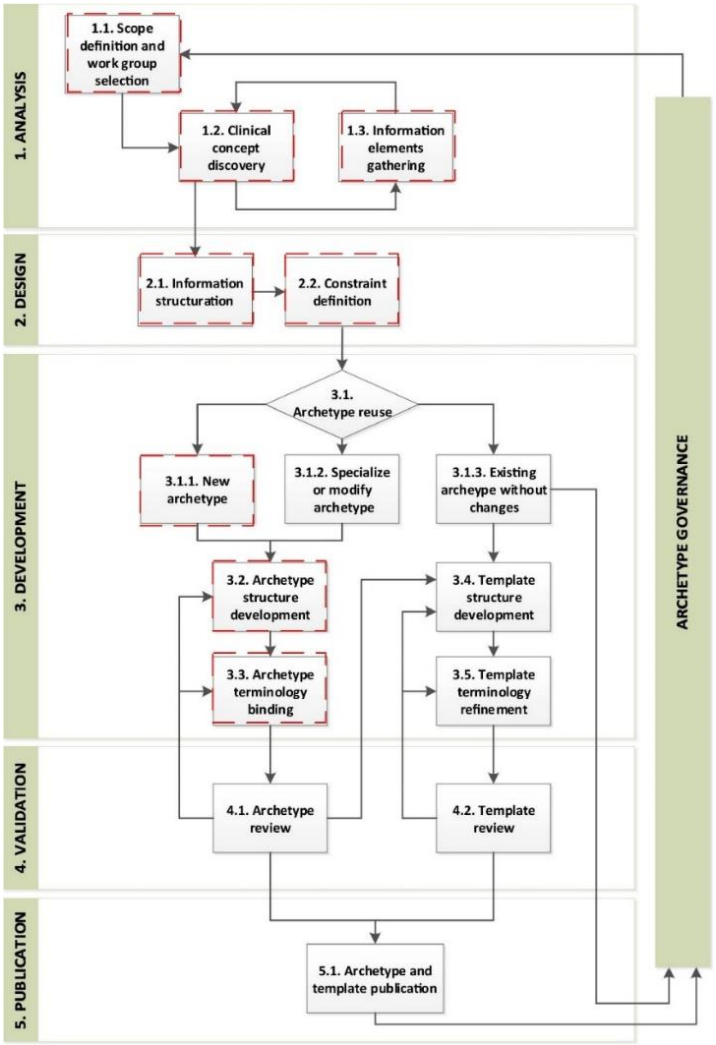


Figure 6.1 Archetype Modelling Methodology (AMM) developed by Moner et al. (2018). Image reproduced from Moner et al. (2018) and activities highlighted using annotations by the author in red, which identify activities relevant to developing archetypes for the scenario evaluations here.

6.1.1 Archetype Modelling Phases

Each relevant modelling phase shown in Figure 6.1 is described briefly in turn below and the activities associated with each step are also described. The descriptions below are adapted from the descriptions presented by Moner et al. (2018).

6.1.1.1 Phase 1 – Analysis

In the analysis phase (Figure 6.1), the scope of the modelling is defined, and initial domain concepts are discovered. Also, initial information elements are captured. The activities involved in the analysis phase include the following (Moner et al., 2018):

a) **Scope definition**. Here the usage scope of the archetype is defined, i.e. for what scenarios of use is it appropriate to use the archetype. Defining an overly limited scope here may result in archetypes that are limited to a very specific scenario and not useful to the broader community. Too broad a scope may result in a large set of archetypes. Or archetypes that try to document too much. Here it is important to precisely define the limits of the scope and use-cases to be covered.

b) **Domain concept discovery**. This activity involves discovering the domain concepts within the scope of the work. A mind map is typically used here to aid the discovery process. Domain concepts are generic groups of related information involved in the modelled scenarios of use. Multiple archetypes can potentially be derived from a single concept in this phase. The list of concepts must cover the complete scope and requirements defined in part a)

c) **Information elements gathering**. Here the list of specific information elements that are associated with each domain concept is collected. Information elements are *atomic* data items (i.e. they can't be broken down further and represent the lowest level of detail). This activity results in a collection of information elements that will become part of the archetype definition.

6.1.1.2 Phase 2 – Design

After the analysis phase, the next phase is the design phase. In this phase *information structuring* takes place, along with constraint definitions (Moner et al., 2018). The activities listed below are required, and for each activity a template table can be used to aid modelling:

a) **Information structuration.** Here the information elements discovered in phase 1 activity (c) are further organised into archetype definitions. Firstly, those domain concepts that have been further considered to constitute an archetype are captured at a high level using a set of tables based on the table template shown below (Table 6.1).

Table 6.1 Archetype Design Table 1

Archetype Description [ID:]	
Name	
Description	
Recommended use	
Leader	
Participants	
Notes	

b) **Constraint definition.** Once the various archetypes have been agreed on, a more detailed design step takes place. Each archetype is further defined based on the details shown in Table 6.2 below. Note, these are the additional constraints that will ultimately be employed by the system against the reference model while instantiating information objects of those types defined within the reference model. Each archetype is a constraint model against already existing concepts within the reference model (or 1st level within the two-level modelling approach).

Table 6.2 Archetype Design Table 2

Archetype Design [ID:]					
Information Element	**Description**	**Mandatory**	**Repeatable**	**Class/Data Type**	**Domain**

6.1.1.3 Phase 3 – Development

Next, is the development phase. Development consists of archetype structure development, terminology binding and template structure development (Moner et al., 2018).

For this work, no pre-existing archetypes are assumed (initially) and so the activities shown in Figure 6.1 are reduced to those highlighted in red (in Figure 6.1), namely:

a) Archetype structure development

b) Archetype terminology binding

c) Template structure development

This phase requires the use of an archetype modelling tool such as those discussed in chapter 3. The modelling tool used supports the reference model for the domain and the archetype definitions captured using design table 2 (Table 6.1 above) are used as the archetype reference details to be captured using the archetype modelling tool.

Of note, is that the process thus far is not overtly technically challenging and is normally carried out by domain experts using a community consensus approach (see Figure 5.2, domain experts interacting with the system development view by way of an archetype editor). During the modelling process, archetype modelling sessions are organised with domain stakeholders as participants, where discussions and deliberations around archetype definitions are typically moderated by a suitably experienced two-level

modeller. This contrasts with the object-oriented design process that relies on more complex modelling concepts and OO visual language symbols and concepts.

The key enabler of this process is the non-technical and accessible mind map approach used, which as discussed throughout this thesis, enables domain experts to become the primary drivers of the information modelling process.

6.1.1.4 Phase 4 – Validation

Archetype validation consists of a review of both the developed archetypes and associated templates to ensure accuracy and adherence to the agreed design tables as part of phase two (described above).

6.1.1.5 Phase 5 – Publication

Validated archetypes and templates are published within the appropriate community repositories. Archetypes and Templates are made available to system developers and domain experts to facilitate their specialization and reuse. In this work GitHub serves as the archetype library repository (or DKM, see section 5.1.4).

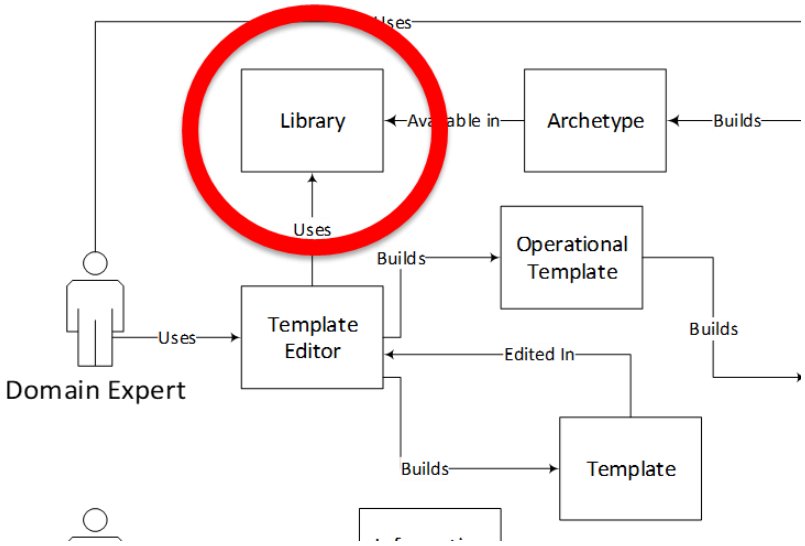


Figure 6.2 Archetype library highlighted within the context of system level view presented in chapter 5 (Figure 5.2)

For this work a set of paper-based templates were produced to enable a low-tech modelling process to be undertaken (Figure 6.3). This process was presented to participants during a tutorial delivered by the author to attendees at the IEEE 5th World Forum on Internet of Things (in Limerick, Ireland, 2019) (Stacey and Berry, 2019b).

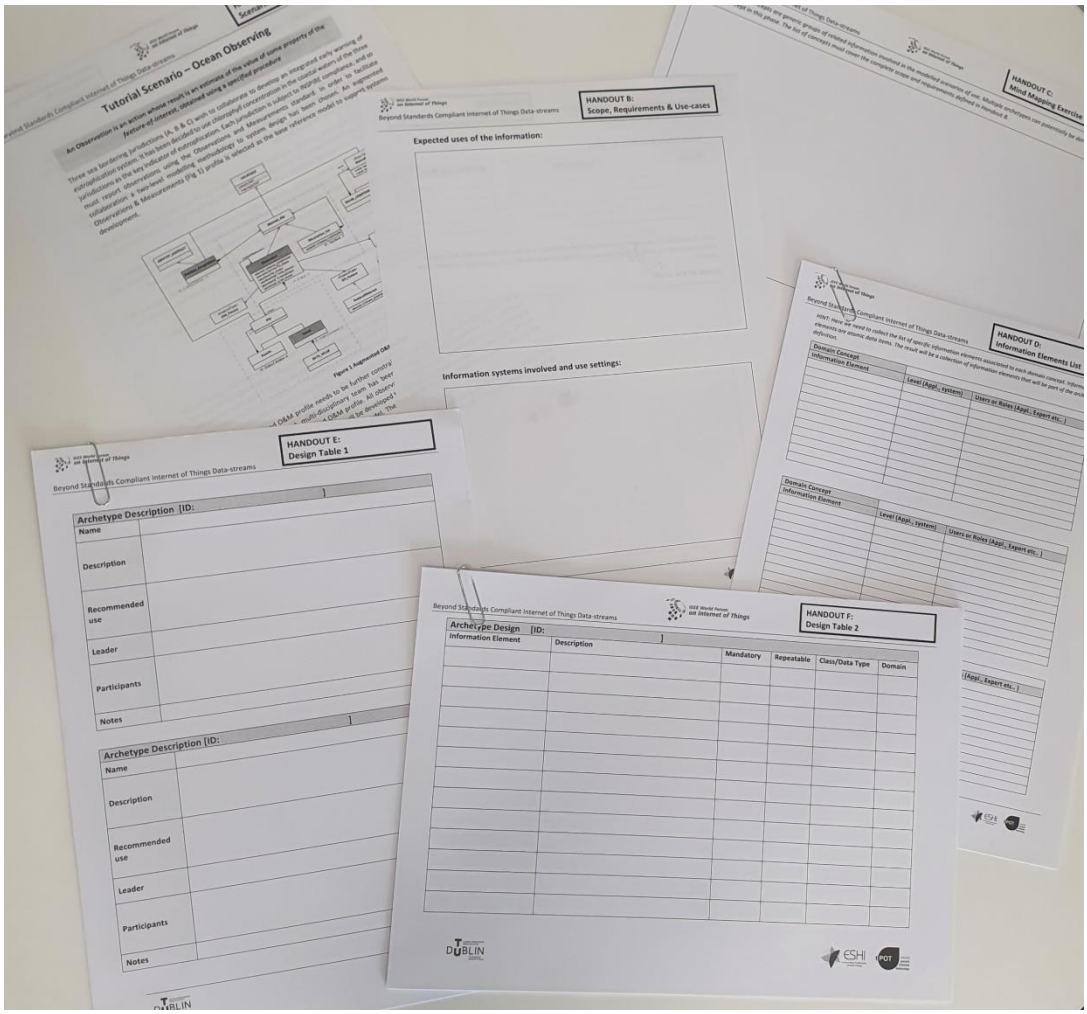


Figure 6.3 Templates produced to support a paper based geo-spatial archetype modelling methodology (Stacey, Berry 2019b)

For both evaluations presented in this chapter, the paper-based process was used for Phases 1- 3 described above (using templates shown in Figure 6.3). It should be noted that, archetypes developed within the air quality scenario were developed by technical participants and not air quality domain experts. Once archetype structures were developed, the LinkEHR editor was used to define the ADL representations of the

required archetypes (described previously in chapter 4, section 4.4, and Figure 4.8). The augmented O&M reference model defined in chapter 3 (Figure 4.5) served as the two-level modelling reference model. The serialised form of the O&M based reference model was loaded into the LinkEHR archetype editor to enable archetype definitions within the tool (Phase 2, activity [a]).

Next, the two use-case evaluations are presented, starting with a basic air-quality/IoT use-case before a more detailed and complex ocean observing scenario is described.

6.2 Interoperable Smart Cities Evaluation

Before the details of this evaluation are presented, the rationale and background to the scenario are described. These introductory descriptors preceding both scenario evaluations are included to show the relevance of each scenario to the overall motivations for this work described in chapter 1 (section 1.1).

The United Nations Population Fund (UNPF) noted the year 2008 as the transition point beyond which more than half of the World's population now lives in urban areas. This trend is expected to continue for the foreseeable future. According to the UNPF, by 2050 the current trend towards greater urbanisation will see another 3 Billion people added to the worlds already densely populated city environs.

Managing complex city infrastructures to meet sustainable development goals requires data and the realisation of *smart cities*. There are many accepted definitions of a smart city. In the context of this work, the definition of Smart City proposed by (Harrison et al., 2010) is assumed:

> *"An instrumented, interconnected and intelligent city .... Interconnected means the integration of those data into an enterprise computing platform and the communication of such information among various city services. Intelligent refers to the inclusion of complex analytics, modeling,*

optimisation, and virtualisation in the operational business processes to make better operational decisions".

Harrison et al. (2010) clearly articulates the barriers to achieving smarter cities. Open standards are highlighted as the foundation for avoiding what they refer to as "frankenmodels"; models composed of incompatible components producing invalid simulations. Metadata semantics, based on existing standards, and extended where necessary are advised (Harrison et al., 2010). This is in keeping with the benefits of adopting two-level models.

This study evaluates the applicability of the methodology described in this work and the support framework for a smart city use-case. The aim of the study is to further evaluate the approach described in previous chapters as applied to resource constrained applications and deployments (in this case a smart city deployment). This evaluation also aims to demonstrate how the solutions described in this thesis meet objectives 3, 4 and 5 (chapter 1, Figure 1.2 and sections 1.5.3/4/5) and to show how the approach can address some of the open research questions with specific Earth system science based in situ observational scenarios (such as sustainable management of city resources using smart city technologies).

Firstly, a basic use-case scenario is presented (below), which informs the modelling requirements. Then, an assessment of the domain in terms of available information models is performed, and a mapping exercise between information structures is described.

6.2.1 Smart City Modelling Scenario

In this scenario[78] a requirement to develop a city scale IoT deployment to monitor environment (air quality & noise) and mobility is assumed. The sensing aspect of the system is deployed to observe the relevant variables for each of the desired phenomena.

6.2.1.1 Scenario description

Limerick City's Digital Strategy seeks to enable Limerick to become a smart, sustainable city. The digital strategy aims to raise Limerick to level 4 "advanced" digital maturity by 2020. Six smart Limerick domains have been defined, including "Urban Places & Spaces" and "Environmental Practices". Several programmes are being implemented to advance the smart limerick domains. Programme 5 "Data & Analytics" has a number of projects, of which the output can be seen here: http://insight.limerick.ie/.

For the purposes of this evaluation, a hypothetical new project: "INSIGHT Limerick – Air Quality" is assumed. The aim of this new project is to provide fine-grained detail of the air quality at key urban locations & spaces, and to inform decision making about environmental practices within the Limerick region. Air quality data will be published under a Data-as-a-Service framework based on the SensorThings API. Allowing all citizens to access and contribute to the service.

Limerick City has an obligation to publish open data and is subject to INSPIRE compliance, and so should report observations using the Observations and Measurements standard (under the environmental facilities theme, discussed in chapter 3). To facilitate collaboration, let us assume that a two-level modelling methodology to system design has been chosen. An augmented Observations & Measurements (see Figure 4.5) profile as described in section 4.4.3 is selected as the base reference model to support systems

[78] This scenario was originally developed as part of a hands-on tutorial delivered at the 5th IEEE World Forum on Internet of Things, Limerick in April of 2019 (Stacey and Berry, 2019b).

development. Data will be modelled based on O&M, the SSNO vocabulary[79] and using terms within the EF INSPIRE theme.

The augmented O&M profile needs to be further constrained to ensure semantic interoperability across heterogenous systems. All reporting platforms will report observations adhering to a data quality constraint model. Ultimately the work here should enable a future application to be developed to consume the air quality observations and generate alerts and information based on an Air Quality index. The first task is to review the application domain before appropriate archetypes for the system to use can be developed using the AMM described above.

6.2.2 Application Domain Review

Before the two-level modelling approach is applied to the scenario above, a technical review of the application domain was performed. The purpose of the review was to ascertain a realistic baseline of typical deployment technical details and typical standards adoption with the described scenario. To promote adoption of two-level modelling within a new domain, it is important to show how the approach can complement existing technologies and standards to encourage up take and *buy-in* within the application domain community.

The results of the review are described below, where a typical air quality sensing platform is defined, and relevant data models within the domain have been identified for further review against the two-level modelling approach.

6.2.2.1 Air Quality Sensing Platform Description

An air quality sensing platform will be deployed by the council consisting of sensors to observe the following properties:

[79] https://www.w3.org/TR/vocab-ssn/ (https://lov.linkeddata.es/dataset/lov/vocabs can be used to find other useful terminologies and ontologies).

- Temperature; Precipitation; Wind Speed; Wind Direction; Luminosity; Noise; Particles; CO (Carbon Monoxide); NO2 (Nitrogen Dioxide)

For the purposes of the scenario, twenty of these platforms will be deployed initially at various locations around the City. The platforms may be moved to different locations from time to time. The platforms will be calibrated regularly based on a defined calibration strategy. The system will produce an Air Quality Index based on the Ambient Air Quality and Cleaner Air for Europe (CAFE) Directive[80]. The system is scalable to allow other third-party organisation and citizen deployed platforms to contribute to the air quality dataset.

In considering this scenario, a review of other related projects found that although the INSPIRE framework mandates that O&M be employed in these monitoring scenarios by 2021, there is little up take of a standardised approach to data representation within similar systems. For example, the Ireland based iSCAPE project[81] (Smart Control of Air Pollution in Europe) does not adhere to a standard data model for the publishing of its observations. Similarly, other air quality monitoring activities such as those performed by the Copernicus programme using the Sentinel-5P satellites, again do not publish their datasets with observational data adhering to a widely agreed data model such as O&M. While Sentinel-5P air quality datasets are disseminated using the netCDF format, the information-model that is used within the netCDF format does not conform to a standard data model such as O&M[82]; despite the provision of an EU Ambient Air Quality Reporting Data Model[83] within INSPIRE. The netCDF format is discussed in more detail

[80] https://www.epa.ie/air/quality/standards/
[81] https://www.iscapeproject.eu/about/
[82] https://scihub.copernicus.eu/
[83] https://aqportal.discomap.eea.europa.eu/toolbox-for-e-reporting/data-model-and-schema/inspire-data-specifications/

as part of the ocean observing evaluation presented in section 6.3 and not as part of this evaluation.

An example of best practice was found by Kotsev et al. (2016), where they describe the architecture of the AirSenseEUR platform, including results from deploying the platform. The AirSenseEUR platform seeks to tackle interoperability issues in air quality monitoring using low cost and open hardware and software by adhering to the INSPIRE directive implementing rules and using OGC compliant standards and service interfaces. The backend system of AirSenseEUR uses the 52 degrees North open source libraries. Kotsev et al. (2016) also directs the reader to the paper Castell et al. (2013) which provided a comprehensive review of similar types of air quality projects at the time. For a more up to date review the reader can refer to Morawska et al. (2018).

In any case, the various deployments described within the literature were found to contain inconsistencies in their implementations of standards. Also, standards have again progressed and evolved beyond the adopted implementing approach for the systems reviewed. Since O&M and INSPIRE have been defined, the SensorThings API data model (Liang et al., 2016) has emerged as a new IoT based profile of O&M with a standardised supporting a RESTful architecture. It can be said that the community standards (i.e. SensorThings API as an evolution of O&M) have evolved beyond what systems adoption has already occurred in the field. This problem was referred to by Beale (2002) as ‘creeping system obsolescence’ (see chapter 3). As such, for this scenario, the author has assumed that the most promising data model standard applicable here is the O&M based SensorThings API data model, which is presented next.

6.2.2.2 SensorThings API

The OGC SensorThings API is divided into two main parts, the *sensing part* and the *tasking part*. The tasking part is the subject of future work within the OGC. This study is concerned with the more mature, sensing part (shown in Figure 6.4 below).

The SensorThings API follows a rich set of principles, conventions, and protocols, specifically aimed at resource constrained sensing devices. For example, the API defines a RESTful based standard to enable CRUD (create, read, update, delete) based interactions for the requesting and reporting of sensed data, similar to OGC's Sensor Observation Service. The sensing part also defines a data model that is based on the ISO/OGC O&M data model. The alignment with O&M can be seen in the entities defined within its data model, specifically *Observation* and *FeatureOfInterest*. In addition, the following entities are also defined: *Thing*, *Locations*, *HistoricalLocations*, *DataStream* & *Sensor* (Figure 6.4).

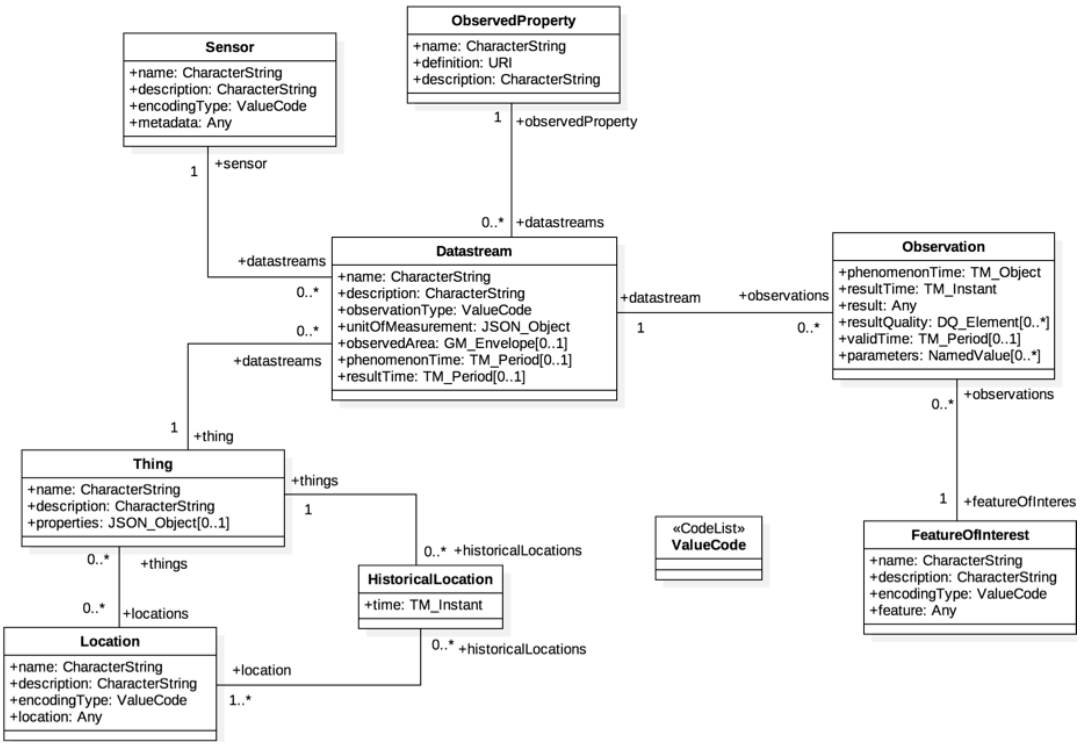


Figure 6.4 SensorThings API Data Model. Image reproduced from (Liang et al., 2016)

Much like O&M, the SensorThings API data model enables syntactic interoperability between heterogeneous IoT systems. Semantic interoperability is however limited. As discussed throughout this thesis, semantic integration goes beyond combining data points solely based on syntactic representation. Typically, ontological bindings - within datasets - are used to record the meaning of the captured data. Of note are an increasing number of ontologies available within the IoT domain that can be used to enable semantic interoperability within IoT scenarios (Bajaj et al., 2017).

6.2.2.3 SensorThings API as a Reference Model

Initially it would appear that the SensorThings API data model could serve as an appropriate reference model to underpin a two-level modelling approach within IoT systems, much like O&M (see chapter 4). To assess whether this is the case, the SensorThings API data model was assessed against the characteristics of a reference model.

As noted previously (section 4.3), reference models should only capture the stable concepts within a domain, at the principles level within a multi-level ontological space (Figure 4.2). In chapter 4, the O&M standard's suitability for two-level modelling was examined. It was concluded that O&M lies just above the principles ontological level (see chapter 4) but given the maturity and wide acceptance of O&M within the community and its adoption within the INSPIRE directive, it is pragmatic to choose O&M to underpin archetype definitions. After examining the SensorThings API data model in detail it was found that it extends O&M beyond the principles ontological level (Figure 6.5).

The review here of the SensorThings API data model concluded that concepts such as *DataStream* are in fact lower level organizational concepts within the IoT domain, and so should be defined within the archetype model and not used as reference model concepts. Therefore, for this evaluation it was concluded that the augmented O&M reference model

defined within this work should be the base reference model to develop the air-quality monitoring system against. However, O&M extensions within the SensorThings API data model should be re-used as part of the 2nd level. To achieve this, a concept mapping exercise was undertaken to redefine the concepts at their respective two-level model levels (reference model and archetype model).

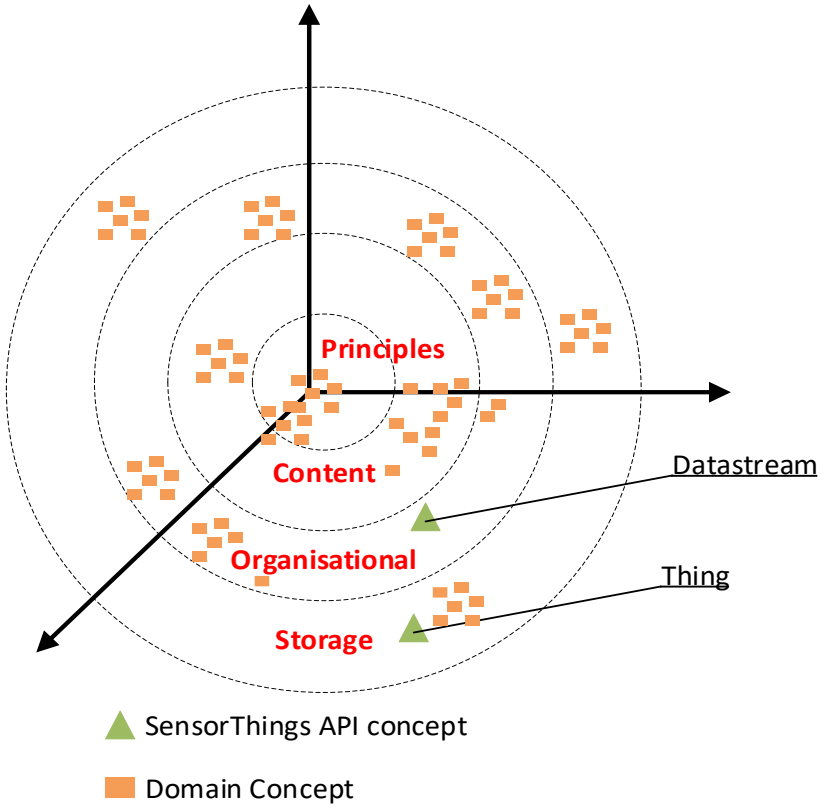


Figure 6.5 SensorThings API Ontological levels. Shown is that SensorThings API concepts lies within the content, organizational and storage levels in a multi-level knowledge space (see chapter 4, Figure 4.2 & Figure 4.6). Table 6.3 below shows a more detailed concept mapping.

6.2.3 Concept Mapping

The domain concepts provided by the SensorThings API were mapped to base concepts available within the augmented O&M reference model (Figure 4.5). The results of the concept mapping exercise are shown in Table 6.3 below.

During the mapping exercise it was found that each new concept introduced by the SensorThings API sensing part can be characterised as a constrained version of an O&M

based reference model concept. For example, it can be seen in Table 6.3 below that *Thing* was mapped as a constrained storage level concept (referred to as a COMPOSITION) and *DataStream* can be mapped as a constraint definition of the organisational concept Observation_set (referred to as a SECTION, see also Figure 4.5).

Table 6.3 SensorThings API Concept Mapping, SensorThings API sensing part 1 (Liang et al., 2016) is mapped to the Augmented O&M base concepts (Figure 4.5).

SensorThings API[84]	Definition	Augmented O&M Base	Comments
Thing	A representation of some physical or virtual entity, equipped with one or more sensors. Sensor Platform	COMPOSITION – GeoData_Composition *(Storage concept see Figure 6.5)*	Thing is a domain concept that is a specialization of the reference model concept GeoData_Compostion
Datastream	A concept that groups Observations	SECTION – Observation_set (*Organisational concept see Figure 6.5)*	Datastream is a domain concept that is a specialization of the reference model concept Observation_set
Sensor	The procedure used in the observation	OM_Process *(Content concept see Figure 6.5)*	Sensor is a constraint on the empty O&M class OM_Process, which is defined using SensorML.
Location	A representation of the Thing's location	Details_COMPOUND (*Content concept see Figure 6.5)*	Geodata_Composition contains an attribute "details" of type Details_COMPOUND which is an aggregation of Details_ELEMENT.
Observation	Act of measuring or otherwise determining the value of a property	Observation *(Content concept see Figure 6.5)*	Semantically equivalent
FeatureOfInterest	The focus of the observation	FeatureOfInterest (*Content concept see Figure 6.5)*	Semantically equivalent
ObservedProperty	The property observed of the feature of interest	ObservedProperty *(Content concept see Figure 6.5)*	Semantically equivalent

[84] http://docs.opengeospatial.org/is/15-078r6/15-078r6.html

It is important to note that these mappings have a deeper consequence that may not be obvious to the reader at first. In chapter 4, the augmentation of O&M with additional design patterns (namely recursive aggregation patterns) was undertaken to transform O&M from a model of reality to a model of documentation. According to the methodology set out in chapter 4, *Thing* and *Datastream* have been mapped to both storage and organisational concepts (see Table 6.3 and Figure 6.5 above). This mapping changes the nature of those entities. As both *Thing* and *Datastream* are mapped to documentation concepts (for example GeoData_Composition), this also transforms the SensorThings API towards a model of documentation, which in turn changes the intention of the original SensorThings API information model. Popper's three worlds theory presented in chapter 3, provides the basis for illustration.

For example, *Thing* within the SensorThings API model is of world 3 (as it is symbolic of a world 1 phenomenon (see Table 6.3 above), and in its current structure has a direct relationship to world 1, i.e. it is a concrete representation of a world 1 phenomenon. Therefore, world 1 directly contributes to the world 3 *Thing* entity. Through the mapping process, *Thing* remains within world 3, but the direct relationship to world 1 is removed. *Thing* is now contributed to directly or via world 2. This is an intentional consequence of the mapping through the translation methodology described in chapter 4, and is in line with the objectives of this work, namely to provide a rich framework for the recording of knowledge produced within the geospatial domain (see chapter 1, section 1.5 and chapter 3, section 3.1.1).

Through the concept mapping process, Thing and Datastream now become *represented knowledge* and therefore should be encoded as archetypes or constraints of the reference model classes GeoData_Composition and Observation_set and not within

the reference model itself (i.e. they are more volatile knowledge concepts, see section 3.5).

In the augmented O&M reference model (Figure 4.5, chapter 4), an Observation_Set can be composed of numerous Observations of different ObservedProperty instances. However, within the SensorThings API data model this needs to be further constrained to only allow Datastream to contain Observations of a singular ObservedProperty; *Thing* then contains numerous Observation_Sets or collections of Observations.

Sensor is the procedure used in the measuring of, or otherwise observing of a property of the feature of interest. It can in fact be mapped as a constraint on the reference model concept OM_Process (Figure 4.5).

6.2.3.1 SensorThings API as an Archetype Model

The resulting SensorThings API archetype model contains numerous resulting archetype definitions, that were defined using the LinkEHR tool (Figure 6.6 and Figure 6.7).

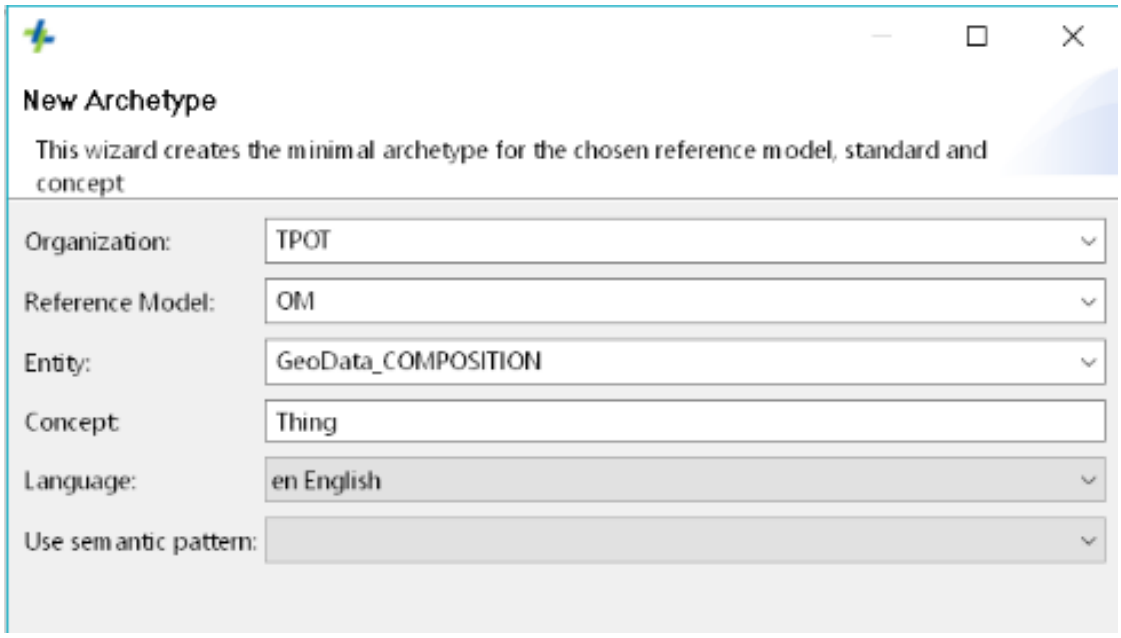


Figure 6.6 Using the LinkEHR multi-reference model editor, an XSD representation of Figure 4 is used to define the SensorThings API archetype model. Here the concept Thing is a set of constraint statements on the reference model concept GeoData_COMPOSITION.

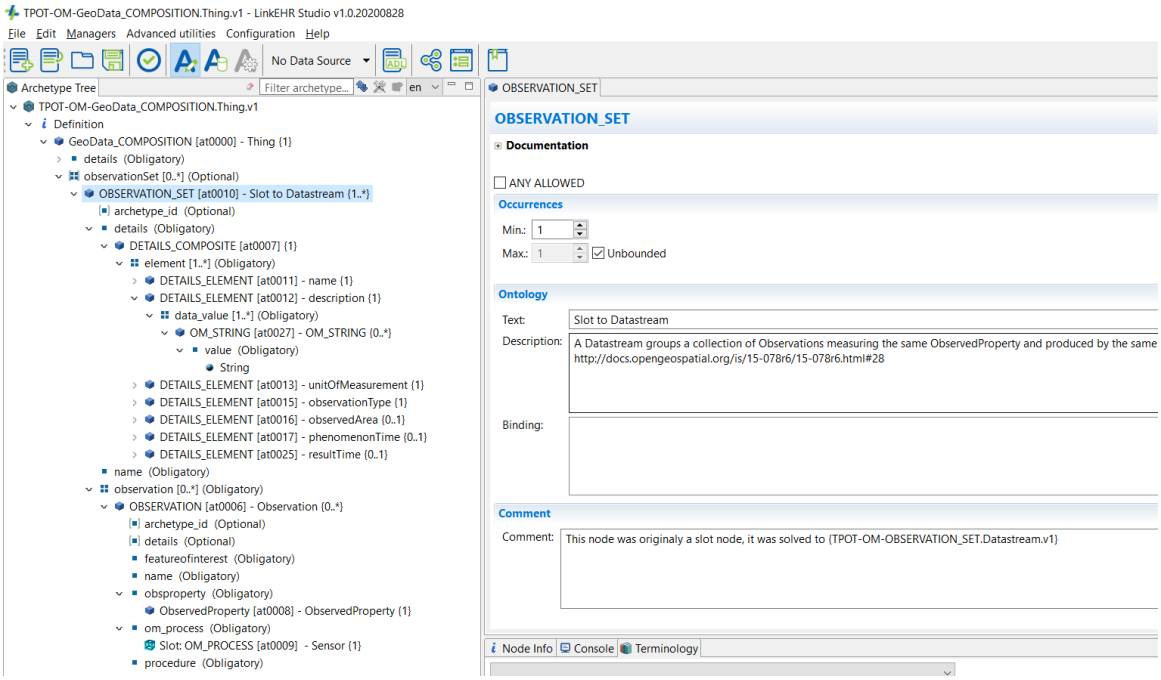


Figure 6.7 LinkEHR defining the constraint Thing: Thing may have 1..* Datastreams. Datastream is a reference model type Observation_set, and here an archetype_slot is created to plug in an archetype of type Observation_set. Archetypes are bound by the underlying reference model.

The resultant SensorThings API Archetype Model can be found at the GitHub based Domain Knowledge Management archetype library[85].

Next, the archetype modelling methodology described in section 6.1 above was applied to the scenario described in section 6.2.1. For brevity, the description below follows the development of only a small number of newly defined archetypes and archetypes that are specialisations of SensorThings API based archetypes. The data specification defined within the INSPIRE Environmental Monitoring Facilities is used to inform concept naming[86] and SSNO is used for term bindings.

Below, some of the archetypes developed during the archetype modelling and mapping exercise are listed[87]. The development of these archetypes was partly informed by participants of the workshop mentioned in section 6.1.1.5, and then further refined by the

[85] https://github.com/pstacey/geo-archetype-library/tree/master/SensorThingsAPI

[86] https://inspire.ec.europa.eu/documents/Data_Specifications/INSPIRE_DataSpecification_EF_v3.0rc3.pdf (see section D.3.1.3).

[87] A full listing of Air Quality archetypes can be found in the GitHub hosted DKM here https://github.com/pstacey/geo-archetype-library/tree/master/air-quality

author. It should be noted that the workshop participants were largely made up of technical experts, rather than air quality domain experts, and therefore the resultant archetypes are for illustrative purposes only. This modelling limitation is discussed further in chapter 7.

As can be seen, some archetypes are specialisations of the SensorThings API data model-based archetype concepts, and some are new archetypes which constrain concepts defined within the augmented O&M model. For example, *AQ_Station* highlighted in green below is a specialisation of the concept *Thing* which is a constraint model on the reference model concept *GeoData_Compostion*.

- `TPOT-OM-ObservedProperty.AirQuality.v1`
- `TPOT-OM-Observation_Set.DataStream.v1`
- `TPOT-OM-GeoData_Composition.Thing-AQ_Station.v1`
- `TPOT-OM-Details_COMPOUND.Pollutants.v1`
- `TPOT-OM-Details_ELEMENT.NO2.v1`
- `TPOT-OM-Details_ELEMENT.CO.v1`
- `TPOT-OM-Details_ELEMENT.VOC.v1`
- `TPOT-OM-OM_PROCESS.Sensor.v1`

6.2.3.2 Scenario Operational Template

A hypothetical operational template (.opt) file was subsequently generated from the resulting archetypes defined in the previous step. This operational template represents the specific scenario of use defined within a template document *`TPOT-OM-Geo_Data_Document.LimerickCityAQ_Report.v1`* (see appendix C).

The resultant .opt file can be used to build real systems using the software components that support the augmented O&M reference model and linked data approach presented in chapter 4 and 5. The resultant operational template can allow air quality monitoring

stations to constrain information objects during runtime against the O&M based reference model concepts within the context of the linked data approach and report constrained data quality rich observations to the backend supporting infrastructure described in chapter 5.

6.2.4 Smart City Domain Findings & Discussion

It can be seen from the literature that progress towards interoperable city scale monitoring is slow. Most research in the area is still making progress towards INSPIRE compliance with only a few projects going beyond INSPIRE compliance to handle variance within specific use-cases.

The problem of system obsolescence was observed against evolving standards within the domain under investigation. For example, within the INSPIRE EF Data Specification an Environmental Monitoring Facility application schema is provided. A new class called EnvironmentalMonitoringFacilities is included which has a relationship of 0..* with the O&M class Observation. Also, within the Air Quality EF technical guidance, it is recommended that AQ_Sensor is a specialisation of EnvironmentalMonitoringFacilities. The approach developed within this evaluation is flexible enough to capture this technical guidance using archetypes. `EnvironmentalMonitoringFacilities` could be captured as an Archetype called *OM-TPOT-GeoData_Composition.EnvironmentalMonitoringFacilities.v1* instead of a hardcoded system implementation and the AQ_Sensor could in turn be captured by a new archetype: *OM-TPOT-GeoData_Composition.EnvironmentalMonitoringFacilities.AQ_Sensor.v1*. Again, this shows the flexibility of the approach and further validates the wide applicability of the augmented O&M model shown in Figure 4.5.

The concept mapping and subsequent encoding of the SensorThings API data model as a set of archetypes, or constraints on the augmented O&M also showed the issue of inconsistencies within standards evolution, causing creeping system obsolescence. For

this work Datastream has been modelled as a set of constraints (or archetype) specialising the base augmented O&M concept Observation_Set. However, it was found that it was not appropriate to capture the exact definition of Datastream as is defined within the SensorThings API as an archetype. It was found that the full description of the SensorThings API represents a specific scenario of use that is overly specialised and thus represents a *template* definition.

In summary, it was found through this exercise that the approach developed within this work was shown to be flexible enough to meet the requirements of the specific domain use case under investigation. Moreover, the approach shows potential to improve current domain specific interoperability efforts and enable future proofing of systems in the face of evolving standards within the domain under investigation i.e. the two-level modelling approach provides additional control over the evolution of standards. Once the base reference model is appropriate and stable, the two level modelling approach if adopted in this scenario would provide an evolutionary approach to standards development that avoids generating inconsistencies between community standards and thus slowing or perhaps avoiding the creeping obsolescence associated with diverging standards. The wider impact, limitations and implications of this exercise are discussed further in chapter 7.

Next, the second (ocean observing) use-case scenario evaluation is presented. The air quality monitoring scenario evaluation - presented above - did not examine the approach in the context of the geo-spatial two-level monitoring infrastructure presented in chapter 5 and so focused on validating the modelling approach in the context of a simple (non health) geospatial scenario. The ocean observing evaluation - presented next - goes further than the air quality scenario evaluation to investigate the approach using real observational datasets deployed within the physical infrastructure presented in chapter 5.

6.3 Interoperable Ocean Observing Evaluation

Combining the findings from chapter 5 and the evaluation in section 6.2 above, it can be concluded thus far that the theories described in chapter 3 show good potential to be applied to real-world scenarios, and real-time in situ constrained observing platforms. To further investigate this assertion, the approach is now applied to a second observing scenario.

The purpose of this additional evaluation is to further investigate the wide applicability of the archetype modelling approach within Earth system science-based domains. This evaluation investigates how the translated two-level modelling approach, defined in this work, performs with harmonising real-world ocean observing datasets; that are deployed on physical embedded boards (observing platforms).

In this final evaluation as part of this thesis, the benefits of two-level modelling in medium and large-scale ocean observing scenarios are investigated. The aim of this study is to demonstrate, investigate and evaluate the two-level modelling approach's ability to enable the automatic backward federation of ocean based observational data flows, governed by the use of community agreed archetypes using the constrained, linked-data supporting infrastructure (described in chapter 4 and chapter 5). A comparative analysis is used to evaluate the approach against current state-of-the-art deployments (see section 6.3.5 below).

Within this evaluation the approach is again developed and refined as part of the design science paradigm (see chapter 1, Figure 1.1) to justify and evaluate its applicability in helping domain experts to better understand and estimate the mechanisms governing chlorophyll-α concentrations within a defined sea region. A scenario rationale and background, showing alignment with the overall research motivations for this work (chapter 1) is presented next.

It is believed that anthropogenic warming of oceans is increasing the level of phytoplankton in the water column (Barnett et al., 2005). Phytoplankton are microscopic algae and are an important source of aquatic food. However, in large concentrations, algae can have a detrimental effect on marine life and water quality (Deltares, 2018). Excessive algae growth can starve aqua-culture sites of dissolved oxygen and consequently devastate fish stock (Abdel-Tawwab et al., 2019).

Chlorophyll-α (Chlfa) is a photosynthetic pigment and common to all phytoplankton (Deltares, 2018). Chlfa concentrations are used to quantify levels of phytoplankton within water (Schalles et al., 1998) (Honeywill et al., 2002). and can be measured using in situ sensors known as fluorometers or satellite-based sensors. High levels of Chlfa can indicate an algae bloom and is an important indicator of eutrophication (Deltares, 2018). There are many drivers of excessive phytoplankton growth. Typically, there are two primary production drivers, light (irradiance) and nutrients within the body of water (Deltares, 2018).

The development of accurate Chlfa estimation models and prediction systems for individual sea regions is an important area of research. The focus is often on developing computationally efficient estimation models, using other oceanic parameters to estimate Chlfa levels. For example, Irwin and Finkel (2008) have shown that sea-surface temperature combined with latitude/longitude, surface nitrate and irradiance can predict 83% of the log variance in chlorophyll-α in the north Atlantic sea region (Irwin and Finkel, 2008). In Blauw (2015) it was found that sea surface temperature is the best single predictor of log chlorophyll-α.

Observations are key inputs to Chlfa estimation models. Pearlman et al. (2019) summarise ocean observing as a chain of processes addressing *why*, *what*, and *how* to

observe, as well as how to *integrate*, *use* and *disseminate* the outcomes of the observing process. The latter being of relevance to the two-level model approach.

As discussed in chapter 2, satellite-based sensors are an important source of observational data but can only make remote observations at or close to the sea surface. Therefore, marine scientists require in situ ocean observing platforms to be deployed to read below the surface, throughout the water column. Given the platform deployment environment associated with marine monitoring, platforms are often technologically constrained (in terms of access to battery power, communications and on-board processing power and storage). This limits the ability of ocean observing platforms to ensure data quality is enforced at the point of capture (see chapter 2).

The following scenario has been developed to further investigate the applicability of two-level modelling within technological constrained in situ ocean observing platforms. As in the previous use-case evaluation above (smart city, air quality monitoring scenario), an ocean observing scenario is defined below. However, in this instance, real observed historical datasets are used within the evaluation to go beyond the modelling and information requirement definition phase (see section 6.1.1).

6.3.1 Ocean Observing Scenario[88]

For this evaluation, consider the scenario that for the purposes of protecting marine resources, three sea bordering jurisdictions (A, B & C) wish to collaborate to develop an integrated early warning of eutrophication system (see section 2.5.3). Let us further specify in the scenario that it has been decided to use chlorophyll-α concentrations in the coastal waters of the three jurisdictions as the key indicator of eutrophication.

Each jurisdiction is subject to INSPIRE compliance, and so must report observations using the Observations and Measurements standard (see section 2.5.1). Also, to facilitate

[88] This scenario is informed by the INSPIRE Marine Pilot described in chapter 3, section 2.5.1.

collaboration and data interoperability in our scenario a two-level modelling methodology to support the information system design has been chosen. An augmented Observations & Measurements (Figure 4.5) profile is selected as the base reference model to support systems development.

As discussed throughout this thesis, the augmented O&M model is composed of general principles level concepts, and is designed to be very flexible, allowing the same concept to be represented in a variety of ways, so adoption of this model is not a guarantee of semantic interoperability, as conformant implementations of O&M may differ substantially from each other. Therefore, the augmented O&M profile needs to be further constrained and bound to common vocabularies and ontologies to ensure semantic interoperability from all three jurisdictions. Appropriate constraints to the augmented O&M profile must be defined for the given scenario.

All observation moorings will subsequently report observations adhering to the shared constrained model for the given application. This will allow applications to be developed to consume the observations and generate alerts and higher-level information based on more accurate estimation and prediction models outputs.

The first task is to develop the appropriate archetypes for the system to use. These archetypes will ensure the observational dataflows adhere to a concise data model. Once standardised dataflows governed by archetypes (and implemented using the federated two-level modelling approach developed within this work) have been established, these will be available for consumption by third party applications. In this scenario, the standardised dataflows will be used to aid chlorophyll-α estimation within a particular sea area. These estimates will be produced by applications using some estimation model primarily using observed temperature & salinity values (discussed later in the next section).

6.3.2 Application Domain Review

Before engaging in the archetype modelling phases (see section 6.1.1), again a review of the application domain was performed to ascertain a realistic baseline of typical deployment technical details and typical standards adoption within the domain of interest.

The details of the review are described next, beginning with a review of the deployment environment and related work within the specific use-case application domain. First, the ocean observing platform deployment locations (sea regions) are considered.

6.3.2.1 Sea Regions

The North West Shelf (NWS) sea region covers a large area. Sub regions include the Irish Sea and Southern North Sea, among others. The NWS operational oceanography organization (NOOS) (Holt, 2003) includes nine countries that collaborate together to develop ocean observing and prediction systems for the NWS area. The NWS data portal[89] is one product arising from NOOS. NOOS is also part of the European Global Ocean Observing System (EuroGOOS) (Woods et al., 1996).

NOOS operates in the context of the Global Ocean Observing System (GOOS) (Dexter et al., 2010). One of the core goals for GOOS and associated GOOS Regional Alliances (GRA) (Malone, 2006) (of which EuroGOOS is part of), is to develop advanced ocean model-based products. Today there is now a wealth of ocean dynamics models available. The EuroGOOS Ocean models Web tool[90] provides a convenient way to browse and filter the various ocean models that are available for the EuroGOOS area.

A wide variety of ocean models are available for the NWS area. The Dutch Continental Shelf Model (DCSM) model is a well-established hydrodynamic model that was

[89] http://nwsportal.bsh.de/
[90] http://eurogoos.net/models

developed by the Dutch government to improve accurate water-level forecasting (Gerritsen et al., 1995). The Nemo Nordic model (Hordoir et al., 2019) is a specialized model for the Baltic & North Sea, based on the well-known NEMO ocean engine. The GEM/BLOOM model developed by Deltares can be used to estimate chlorophyll-α concentrations and water quality in the North Sea (2008). Other generalized statistical models such as the Generalized Additive Model (GAM) (Hastie and Tibshirani, 1990) (Hastie, 2017) are also often used as a linear predictive model for ocean dynamics.

For this work, the Southern North Sea region was selected as the focus for deployment and investigation. The use-case is motivated by the previous work performed as part of the INSPIRE Marine Pilot (European Commission, 2016). In this use-case, for simplicity, salinity and temperature observations are the data flows of choice. It is reasonable to focus on salinity and temperature as they have been shown to have a strong correlation with chlorophyll-α concentrations in the NWS sea region (Irwin and Finkel, 2008). Also, typically salinity and temperature in situ observations are more readily available within sea regions.

6.3.2.2 Predictability of chlorophyll-a fluctuations

Blauw (2015) shows how the predictability of chlorophyll-α concentrations from environmental variables increases greatly when environmental variables monitored from in situ mooring stations are included within GAM models. Blauw highlights the need for fine grained monitoring of ocean regions through the deployment of in situ observing platforms. Blauw's results also show that the driving forces for Chlorophyll-α fluctuation differ in different regions of the North Sea. This gives weight to the need for high density deployments and harmonised ocean observational datasets.

For this work the adopted approach is: *simple method and lots of observations*. If the model is simple, it is less computationally intensive. Maximizing observations also means

less grid interpolation is necessary. Therefore, the approach seeks to harvest as many useable observations as possible for an area of interest. For the purposes of investigation, a deliberately overly simplified GAM model is used (equation 1). It is assumed that there is an ideal and simplified linear relationship between temperature, salinity and chlorophyll-α concentrations within the southern North Sea region. In equation (1) μ represents mean chlorophyll-α concentrations from previous model runs. A 2-dimensional square grid with 6 grid points is used, and constant depth is assumed.

$$Chla = \mu + f_1(salinity) + f_2\left(sea_{surface_{temp}}\right) + f_3(lng, lat) + f_4(month)$$

Equation 1

A Kalman filter (Kalman, 1960) is used for assimilation of observations into the model. Kalman filtering is a commonly used approach for the assimilation of time series water quality data (Pastres et al., 2003), where a series of measurements observed over time, which contain inaccuracies are used to estimate unknown values (discussed further below). As new observations are discovered using the additional semantic search capabilities provided by two-level modelling - using the OPTaaS system - they are automatically assimilated in real-time into the GAM model. The OpenDA framework is used for this purpose here (Figure 6.8).

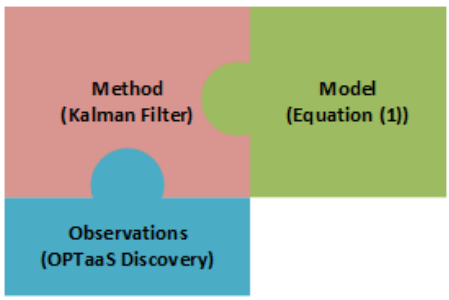


Figure 6.8 The OpenDA model. OPTaaS is used to collect interoperable and harmonised ocean observations adhering to a set of defined archetypes. The Kalman filter is used to assimilate the observations into the GAM model shown in equation 1.

6.3.2.3 Data Assimilation

Data assimilation (DA) is commonly used with ocean models to improve model estimation. Data assimilation optimally blends all information available about a geophysical system to give a consistent picture of its state (Pham, Verron and Rouban, 1998). The most useful information to improve ocean models is obtained from in situ sensor-based observations.

Data assimilation uses measured observations in combination with a dynamic system model to improve the estimates of an ocean system's states (Markensteijn, 2017). Lopez at al. (2016) note the importance of assimilation of appropriate and relevant observations when estimating hydrological variables. However, the discovery, interoperability and thus assessment of an increasing the number of observations and observation points that are assimilated into estimation models greatly improves model forecasting results. In situ observational data are typically considered more accurate and timely and thus once properly described and supported by context information that is semantically coherent across the region of study, they can present an opportunity for more accurate estimations (Ridler, 2014).

Verrier et al. (2017) have shown that a seven-day forecast for sea levels and ocean currents was significantly improved when moving from one altimeter to two. Numerous methods are used for assimilating observations with ocean models. The two main categories are variational methods and sequential methods. Sequential methods are used when assimilation takes place when new observations become available.

Improving the assimilation process is an active area of research. The ensemble Kalman filter is an updated version of the extended Kalman filter and is more computationally efficient. Today ensembles are used to improve forecasting. Ensembles are the combination of results from numerous models. The singular evolutive extended Kalman

filter (SEEK) (Pham, Verron and Rouban, 1998) further improves the assimilation process for oceanography. These developments are largely driven by the increasing availability of ocean observational data, such as satellite oceanography (Parkinson, 2006) and the ability of the filter to evolve as new data becomes available.

There are many tools to aid assimilation such as OpenDA (Verlaan et al., 2010). MOVE (Usui et al., 2006), ECMWF (Balmaseda et al., 2013) and PEODAS (Yim et al., 2011). OpenDA is a free open source data assimilation toolbox primarily written in Java. OpenDA is actively used in several other assimilation projects and tools such as SANGOMA (Van Leeuwen et al., 2011).

Next, a review of state of standards development and data interoperability efforts within the featured domain is presented. As in section 6.1, it is important to understand the complexities of the domain data to which the approach is being applied. Within these domain evaluations there already exists a wealth of data, standardisation work and deployed observing systems and SDIs. The domain data interoperability assessment below ensures that the approach aligns with work already progressing within the domain.

6.3.2.4 Domain Data Interoperability Assessment

Blauw et al. (2012) illustrate the complexity of working with in situ observed ocean datasets. In their work they obtained observations from the Cefas operated WARP (TH1) NMMP SmartBuoy (WARP CEFAS- 62010720). The observations obtained were subsequently used to examine the interplay between coastal phytoplankton and the tidal cycle. The observations were downloaded directly from the Cefas website[91]. Based on the instruments used and the calibration information available, several data cleansing steps were required to ensure the data were suitable for analysis.

[91] https://www.cefas.co.uk/cefas-datahub/cefas-data-hub-apis/

For this evaluation, datasets for WARP CEFAS-62010720 obtained from the EMODnet-physics portal were examined by the author (see Appendix C). The datasets include the quality check data from the CMEMS INS-TAC processing centres (discussed in chapter 2, section 2.5). These quality checks perform several functions such as spike detection and statistical controls; more details can be found in (Wehde et al., 2016). However, the additional information required for the data cleansing steps conducted in (Blauw et al., 2012) is not encoded either directly or indirectly in the dataset; even O&M extensions do not mandate this level of interoperability. This example illustrates the requirement for a mechanism that allows organizations to further constrain and describe their information based on individual platform deployments; referred to as an extensibility mechanism within a digital Earth system (see chapter 1).

INS-TAC regional centers, described previously in chapter 2 (see section 2.5.2), provide additional quality and validation of datasets, and produce a final "quality checked" (QC) data product from the raw observational data received. The regional centers use the *oceanotron* server, which disseminates the QC observational data flows using the OceanSITES for Copernicus standard, consisting of netCDF CF and to an extent O&M compliant data representation.

The OceanSITES for Copernicus standard is hard coded into the oceanotron software. Therefore, oceanotron will be subject to the creeping obsolescence described by Beale (2002) and noted in the previous domain evaluation (see section 6.2); as ESS data standards evolve based on the rich and growing community of supporters. This is already evident as oceanotron uses CF conventions version 1.6. CF conventions are at version 1.9-draft[92] (checked December 2020). This requires the oceanotron software to be updated and re-distributed to centres. Presently, this is not a difficult task as the number

[92] https://cfconventions.org/documents.html

of centres using the software is small. However, the scalability of this approach must be questioned. Ideally integration services such as CMEMS INS-TAC should happen in a more distributed manner, using a total data quality approach from the point of capture such as that provided by two-level modelling approach developed as part of this work.

The EMODnet-physics hosted platform WARP CEFAS- 62010720 has undergone the CMEMS INS-TAC integration process. At the platform's dashboard, SOAP API, GEOSERVER OGC, THREDDS and ERDAP services are provided. Also, a sensorML descriptor is provided. The OGC and SensorML descriptors are provided at a minimum requirement level for compliance. SensorML provides a mechanism to further describe the sensing process that is used to obtain observations, such as sensor calibration data. However, this level of detail is not available for this platform. WMS and WFS (see chapter 2) minimum compliance are provided. Within the Copernicus hosted platform page, Sensor Observation Services are not yet available and full O&M compliance is not observed. For example, the feature-of-interest was found to be encoded in an *non-O&M-compliant* manner. Two other ocean observing platforms (listed below) were also examined using the data flows obtained from the EMODnet-physics downstream service.

- EMODnet-physics hosted TWEms BSH – 10004 platform.
- EMODnet-physics hosted FoxtrottLightship Met Office – 62170 platform.

These additional datasets also followed the oceanotron metadata standard, and contained similar deficiencies.

6.3.3 Archetype Modelling & Concept Mapping

Having reviewed the application domain, an archetype modelling exercise was performed following the methodology described in section 6.1. As in the previous scenario (section 6.2) this archetype modelling exercise has been contributed to by way of the workshop mentioned in section 6.1.1.5, with the same limitations (described in section 6.2.4 and

chapter 7) to be noted. Also, as part of progressing from the domain review to initial archetype modelling phases, the author engaged with an ocean observing domain expert within the Marine Institute Ireland. This engagement took place over two separate onsite visit days and the purpose of the engagement was to review the overall approach being undertaken and to validate the initial selection of archetypes and appropriateness of the scenario for an archetype modelling approach. Leading on from these discussions, several archetypes were identified in the design phase and these identified archetypes were further refined and developed within the context of the ocean observing application domain and current ways of working by the author and informed by discussions with the Marine Institute domain expert.

One of the key considerations identified through domain expert engagements was the need to consider current real-world documentation and container systems used for ocean observing datasets as part of the modelling process. This is especially important here, as this evaluation uses real datasets.

NetCDF is the primary standard for packaging environmental datasets (see chapter 2). NetCDF was thus examined in the context of the next phase within the archetype modelling methodology (phase 3, archetype development, see section 6.1).

For the platform-based observations under investigation, the netCDF data model essentially acts as an *organizer* (see chapter 2, section 3.5), it does not represent a documentation model or a conceptual data model. However, as discussed in chapter 3, archetypes and two-level modelling provide a way to model and organize documentation about topics of interest in a standardized way.

As discussed previously in chapter 3, in two level modelling, *compositions* represent storage concepts; *sections* represent organization concepts; and an *entry* represents content concepts (see section 3.5). COMPOSITION, SECTION, and ENTRY were

shown previously, highlighted in the augmented O&M model in Figure 4.5 (chapter 4). It was also noted previously that identity and topic-of-information must also be modelled, this is also shown in Figure 4.5 and is considered in some more detail here.

For this work, after careful examination, it was found that the concept *region* can serve as the (basic) identity-model (see section 4.4.3 for discussion of the basic flexible identity model provided as part of the augmented O&M model). *Sea region* is a sub theme of region and *OceanRegion* within CMEMS and INS TAC. The CF standard-name for the region under investigation is used - north_sea -, meaning the *north_sea OceanRegion* is the topic of information for this study (see Listing 6.1).

A COMPOSITION concept can be described here as a transaction and a unit of committal (or a contextually complete and standalone "document"). Within the reference model (Figure 4.5) GeoData_Composition represents the stable composition concept from which further concepts can be defined using archetypes, as shown in the previous air quality monitoring evaluation.

```
archetype (adl version 1.4)
  TPOT-OM-Geo_Data_Document.north_sea.v1
concept
  [at0000]
Language original_language = <[ISO_639-1::en]>
Description original_author = < lifecycle_state = <"Draft">
        details = <["en"] = <language = <[ISO_639-1::en]>>
    >
definition
  Geo_Data_Document[at0000] occurrences matches {1..1} matches {  -- north_sea
  archetype_id existence matches {0..1} matches {*}
  details existence matches {1..1} matches { ......}
  geoDataComposition existence matches {0..1} cardinality matches {0..*; unordered; unique}
  matches {
    GeoData_COMPOSITION[at0001] occurrences matches {0..*} matches {  -- Slot
       observation_Set_ existence matches {1..1} cardinality matches {1..*; unordered;
       unique}
      matches {
        OBSERVATION[at0002] occurrences matches {0..*} matches {  -- Slot
          featureofinterest existence matches {1..1} matches {..}
          obsproperty existence matches {1..1} matches {
            ObservedProperty[at0006] occurrences matches {1..1} matches {*} --Slot
              details existence matches {1..1} matches {
           DETAILS_COMPOUND [at0008] occurrences matches {*} -- Slot
         }
          }
          resultTime existence matches {1..1} cardinality matches {...}
          results_cluster existence matches {1..1} cardinality matches {1..*; unordered;
          unique} matches {
            Results[at0009] occurrences matches {1..*} matches {*}  -- Slot
          }
             procedure existence matches {1..1} matches {*}
        } } } } }
ontology
  term_definitions = <
    ["en"] = <
      items = < ....
      ["at0001"] = < . . . < . . solved to {TPOT-OM-GeoData_COMPOSITION.platform-
            oceanSITES-moorings.v1}">>
      ["at0002"] = < . . . < . . solved to {TPOT-OM-OBSERVATION.PSAL_Obs.v1}">>
      ["at0006"] = < . . . < . . solved to {TPOT-OM-ObservedProperty.PSAL.v1}">>
      ["at0008"] = < . . . < . . solved to {TPOT-OM-
            DETAILS_COMPOUND.ComplexProperties.v1}">>
      ["at0009"] = < . . . < . . solved to {TPOT-OM-Results.PointTimeSeries.v1}">>
>  >  >
```

Listing 6.1 ADL Snippet of an archetype for the north_sea. The north_sea archetype is constructed using many other archetypes, a number are shown here in the summarized ADL file. Where concepts are described as external archetypes these are labelled as – Slot. Slots are bound to external archetypes using at-codes. For example, above it can be seen that the details attribute at0008 is in fact governed by the complex properties archetype.

As observing platforms may have short deployment times and therefore may only exist temporarily, for this work an observing platform deployment is considered a unit of committal. Its purpose in this evaluation is to capture a passing ocean observing event or a longer-term observing deployment.

Thus the following archetype is defined, which is a specialisation of the concept platform and OceanSITES (platform): *TPOT-OM-GeoData_COMPOSITION.platform-oceanSITES-moorings.v1*, shown as Archetype B in Figure 6.9 below.

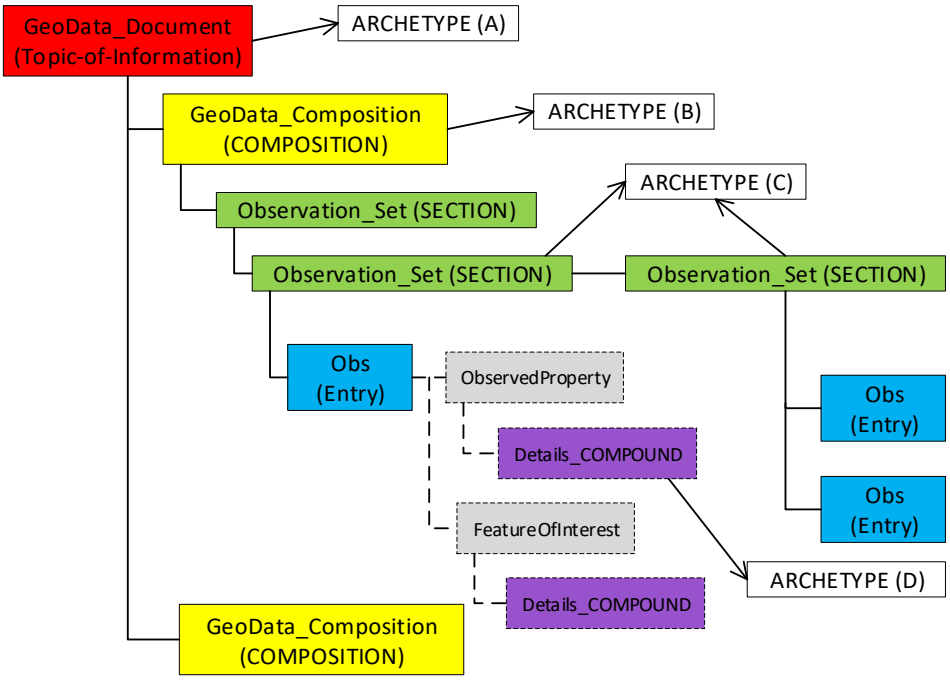


Figure 6.9 Archetype Definition Extent. Shown is the extent to which each archetype defines the overall model. GeoData_Document represents the top-level document, which contains an aggregation of compositions. Compositions are storage level concepts, in this case the document about the north_sea has numerous observing platforms which are COMPOSITIONS and governed by Archetype B. Archetype C is defined based on part of the OceanSITES netCDF model where observations are organized daily. Archetype D represents the INSPIRE defined complex properties profile of O&M, which has been further specialized.

A SECTION represents an organization concept. Within the reference model Observation_Set represents a stable section concept. The purpose of a netCDF file (see section 2.3) is somewhat analogous to a section. Here a section is an ordered list of content items, this is also true of netCDF files, however netCDF files contain much more information besides. In fact, much of the additional metadata within a netCDF file is repeated per netCDF file.

Sections may contain more sections or entries. For this study the netCDF *variables.attributes* concept is chosen as a constraint on the Observation_Set reference model concept. For convenience, the archetype name netCDF-attr is used. Therefore, the

following archetype is defined: *TPOT-OM-OBSERVATION_SET.netCDF-netCDFattrdaily.v1* (Shown as archetype C in Figure 6.9).

An ENTRY represents details of data elements. Within the reference model (Figure 4.5) OM_Observation represents a stable ENTRY concept. Here the practical salinity concept is mapped to an OceanSITES/INSPIRE/O&M compliant data model using the following archetypes:

- `TPOT-OM-OBSERVATION.PSAL_Obs.v1`
- `TPOT-OM-ObservedProperty.PSAL.v1`
- `TPOT-OM-OM_Observation.oceansitesObs.pointtimeseries.v1`

It can be seen in the O&M based reference model (chapter 4, Figure 4.5) ObservedProperty contains a COMPOUND type attribute called details. Details_COMPOUND allows for the further constraining and specialization of observed properties. As mentioned previously in chapter 2, INSPIRE already defines an O&M extension called the complex properties model (see section 2.5.2). Here the complex properties model is redefined as an archetype *TPOT-OM-Details_COMPOUND.complex_properties.v1* (Shown as Archetype D in Figure 6.9). Redefining the complex properties model as an archetype allows for further managed specialization and helps address the issue - described in Leadbetter and Volden (2016) (see section 2.5.2)- of the complex properties model being overly abstract.

6.3.3.1 Archetype Domain Expert Review

Having developed a reasonable number of new archetypes (described above) appropriate to the given scenario, a basic qualitative review of the archetype modelling outputs was performed. The purpose of the review was to gain further input to the initial archetypes under development and to ensure the given archetypes for this scenario represent an

appropriate maximal dataset needed for this evaluation. Reviewing archetypes is a normal stage of any archetyping exercise and is a pragmatic way to achieve higher quality archetypes by way of consensus (Min et al., 2018).

This review involved a one-to-one review session with an additional ocean observing domain expert. The domain expert was part of an ongoing state-of-the-art marine monitoring development project team, consisting of both academic and industry stakeholders. The review was conducted in the context of the scenario presented, but also in terms of an advanced water monitoring system and decision support system (which was attempting to adhere to INSPIRE compliance) that is currently under development. Two aspects of the archetypes under review were examined: *domain concepts* and *information representation*. The domain expert participated in the review by way of a guided video call performed by the author.

The domain expert was not familiar with the method prior to the review and so a high-level overview was provided to the domain expert including reading material. The review session was performed over two separate sessions to give the participant time to reflect on the approach and the initial review session. Following the review, several additional concepts and constraints were identified which would be required in the context of the domain expert's work on an advanced water monitoring system, but these could be accommodated through specialisation of the archetypes presented. In general, it was reported that the domain concepts were valid from the domain expert's perspective on the given scenario, and the information representation provided adequate coverage for the given scenario with enough flexibility for future scenarios (through specialisation). It was commented that the approach could potentially solve several ongoing issues experienced by the domain expert in an ongoing water quality monitoring development project.

Next, the supporting technical system and deployment are described.

6.3.4 Evaluation System Deployment

A proof-of-concept architecture and deployment environment was described previously in chapter 5 and shown in Figure 5.1. Here, to support this evaluation, the deployment environment and proof-of-concept system remains much the same (i.e. backend system hosted on DigitalOcean based droplets), however the deployment of the observing platform differs to the work described in chapter 5.

In chapter 5 the FIT IoT Lab infrastructure was used for evaluation purposes, with ARM M3, Contiki-NG based nodes deployed as observing platforms. Here the observing platforms are realised using three ARM A8 based boards (as recommended following the findings from chapter 5), described in more detail below.

Before the observing platform deployment is described, the further refinement of the digital mist platform arising from evaluation findings in chapter 5 are described below.

6.3.4.1 Knowledge Framework Implementation

The architecture and software components described in chapter 5 are again employed within this evaluation. Specifically, for this evaluation a basic Web application has also been developed to provide a visual interface and a front-end view of the experiment (Figure 6.10).

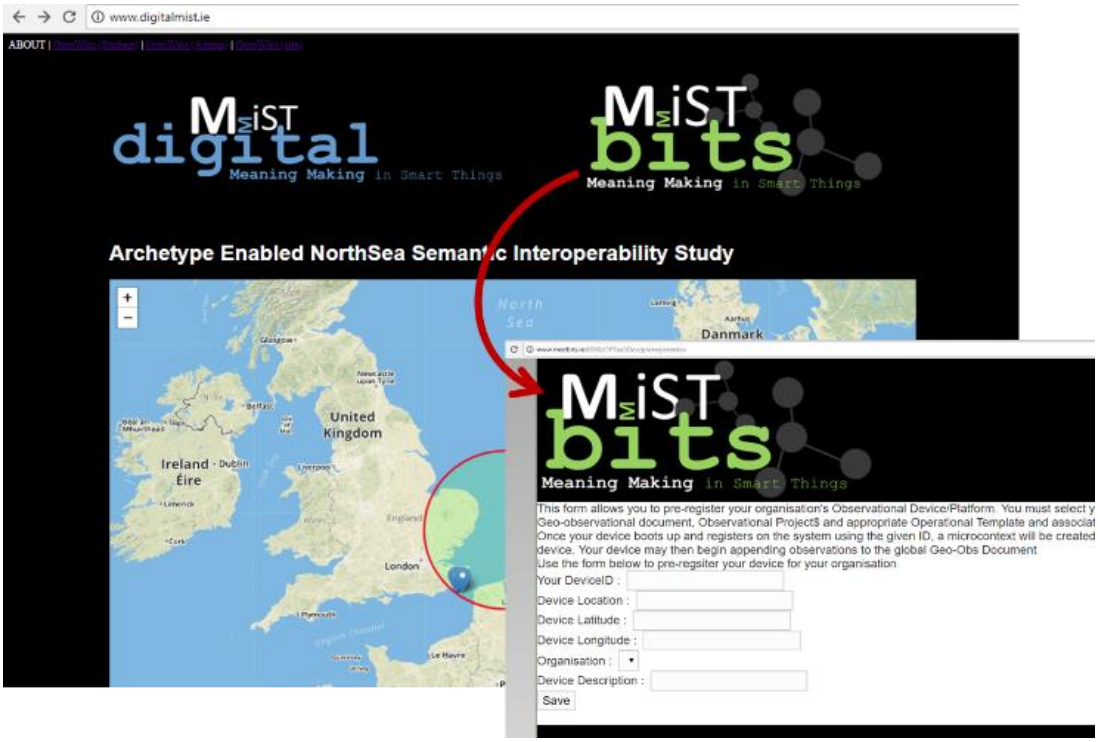


Figure 6.10 The OPTaaS backend infrastructure is implemented as a set of RESTful Web services using Groovy/Grails and Java. New platforms can register against community agreed archetypes/opts where the platform then receives a micro-context template to constrain their observational data.

6.3.5 Evaluation Overview & Analysis

To ensure a robust frame of reference for this evaluation, real marine observational datasets and SDIs were considered. A review of publicly available ocean observational portals was performed, following on from SDIs detailed in the literature review (see chapter 2). Of the portals reviewed the EMODnet-Physics data portal (Novellino, 2015) was chosen to support this evaluation and the subsequent comparative analysis. EMODnet-Physics was chosen as it represents the state-of-the are in ocean monitoring SDIs (see chapter 2).

Three ocean observing platforms were selected within the area of the southern North Sea. This sea area is chosen as it is composed of several bordering jurisdictions (UK, Netherlands, Belgium, France) who are subject to EU INSPIRE compliance (INSPIRE, 2007) (see section 2.5.1). This approach aligns with the scenario description presented in

section 6.3.1 above, while using the INSPIRE directive also provides a useful lens to compare the current state-of-the-art deployments with the potential benefits of adopting the two-level modelling approach developed as part of this work.

To perform the evaluation, observational data for a 60-day period was downloaded from each of three ocean monitoring platforms through the EMODnet-Physics data portal. The data was retrieved in netCDF format (see section 2.5.2). NetCDF data files were converted to JSON using the netCDF operator tool suite NCO toolkit (Zender et al., 2012) for ease of parsing and assessment. The assessment of the retrieved datasets examined adherence to common standards and interoperability traits using mapping tables of data concepts and their representation contained within the netCDF files.

The mapping tables were then used to perform a transformation of the datasets to produce harmonised, INSPIRE (and O&M) compliant data flows (see Table 6.4 below, column 2).

To further validate and then analyse the overall two-level modelling translation approach developed within this work, the archetypes listed in section 6.3.3 were combined to create an operational template (.opt file). Further constraining and transformation of the now INSPIRE compliant datasets using notional *community* agreed archetypes and the O&M profile was performed (see Table 6.4 below, column 3).

When an observing platform is ready to come online and begin reporting observations, the platform is pre-registered on the OPTaaS backend system using the *Mistbits* registration form shown in Figure 6.10, relevant templates for the platform were associated with each platform. A pre-registration ID is returned. This pre-registration ID was then used by the platform to register fully on the backend system when the platform is fully setup.

During the evaluation, platforms register by calling the following URL and passing their unique pre-registrationID: *http://mistbits.ie:8080/OPTaaSDev/register/{pre-red-ID}*. The OPTaaS backend system then builds a constrained micro context which acts as a micro template for the platform to create information instances. For example, a snippet of the micro context for the WARP CEFAS- 62010720 is shown in Listing 6.2 below.

```
"@Context" : {
  "obj_store" : " coap://[2a03:b0c0:1:d0::c61:1]/obj_store/",
  "obj_id" : {
      "@id" : "obj_store: 6b73517a-0efa-11eb-adc1-0242ac120002", //
     "@type" : "@id"
  }
  "at0000" : "obj_id:at000/",
  "at00001" : "at0000:at0001/",
  "at0002" : {
         "@id" : "at0001:at0002",
         "@type" : "@id"
  },
  "DV" : {
         "@id" : "at0002:#at0006",
         "@type" : "@id"
  },
  "resultTime" : {
         "@id" : "at0002:#at007",
         "@type" : "@id"
  }
```

Listing 6.2. Micro-context returned from the OPTaaS backend once the platform WARP CEFAS-62010720 has registered. The object has a UUID of which is the TPOT-OM-OBSERVATION.PSAL_Obs.v1/[at000]/[at0001][at001]; which is a PointTimeSeries data object governed by the archetype TPOT-OM-Results.PointTimeSeries.v1 for the practical salinity measurement

When the observational platforms report new observations, they use the OPTaaS observations append web service. Platforms call the URL below, using a POST method and passing the observations in the format defined in the platform's micro context template. `coap://[2a03:b0c0:1:d0::c61:1]/obs-append/{platformID}`.

The observation append Web service appends the new observations as a new SECTION with associated entries for the COMPOSITION relating to the reporting platform. The act of appending observations involves a validation step to ensure the information instance adheres to the platforms associated operational template. It is important to note that appending observations adds information to the overall document about the topic-of-interest. In this case the *north_sea*.

Prior to running the evaluation simulation, each dataset was loaded onto the external flash memory of three separate ARM 1GHz Cortex A8 processor-based boards with wired LAN connectivity (Figure 6.11). Each board represents each dataset's source observing platform. Experimental time spin-up was of the order of 60:1, meaning the 60-day period of data was re-run over a 24 hour period. The data was reported using the operational-templates-as-a-service (OPTaaS) and Linked Data knowledge graph method described in chapters 4 and 5 (and above). Data assimilation was performed using the OpenDA toolbox (discussed above), with experimental real-time assimilation of the reporting test rig system performed to *tune* the GAM estimation model (equation 1) parameters as new datasets were discovered.

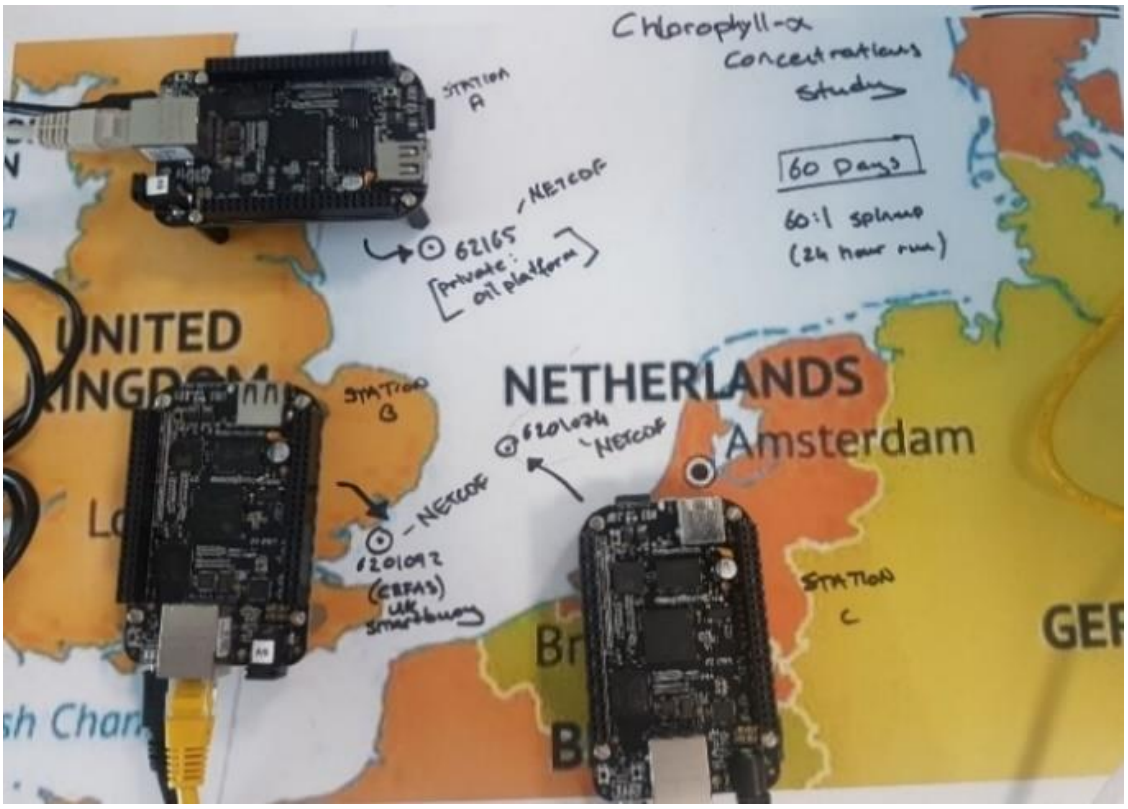


Figure 6.11 Test rig. Each board represents a real deployed platform. Data for each platform was acquired from the EMODnet-physics portal.

Because of the two-level modelling approach used, AQL (see section 5.1.3) could ultimately be used here to perform a fined grained automatic assessment of newly

discovered data-flows relevant to an application. This is enabled by the rich metadata associated with each information object, standardized to meet the community agreed constraints. The testing framework does not support AQL yet. However, an example AQL statement using the developed archetypes is shown for illustration below in listing 6.3 (this will be the focus of future work).

```
SELECT c/…/wmo_platform_code
FROM GDR [include specific scoping here]contains
 GeoData_COMPOSITION c [TPOT-OM-GeoData_COMPOSITION.platform.v1
contains OM-Observation_Set […]
contains OM_Observation obs [TPOT-OM-OM_Observation.PSAL_obs.v1]
WHERE  obs/data[at0001]/details_COMPOUND[at0002]..
 /items[at004]/value = "hourly"
```

Listing 6.3 AQL example statement

As the OPTaaS backend system uses a linked data approach to build information instances, enabled by Apache Jena (see chapter 3), SPARQL end points are provided by Fuseki (see section 3.2). Fueski allows the data to be queried using a semantic search approach (i.e. using SPARQL). In place of AQL, SPARQL was adopted during this evaluation to demonstrate the automatic discovery of relevant observing platforms against their rich metadata provided by the two-level modelling approach. A SPARQL query example is shown in Listing 6.4 below. Note in the example below the archetype appears as an OWL schema (see section 3.2.1), converted from its original ADL representation.

```
PREFIX sea_region: <http://digitalmist.ie/optaasdev/archetypes/tPOT-
REGION.sea_region.v1.owl#>
SELECT DISTINCT ?sea_region WHERE {
?sea_region sea_region:at0000.1_..... "north_sea"
} ORDER BY ?sea_region
```

Listing 6.4 Archetype based SPARQL query. Here the simple SPARQL query will return all platform wmo codes where platforms are located within an area of interest, governed by their longitude and latitude coordinates.

As new platforms come online and are discovered a simple *quality reasoner* decides whether to integrate the new data flow. In this evaluation, as each platform becomes live,

it is discovered using the SPARQL query in Listing 6.4. The dataflow is assessed for relevance to the application using fine-grained standardised search terms against the platforms governed archetypes. For this evaluation, a quality reasoner has not been developed (yet) and the system is configured to accept a dataflow, assimilate it and continually produce chlorophyll-a estimates using the combined Kalman filter and GAM.

Note, that the purpose of this evaluation is to solely assess the improvement in interoperability and findability of datasets, using the two-level modelling approach, which in turn should enable better model estimation. The assimilation process and GAM used do not provide accurate estimated datasets, and so an assessment of, for example, the ability of the approach to reduce root mean square error (RMSE) using additional observations has not been performed, nor is it useful here to achieve the work's objectives. The aim of this evaluation is to determine whether the overall two-level modelling approach described in chapter 4 meets the research objectives (see section 1.5). The evaluation scenario described here seeks to enable a validation and evaluation of the approach at a highly domain specific application level, and further show the ultimate alignment of the approach to meet the objectives of this work.

Therefore, the accuracy of the estimated Chla values are not in themselves important at this point. However, it is worth noting that, as shown in Figure 6.12, reasonable Chla values were produced when compared to the satellite based Chla readings obtained for the same time period (16^{th} of August 2016). This is shown in the plot at the top of Figure 6.12. As can be seen in Figure 6.12, the estimated value of Chla using the GAM model at the time period highlighted in orange is 2.54 ug/l of chlorophyll concentration and the satellite reading for the same period is ~ 2.4 ug/l (or mg m^3).

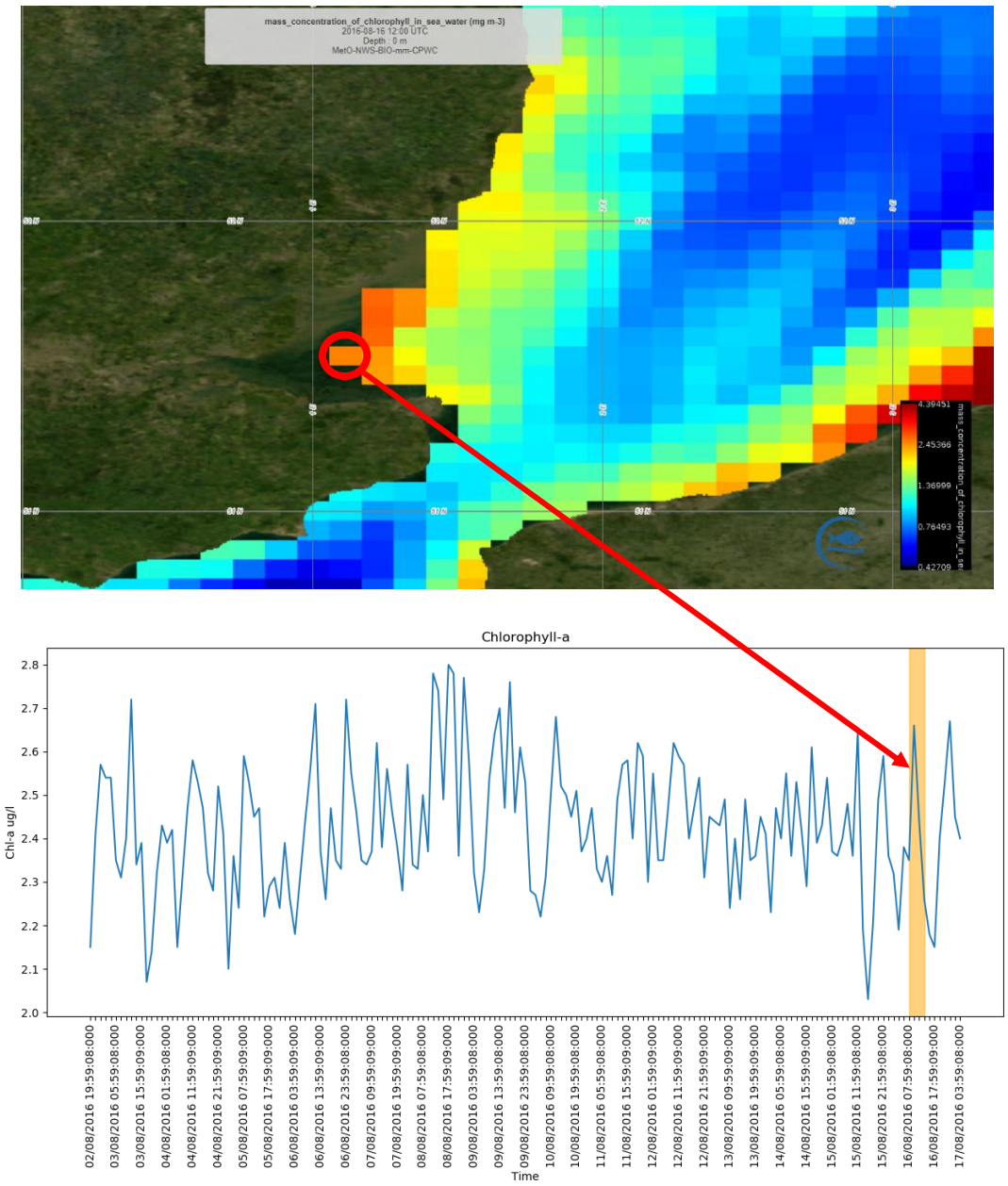


Figure 6.12 Chlorophyll-a prediction over time resulting from the GAM model and assimilation of ocean observations from the three observing platforms. Shown is the output of the OpenDA simulation for 1 singular point (lon = 1.11E; lat=51.52N). On the Y-axis are chlorophyll-a concentrations, measured in ug/l. On the X-axis is time in units of days-hours. The first 16 days are shown from the 60-day observing time period. The predicted Chl-a values highlighted in orange on the bottom graph correspond to the values observed by satellite on the same date shown in the map on top, highlighted by the red circle. The top plot was produced using the Copernicus data portal resources tool. The bottom plot was produced using Python libraries Numpy and Matplotlib.

The important point to note is that the estimated values produced are now documented and recorded in a semantically interoperable way and the approach has provided a richer mechanism to integrate, use and disseminate the outcomes of observing, see section 6.3 and Pearlman et al. (2019). This means that the values can be interpreted properly at a

later point (discussed later). In many cases, when datasets are used for secondary use, such as the production of new estimated derived values of Chla, the resultant datasets are not documented with sufficient context. The two-level modelling approach described in this thesis provides a solution to carefully documenting the evolution of the data, as it's used and reused throughout the data value chain (see section 1.2, problem statement, and section 1.5 research objectives).

Using the two-level modelling approach, the GAM parameters used to generate the Chla estimation values can now be documented in an interoperable and machine-readable way by the data provider (in this case the author). Often this level of documentation is provided in spurious, non-standardised reference manuals (PDF files) or not at all. This could mean that the inaccurate Chla values produced here may be misinterpreted, leading to incorrect conclusions and conflation (see section 1.2, problem statement), one of the core arguments for *born semantic* data (see section 3.2). For example, the additional documentation of the *procedure* used to arrive at the Chla values can be captured using archetypes. The archetypes listed below are defined to further specialise the O&M based procedure concept which constrains the OM_PROCESS concept from the augmented O&M model (Figure 4.5).

- `TPOT-OM-OM_PROCESS.procedure.Sensor.v1`
- `TPOT-OM-OM_PROCESS.procedure.SimpleProcess.v1`

Within the SWE (see section 2.2), procedure is normally encoded using sensorML (see section 2.4.1). The further constraining of procedure can follow both sensorML and further extend it to capture the GAM parameters (see Appendix C) in an interoperable way.

A sample of the WARP CEFAS-62010720 platform's dataset, standardised through the EMODnet-Physics portal is provided in Appendix C. The observational dataset has

been standardised using the Oceanotron software and includes QC indicators and standardised CF naming conventions. However, the dataset is not INSPIRE compliant, and the data part of the dataset is not O&M compliant. Using the two-level modelling approach described in this work, a set of archetypes was developed to provide fine-grained control over how the dataset is described. This has enabled a comparative analysis of the impact of two-level modelling on the datasets used within this evaluation. The INSPIRE directive and implementing rules have been used as a lens to analyse the dataset's transformations (discussed in the next section).

6.3.5.1 Data Transformation Comparative Analysis

To map datasets retrieved from the EMODnet-Physics data portal to be INSPIRE compliant, the first task was to map between *sub themes*. The INSPIRE application *Find Your Scope*[93] was useful to aid navigation of the large array of specifications that are defined under INSPIRE. EMODnet-Physics is organised under *topics*, whereas INSPIRE is organised into clusters (nine thematic, and two cross-domain). The INSPIRE thematic cluster on *Metocean* works closely with EMODnet-Physics to align both community practices. However, as noted previously in chapter 2, this is still a work in progress. It was found that the ocean observing datasets obtained from the EMODnet-Physics portal (in the specified 60-day period in 2016), did not align with the INSPIRE sub themes of oceanographic features or otherwise (i.e. MC/MF, see Table 6.4 below).

Transforming datasets to be INSPIRE compliant was seen to improve the interoperability of the datasets. For example, in Table 6.4 below it can be seen that the O&M standardised term *featureofinterest* is now assigned the value http://vocab.nerc.ac.uk/collection/C16/current/04/ in colmn 2, which links to the standardised definition of *North Sea*. This enables better "findability" (see Table 6.4

[93] https://inspire-regadmin.jrc.ec.europa.eu/dataspecification/FindYourScope.action

below) by providing syntactic interoperability of attributes, as they are named against the O&M standard. It also improved the semantic interoperability of the dataset as vocabulary servers such as NERC also provide concept relationships. For example, in the case of the dataset shown in Table 6.4, North Sea is *same as* http://vocab.nerc.ac.uk/collection/C19/current/1_2/ which provides a richer definition of the feature of interest (north sea), including *narrower*, *broader* and *related* terms.

The third column in Table 6.4, "two-level modelling approach", shows a snippet of a data instance transformed from the original EMODnet-physics based datasets using the two level modelling approach, and subsequently reported.

In Table 6.4 it can be observed that the data instance in column 3 looks very different to both the EMODnet and the INSPIRE compliant data instances. Many of the human readable terms present in column 1 and 2 do not appear within the data instance in column 3. For example, last_latitude_observation appears as [at0.2] in column 3.

Although, this reduces the column 3 (Table 6.4) data's immediate *human* readability, it does increase the data's *machine* readability. The reason for this is intentional, and is related to how concepts are organised between levels within the two-level model approach. Let us consider again the term *last_latitude_observation*, which can be said to be a volatile concept, and so has been defined within the knowledge level (or 2nd level) i.e. within the archetype:

- `TPOT-OM-DETAILS_COMPOSITE.shape.Point.v1.`

The data object types which appear within the data instance (column 3) can only be constructed from the reference model and are therefore stable concepts (see chapter 3 and chapter 5). This means that information systems, SDIs and data portals that are developed against the two-level modelling approach remain relevant in the face of change at the knowledge level. Systems are prevented from straying from the core stable reference

Table 6.4 Instance Data Transformation Table. Equivalent Archetype governed information structuration across information instances is highlighted (1)-(4). Example equivalent data points across data instances are highlighted using coloured highlighting.

(Col. 1) EMODNet-physics (1)

```
"variables": {
... //data removed
(4) "PSAL": {
      ... //data removed
      "type": "int",
      "attributes": {
        "long_name": "Practical salinity",
        "standard_name":
          "sea_water_practical_salinity",
        "units": "0.001",
        "_FillValue": -2147483647,
        "valid_min": 1,
        "valid_max": 36500,
        "DM_indicator": "R",
        "scale_factor": 0.001,
        "add_offset": 0
      },"data" :[34.225, .. . //data removed
   },
...//data removed
},
"attributes": {
    "platform_code": "6201072",
    "wmo_platform_code": "6201072",
    "source": "mooring",
...//data removed
    "update_interval": "hourly",
    "qc_manual": "OceanSITES User's Manual
                                    v1.2",
    "last_date_observation": "2016-08-
                              17T03:59:08Z",
    "last_latitude_observation": "51.5255",
    "last_longitude_observation": "1.028"   (2)
  }
}
```

(Col. 2) INSPIRE

```
... //data removed
(3) "_omso:PointObservation": {
    "_om:phenomenonTime": {
        "_gml:TimeInstant": {
           "_gml:timePosition": {
             "value": "2011-01-26T19:25:00"
           },
           "_gml:id": "..",
        },
    },"_om:resultTime": {
        "_gml:TimeInstant": {
           "_gml:id": "..",},
    },
    "_om:procedure": {
        "href": "http://vocab.nerc.ac.uk/.. ",
    },
(4) "_om:observedProperty": {
        "href":"http://vocab.nerc.ac.uk/..",
    },
    "_om:featureOfInterest": {
        "href" :
            "http://vocab.nerc.ac.uk/collection
                        /C16/current/04/"
         "_sams:SF_SpatialSamplingFeature": {
             "_gml:id": ".."
                "_gml:shape": {
                   "_gml:Point": {   (2)
                      "_gml:pos": {
                         "value":[[51.5255],
                                  [1.028]]},},
    },"_gml:id": "..",},
    },
    "_om:result": {
        "value": 34.225 . . .},
},

... //data removed
```

(Col. 3)Two-level modelling approach

```
"OBSERVATION_SET":[{
  "archetype_node_Id":"TPOT-OM-
      OBSERVATION_SET.netCDF-oceanSITES.v1"   (1)
...//data removed
  "DETAILS_COMPOSITE":[{ . . .
    "DETAILS_COMPOSITE":[{
      "archetype_node_Id":"[at0.4]"
      "details_ELEMENT":{
         "archetype_node_Id":"[at0.17]"
          "DETAILS_VALUE" : "6201072"
      },"details_ELEMENT":{
          "archetype_node_Id":"[at0.9]"
          "DETAILS_VALUE": "mooring"
      }],}]
...//data removed
"DETAILS_COMPOSITE":[{
   "archetype_node_Id":"TPOT-OM-
       DETAILS_COMPOSITE.shape.Point.v1"   (2)
   "DETAILS_COMPOSITE":[{
      "archetype_node_Id":"[at00001]"
      "details_ELEMENT":{
         "archetype_node_Id":"[at0.2]"
          "DETAILS_VALUE" : 51.5255
      },"details_ELEMENT":{
          "archetype_node_Id":"[at0.4]"
          "DETAILS_VALUE": 1.028}],}]
...//data removed                          (3)
"Observation":{
   "archetype_node_Id":"TPOT-OM-OBSERVATION.
                  PointObservation.v1",
   "observedProperty":{
      "archetype_node_Id":" TPOT-OM.
                  ObservedProperty.PSAL.v1",   (4)
      "details_COMPOUND":[{ ...
   "featureOfInterest":{ . . .
   "results": [ . . . { "result":[{
            "archetype_node_Id":"[at0008]"
             "DATA VALUE" : "34.225"
      ...//data removed
```

model and so remain interoperable with other systems adhering to the two-level model-based reference model. This is a key advantage of two-level modelling over current standardisation approaches, as it allows systems to remain future proof in the face of evolving standards (see chapter 3). The 2nd level provides a dynamic mechanism to allow the volatile concepts (in this case last_latitude_observation) to evolve, using archetypes. This is not the case with current approaches within EMODnet or INSPIRE.

As mentioned above, it now appears that the two-level modelling-based data instance (column 3, Table 6.4) appears to have become less human readable. However, this is not really the case. Using the *archetype_node_id* values (see Table 6.4, column 3), a rich human and machine-readable descriptor (ADL based archetypes) is now easily accessible instead to give context and meaning to the data (see archetype Listing C.3, Appendix C).

Using the ADL encoded archetype, semantic interoperability is further improved beyond what is possible with either EMODnet or INSPIRE. As the archetype in Appendix C shows, fine-grained definitions of the data instances and meaning, beyond that of INSPIRE are contained within the archetype. The archetype also provides richer semantics through its term binding mechanism and through linking to both external and local ontological definitions. Strong data typing is now also observed within the data instance against the governing archetypes, a key requirement to ensure interoperable datasets, and an aspect of the INSPIRE directive that remains a work in progress.

Through this evaluation, it was also found that the use-case scenario modelling process was much improved using two-level modelling, compared to that possible using the INSPIRE portal. However, the modelling process was hampered within this evaluation by the lack of agreed community archetypes available. This was in comparison to the rich schemas available within the INSPIRE Web portal. The improvement in modelling arose when scenario specific constraints were required to be encoded that were not already

captured within the INSPIRE based schemas. Where domain and use-case specific encoding went beyond that defined within INSPIRE, there was no clear path to ensure an integrated definition and extension to the schema that could be community agreed in a short space of time. This is contrast with the two-level modelling approach. Using LinkEHR, an accessible (from a domain practitioners' perspective) and controlled extension capability was possible. This extensibility mechanism (key to any Digital Earth system, see chapter 1) allows for use-case specific standards encoding, while enforcing semantics and interoperability during the modelling process, as only concepts contained within the reference model are allowed (see appendix C, Figures C.1 and C.2).

The evaluation highlighted the limitation of only having the augmented O&M model as part of the underlying reference model. Certain concepts could not be mapped to the reference model but were available within INSPIRE schemas. This was to be expected, as the evaluation is based on only a proof-of-concept implementation of the approach. Upon further development and adoption of the approach, a richer and broader reference model would need to be defined. It was found during this evaluation that rich schemas already available within INSPIRE can provide the basis for this two-level modelling reference model work.

Within EMODnet-Physics, the ability to capture extensions or use-case specific encoding was even less well managed than that of INSPIRE. Use case and domain specific concept details tended to be captured in non-standardised PDF documents, hosted on data providers websites (such as CEFAS, see section 6.3.2 above) and linked to the dataset using the OceanSITES standardised *references* and *institute_references* attributes.

Table 6.4 above also highlights the disparity in data instance size possible needed for the sub object *results* to *make sense*. Allowing platforms report only the minimum data required per observation was a key objective of the linked data approach employed here

(see chapter 4 and 5). Using the federated linked data approach, the results JSON object can be extracted, while remaining linked to a rich set of machine-readable documentation (using micro-contexts, data graph and governing archetypes) to ensure the data fragment retains its semantic meaning (see Listing 6.2 above). This mechanism enabled the observing platform boards, used within this evaluation, to become producers of *born semantic* data objects (see section 3.2) (using the OPTaaS and two-level modelling backend infrastructure, described within chapter 4). This contrasts with how data objects are created by data providers within the EMODnet-Physics ecosystem, where standardisation may happen later (see chapter 3), within the regional processing centres and not at the point of capture.

Table 6.4 also shows how standardisation efforts can be merged and aligned using the two-level modelling approach once the reference model is suitably designed. It can be seen in the two-level modelling data instance (column 3, Table 6.4) that both the EMODnet-Physics based OceanSITES and INSPIRE standards requirements can be satisfied simultaneously using the two-level modelling approach (highlighted in column 3, Table 6.4 (1)-(4)). Data brokers, or converters (such as an SOS broker, Figure 5.7, chapter 5) can be subsequently integrated to two-level modelling-based systems to produce specific encodings of datastreams on request, and thus retaining the requirements of any existing standards requirements within the application domain.

It should also be noted that once platform's observational datasets were consumed by the Chla estimation application and thus reused to produce new knowledge (i.e. Chla estimations), portals such as EMODnet-physics do not provide any additional way to document how the estimated values of chlorophyll-a have been calculated, beyond re-submitting the estimated values to the INSTAC ingestion engine (see chapter 2). Therefore, only producing a similar dataset to that shown in Appendix C (see Warp (TH1)

dataset), while only referencing the estimation process by way of a link to a manual or some other form of non-machine-readable documentation using the following netCDF attributes: *references*, *qc_manual* or *distribution_statement*. Therefore, the *provenance* (see section 2.2.1) of the estimated values will not be documented in a standardised and machine-readable way.

Complex processing of raw observational data is common within the ocean observing domain (Blauw, 2015), but recording this additional information in a standardised way has not been realised yet, beyond what has been proposed as part of ODM2 (discussed in section 3.6). Whereas using the two-level modelling approach, the standardisation process can be continually extended within the community using the specialisation functionality provided by two-level modelling approaches. This specialisation can be seen in the archetype development example in Appendix C (see Figure C.3).

To further evaluate the resultant data outputs, the data transformation was reviewed with the same domain expert as per section 6.3.3.1. The review focused on the potential benefits of the approach to ongoing marine monitoring projects and identifying the limitations of applying the approach to ongoing real-world deployments.

During the domain expert review it was noted that the additional processing overhead may result in scalability issues and that current system users may not perceive the ultimate benefits of having such detailed information instances. It was commented that current industry needs may see the approach as "overkill" and not worth the investment, and more tangible demonstrations of the benefits of the approach to enabling end user applications would be necessary to gain industry buy-in. However, it was noted that the benefits of the approach were clear from a future trends perspective, and the approach shows potential for solving the perceived short comings in initiatives such as the INSPIRE directive.

This study has ultimately found that two-level modelling can be beneficial within the ocean observing domain to better manage and bridge *top down* and *bottom up* standardisation processes (see section 3.3.1) using a stable reference model and archetype constraint definitions. However, the ultimate benefits of adopting the approach versus the effort, increased complexity and cost are not evident enough to promote adoption within industry or real-world deployments and more tangible benefits need to be demonstrated in future studies.

6.4 Chapter Summary & Discussion

Section 6.1 defined an appropriate archetype modelling methodology to support a domain evaluation approach. In section 6.2, a domain based environmental observing evaluation of the two-level modelling approach, supported by the augmented O&M reference model described in chapter 3 was presented. The findings arising from section 6.2 further confirm the suitability of the approach described proposed throughout this thesis, its potential benefits, and how the approach is in keeping with current research agendas within the IoT and smart-city fields.

It was also shown how existing standardisation efforts can be supported with the two-level modelling process and this confirms the flexibility and wide application of the approach. Ultimately the aim of the approach is not to duplicate work already done or create a divergence in approaches, but to provide an alternative mechanism to further the already ongoing standardisation work within geo-spatial communities, while solving identified shortcomings with current approaches.

In lieu of an overly descriptive, and without making light of the current issue with the proliferation of new standards (and to provide the reader with some light relief at this point within the text), Figure 6.7 below captures the issue succinctly.

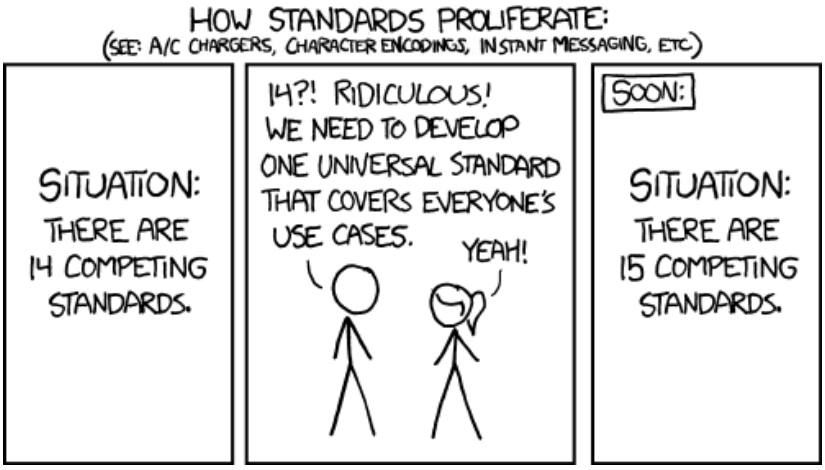


Figure 6.13 How Standards Proliferate. ***Image credit Tor Bjorn Minde, RI.SE lab, Sweeden*.**

In section 6.3 a second ocean observing domain evaluation approach was presented. While using real world data it was found that data pre-processing is an important step when assimilating data from heterogeneous sources. To ensure data sources are truly interoperable, the metadata must be detailed enough for systems to manually assess the dataset's suitability for automatic assimilation into the system. Also, adhering to principle of *collect once. use multiple times, and find-bind-publish*, data providers may wish to re-publish the cleansed dataset including data provenance in an interoperable way.

Retrieving data from current spatial data infrastructures can be a cumbersome process. Current SDI implementations do not allow for easy automatic discovery and consumption of ocean observational data flows. The reason for this is twofold:

1) The lack of a standardized data access mechanisms across different SDIs.
2) The lack of adherence to standard data & information models.

Initial results of the automatic data assimilation exercise results show that discovery and assimilation of data can be automated with a high degree of confidence when systems adhere to community generated archetype models.

Both evaluations have shown the approach to be flexible and robust in real-world scenarios. It has also shown that the approach is in keeping with current interoperability

efforts and is compatible with existing standards. Another advantage of the approach is that is improves the ability of systems to automatically discover relevant data flows and datasets and due to the verbosity of the quality data enables the automatic assimilation of the data into existing monitoring applications a key requirement for an y Digital Earth; discoverable data (section 1.1.2).

Chapter 7

"There is no end to the journey, and that is the mystery, the beauty of it"
(Krishnamurti, 1964)

7. DISCUSSION & CONCLUSION

Chapter Overview: This chapter provides the reader with a summation of the overall outcomes of the research presented in this thesis. The main research objectives from Chapter 1 are revisited and discussed in relation to the findings of this work. The main conclusions arising from this research are presented and discussed. The main contributions of this work are also described, and a future research agenda is presented.

With the evolution of geographical observational data capture, storage and sharing technologies such as in situ remote monitoring systems and spatial data infrastructures, the vision of a Digital Earth, as articulated by Al Gore in 1998 is getting ever closer. As discussed in chapter 3, current data interoperability efforts solve many problems within Earth system science-based information systems and spatial data infrastructures. However, despite the need for high quality "joined-up" information to document the climate crisis and associated global environmental issues, interoperability at a knowledge level to date has not been fully realised. In fact, many information infrastructures are still struggling to provide even the most basic syntactic interoperability, despite far reaching legislative directives such as INSPIRE.

There is a disparity in the pace of standards development and the ability of systems deployed in the field to keep pace with changes. This is partly due to the nature of top-down standards processes, whereas most real-world deployments use bottom up best

practice approaches during implementation and deployment, with a minimal adherence to overarching standard in place at development time. Once deployed, it is difficult and costly to update systems to take account of any modifications in published data standards.

As discussed in chapter 1, the process of capturing volatile domain specific knowledge concepts in an observational system, and in supporting information management infrastructures, invariably leads to a mismatch between the needs of the domain practitioner and the concept definitions. Knowledge within complex domains is always evolving, and standards development and standards-based systems struggle to evolve in tandem. This is at the root of the problem in standards adoption at the system and use case level. All these issues limit the ability of the Earth science community to meet the many global challenges arising from climate change (see chapter 1, section 1.1).

Two-level modelling has been shown to provide the basis for achieving adaptable interoperable knowledge-based systems within the health domain (see section 3.5). This work has shown that with additional design patterns, existing Earth system science-based data models such as the O&M standard can serve as the basis for a two-level modelling reference model, which can be translated beyond health to the geo-spatial domain (see chapter, 4,5 and 6).

Stable and general standardised reference models are a key requirement for a successful translation of this approach to the geo-spatial domain. While O&M only provides a minimal reference model for the Earth system science domain, the success of the O&M profiling work done here to support a two-level modelling approach acts as a proof-of-concept of the translation approach that was adopted in this work. Findings presented in chapter 5 (section 5.5) and chapter 6 (sections 6.2.4 and 6.4) from infrastructure deployment and domain evaluations provide additional evidence of the suitability of the approach and its potential benefits.

The translation approach presented in this thesis can enable a wide and diverse set of domain experts (within the Earth System Science community) to contribute directly to the creation and evolution of consensus-based content models, while ensuring the provision of high quality and accurate shareable knowledge amongst a diverse super-domain. These characteristics indicate that the solution is important in the realisation of a Digital Earth system. For example, the provision of Web based "social network" style management, review and publishing of Digital Earth Knowledge Artefacts in the form of archetypes, can foster greater semantic interoperability between systems.

The continuous process of model evolution contrasts with the relatively lengthy renewal cycle of geographical information-based ISO standards (O&M, WMS etc.). Standards development is typically a top-down process. Outside of standards, communities often also adopt a best practices approach. Best practices are typically a bottom-up approach and many such practices, in time, lead to the creation of standards. Best practices and standards are part of a wider process of community agreements (Pulsifier et al., 2019).

Archetypes and two-level modelling allow the bridging together of top-down and bottom-up processes; allowing implementation best practices to be documented and evolved on top of published standards. The two-level model approach promotes the idea that information can be structured and constrained using archetypes to enable its use in high quality "live" documentation. The experience with multiple national and international initiatives within the health domain has shown that archetype repositories allow best practices to be readily adopted within the community, improving quality of information systems, and increasing interoperability[94].

[94] https://ckm.openehr.org/ckm/

Templates and OPTs offer additional flexibility and specialisation outside of the community-agreed archetype model for local use while still adhering to the community constraints where possible. This provides for situations where disparate domain expert groups may disagree and can lead to archetype alignment issues as the approach matures within the domain. Again, as the development community continues to be richly supported, techniques to overcome this potential for divergence are emerging (Bisbal and Berry, 2009).

This work has shown that the two-level modelling experience should be of interest to Earth System Scientists. Especially those wishing to share interoperable information that is trusted across measurement platforms and sub-domains.

7.1 Objectives and Achievements

Having made the case for the application of two-level modelling to support interoperability and sharing of higher quality document-oriented information in the ESS domain, in this section, the main objectives of the research are reviewed against the work done in trying to meet them. Limitations are highlighted and future opportunities are identified. Firstly, a summary of the work performed to achieve the objectives is provided.

7.1.1 Objective 1

Identify the technical tasks required to translate the two-level modelling methodology from the health domain to the geo-spatial and Earth System Science domain

Using the design science paradigm (chapter 1, Figure 1) the main areas for review were identified, these are categorised within the *environment* and *knowledge base*. This informed a detailed review of the relevant literature and approaches to enable semantic interoperability within the Earth system science domain (chapters 2 and 3). This exploration of the relevant literature allowed a model and experimental methodology (section 1.6.2) to be employed to meet objective 1, where the technical tasks required to

translate the two-level modelling methodology were identified and refined to form a set of proposed theories to be further refined and evaluated.

The technical tasks needed were identified and they subsequently informed the model for how two-level modelling can be translated to the geo-spatial domain. The approach identified was then evaluated through several iterations of a design & build methodology within the context of the design science approach (see chapters 4 & 5). The results of the evaluation showed that the two-level modelling translation approach can be successfully deployed to the geo-spatial domain. The translation approach developed and described in chapter 4 was validated through a proof-of-concept build (described in chapter 5). Through a use-case evaluation approach (described in chapter 6) the translated modelling methodology was found to contain the required expressivity to be able to produce data flows for the defined scenarios of use. The main tasks were identified to enable translation were summarised in section 4.2 and subsequently validated and evaluated throughout chapters 4,5 and 6.

The work thus far has provided a proof-of-concept of the defined approach; however, it is still not comprehensive, and the findings are still not conclusive as to the full applicability of the approach defined, as there are areas still requiring further investigation. For example, the question of developing a generalised identity model for the ESS domain remains an open question.

During this work, it was not practical to form a rich community of supporters and thus accurate archetype models developed using a true consensus-based modelling process. However, the suitability of the initial archetypes developed as part of the ocean observing evaluation (section 6.3) was reviewed and confirmed with an independent domain expert (see section 6.3.3.1). The implications of this are discussed further in section 7.4 below.

The development of a multi-purpose and generic reference model for ESS is not complete. While it was shown that existing data models such as O&M can be translated through a re-profiling, this limits the findings of the work to the area of observing and measuring environmental phenomena. Any all-encompassing reference model requires a significant amount of additional work, indeed the openEHR reference model was refined over a period of 10 years to reach a mature state within the Health domain. However, the success of re-profiling O&M provides a way forward in how to leverage and re-purpose other existing standardisation work to support the approach developed here. The outcomes of the evaluations described in chapter 6 show that INSPIRE provides a rich set of schemas that can readily inform the reference model development to underpin a two-level model approach.

7.1.2 Objective 2

Define a technical architecture to underpin a two-level model enabled spatial data infrastructure.

As per objective 1 above, objective 2 has been met within the context of the design science approach described in chapter 1. Arising from a conceptual prototype model, or set of well-developed theories described in chapter 4, a proof-of-concept technical framework and methodological concept model was defined (described in section 4.5). This ideal model was further evaluated and refined using a build method (as per section 1.6.3), again within the context of a design science paradigm build/evaluate cycle. A final proposed technical framework was refined through the build/evaluate cycle (presented in chapter 5), including a deployment view of the required supporting infrastructure to support the translated two-level modelling theories and approaches described in chapter 4.

Again, the findings are to be treated cautiously, but are ultimately very encouraging. The software components built to validate the technical architecture are not production

ready but do provide the basis for further development of the approach. This is a significant task which requires further investment. However, the findings arising from this work indicate that this work is worth-while and should be further explored within the research community.

7.1.3 Objective 3

Investigate to what extent two-level modelling can act as a solution for geo-observational sensor systems semantic interoperability.

Where chapter 5 describes a build and deployment approach to validate the theories and designs described in chapter 4. Chapter 6 describes two domain-based evaluations of the approach to ascertain the extent to which the approach can be applied to different scenarios. Again, the findings are very encouraging.

The approach was found to be flexible enough to capture all required domain-based concepts identified for given scenarios through an application domain review exercise and subsequent concept mapping and modelling work for the given evaluation scenarios. Also, of note, was that the approach was identified as being suitable to also act as an implementation approach to be used within existing standardisation efforts. For example, to redefine the SensorThings API data model as an archetype extension on top of the augmented O&M base reference model. The work on the SensorThings API also showed how the SensorThings information model can be transformed into a model of documentation (see section 6.2.3).

The Environmental Facilities Monitoring data model defined within the INSPIRE directive implementation guidance was identified as potentially benefiting from being redefined within the context of the two-level modelling approach developed here. Again, these are significant tasks and preliminary findings presented in this thesis suggest that this work should form workstreams within future research agendas.

It was also shown that oceanSITES standard employed within the EMODnet-Physics ocean data portal can be redefined using the two-level modelling approach and many of the identified semantic shortcomings of the netCDF standard could ultimately be addressed using the two-level modelling approach.

The evaluation approach is limited however in its ability to quantify to what extent semantic interoperability has been improved. Through a comparative analysis of the transformation of data instances, it was observed that an improvement in semantic interoperability traits had been achieved within two-level model-based data instance structures (see Table 6.4). However, at this point the true impact of this has not been accurately quantified. It can only be inferred that as the technicalities of the approach have been validated within the geo-spatial domain, the benefits demonstrated within the health domain over many years will also be seen within the geo-spatial domain. To fully investigate this a large community development effort is required, which was outside the scope of this work. However, the recommendation is that this effort is worthwhile and should be explored further.

7.1.4 Objective 4

Develop and make publicly available a library of geo-archetypes that can act as a proof-of-concept of two-level geospatial modelling and thus enable further exploration and adoption of two-level modelling within the geo-spatial community.

Arising from the work described in chapters 5 and 6, a set of geo-archetypes have been developed and made available to the wider community. For example, air quality, ocean observing and SensorThings API IoT based archetypes have been developed. These archetypes have not gone through any form of rigorous validation from a community-based, domain expert perspective and so their quality cannot be assumed to be sufficient for real world usage. They are only proof-of-concept archetypes. However, they do form

the basis for future archetype development and review, as well as providing a tangible example to the ESS community of geo-archetypes to encourage up take of the approach within the community.

7.1.5 Objective 5

Investigate mechanisms to enable a two-level modelling approach to be applied to the edge and beyond of constrained in situ geo-observational sensor systems

Chapters 4 and 5 describe a linked data approach to enable federated data streams, governed by archetypes across a two-level model-based infrastructure. The approach introduces the idea of *micro-contexts* and the *OPTaaS* to support the approach. The theory and designs underpinning this solution were validated and evaluated through implementation using the build and evaluate cycle within the design science paradigm.

The impact of the approach on generated data instances was seen to increase metadata and data instance size (see Table 6.4), however through implementation is was confirmed that this increase in size can be managed on in situ remote sensing platforms using the linked data federated data instance approach defined within chapter 4. The approach was validated against several constrained systems i.e. ARM based M3 and A8 boards running Contiki-NG and Linux. The approach was found to be successful in reducing the size of the data instance while maintaining the benefits associated with the application of the two-level modelling approach. This validation was performed using a Coniki-NG based two-level model kernel, supporting a linked data approach (see chapters 4 & 5).

For this work, the implementation again only serves as a proof-of-concept of the conceptual system design and translation approach defined within chapter 4. The efficiency of the approach in terms of time delay, memory and power usage and the scalability of the approach was only evaluated on a small scale of up to 10 observing and reporting platforms. Ultimately, within the scope of the evaluation performed the findings

showed the approach works well, however more work is needed to refine the approach and verify its scalability and impact on the constrained platform's longevity. This is especially true for observing platforms that are constrained to the level mandated by the ARM M3 architecture and typical communications networks used with remote in situ geo-observational systems.

7.1.6 Research Question Commentary

Can two-level modelling be translated from the health domain to the geo-spatial domain and applied to observing scenarios to improve semantic interoperability within and between spatial data infrastructures beyond what is possible with state-of-the-art approaches?

Translating two-level modelling to a new domain is a large undertaking. This work serves to highlight its applicability to the geo-spatial domain and develop an appropriate translational approach and techniques for realising the approach on technologically constrained systems. This work has shown that two-level modelling can be translated beyond the health domain and that its adoption within the geo-spatial domain has many benefits especially when applied to observing scenarios. The work shows semantic interoperability can be improved within ocean observing based spatial data infrastructures and ocean data portals. But to what quantifiable extent remains unclear and more investigation is required to confirm this.

This work is just the first step in a larger, new research agenda highlighting the applicability and potential benefits of a two level modelling approach to the wider geo-spatial community; and it is hoped these results will encourage wider uptake and investigation within the community.

7.2 Conclusion

Arising from this work it can be concluded that two-level modelling presents a viable approach to achieve semantic interoperability in constrained geo-observational sensor systems. The work presented here constitutes a proof-of-concept of the translational approach defined in chapter 4 and the reference architecture for two-level modelling supporting infrastructure deployment defined in chapter 6. It was found that selected geospatial data models and standards can be re-purposed to support an appropriate two-level reference model (chapters 4, 5 and 6). However, this requires a careful ontological analysis of each concept within selected data models.

Domain evaluations presented in chapter 6 support the hypothesis (section 1.3) that once the approach is translated and deployed, two-level modelling can enable diverse Earth system science domain experts to be the primary drivers of geo-observational sensor based digital artefacts. While the benefits of adopting a two-level information modelling approach to geospatial information modelling are potentially great, it was found that translation to a new domain is complex. The complexity of the approach was found to be a barrier to adoption, especially in commercial based projects where standards implementation is low on implementation road maps and the perceived benefits of standards adherence are low.

Due to limitations within the evaluations performed - especially where there was limited expert user input to the modelling process - the findings of this work are not exhaustive. The author recommends that based on the positive outcomes thus far, several research work streams should be commissioned to further evaluate and develop the research area. These are detailed throughout this chapter. It is recommended that a community-specific real-world pilot be undertaken modelled on the successful INSPIRE directive pilots to further the objectives of this research and that a European-wide

stakeholder group should be formed to seek additional EU research funds to realise a pilot project.

7.3 Future Directions

This section documents recommended future directions and outlines a future research agenda to continue the work. Here several open research questions arising from this research work are proposed and documented.

7.3.1.1 It's all About Community

A key differentiator of two-level modelling compared to other approaches is that it allows domain experts to be the primary drivers of Digital Earth Artefacts, while also ensuring that technical validity is maintained in one highly accessible and integrated process. This enables *extensibility*, a key component in a Digital Earth system. This view has also been expressed by domain experts (Clinicians) in the health domain (Garde et al., 2007). Also of note, from health domain experiences of two level models, are reports of reduced complexity of software and a greater focus on the realisation of useful applications (Chen et al., 2009); arising from a reduced demand for software model authoring tools. Any increased focus on the more convenient realisation of useful applications in ESS only further supports the realisation of the *functionality actions* provided by a Digital Earth system as compiled by Grossner et al. (2008).

Development of a mature, consensus-based repository of community-derived archetypes is a non-trivial task and requires established processes within any domain to ensure proper governance (Wollersheim et al., 2009) (Garde et al., 2007). However, with over 20 years of development experience, the technique is well supported by a strong theoretical and methodological framework. The true benefits of two-level modelling and archetypes are certainly realised when a large community consensus approach is employed, but Hoy et al. (2007) show how smaller local communities can also begin

seeing early dividends from two-level modelling, without a large archetype repository. This offers a way forward for new domains in terms of a parallel introduction of the technique. Should two-level modelling take hold as a preferred mechanism for the development of standard content models, migration of valuable ESS legacy systems to an archetype-based representation is possible due to the rich expressive power of archetypes (Chen et al., 2009). Adoption of a common format for content models leads to rapid and convenient installation and configuration of new metadata. Employment of two-level modelling techniques could potentially facilitate a nationally or conceivably an internationally standardised representation of all ESS content, as is alluded to in the health domain (Bernstein, 2009). This approach is ultimately about facilitating the pooling of high-quality data between Earth System Scientists and helping to develop critical Digital Earth based decision support systems.

7.3.1.2 Geo-Community Modelling Tools

Software tools to realise a two-level modelling methodology are complex to implement. For this work, there was no toolset available that is readily useable outside of the health domain. There are many barriers to the reuse clinical based tools. For instance, clinical domain modelling tools assume a static singular identity model, that of the Patient. However, as the author has demonstrated in this work, open source two-level modelling tools and components that were developed for the clinical domain can be re-used to aid ESS-facing tool development.

As discussed in chapter 4, the openEHR Java Reference Implementation is specifically designed for openEHR archetypes. The LinkEHR editor is a multi-reference model archetype editor which has enabled archetypes to be developed for this work. However, a tool specific to the geo-spatial domain is required.

7.3.1.3 Constrained knowledge Engine

The ongoing work to translate two-level modelling to constrained Earth system science based observational environment will adopt the concrete grammar approach described in (Käbisch, Peintenr and Anicic, 2015) and extend it to help realise a RDF/linked data style for a federated archetype-based instance data. The W3C Web of Things (WoT) Interest Group published a WoT Current Practices draft (W3C, 2017a), which also provides several proposed approaches, which could prove useful for the work presented here.

This work has shown that two level modelling can also be extended to the IoT domain through the mapping of the SensorThings API to appropriate data patterns within an augmented O&M based data model, and consequently encoding the SensorThings API data model as a set of extensible informational artefacts (archetypes, or an archetype model).

To ensure that IoT domain-based data streams are truly interoperable, metadata must be semantically rich enough for IoT systems to automatically bind disparate data streams. The SensorThings API data model provides a rich framework for achieving horizontal integration of IoT silos, enabling IoT systems-of-systems to be realised. However, the abstract nature of the SensorThings API data model means system developers must make local decisions about how to encode data structures for individual use-cases. Section 6.2 demonstrated that by transforming the SensorThings API data model, to a model of documentation, the model can be made less abstract, while retaining its wide use-case applicability.

Once mapped, modelled and published, these artefacts can enable a two-level modelling community of supporters to develop and grow within the IoT domain. Communities can agree on further specialization of the SensorThings API archetype model for individual IoT use cases and again publish these to be used within the

community or to enable systems to semantically integrate through rich querying made possible by the semantically rich datasets.

This approach has implications for the current implementation of SensorThings API. Transforming SensorThings API to a model of documentation changes the intention of several concepts (discussed in section 6.2.3). Mapping concepts to either the reference model or the archetype model ultimately determines the access API. To future-proof systems, the access API should ideally only implement reference model concepts. The wider ramifications of this would require further evaluation, while engaging domain practitioners in further work.

To further evaluate the applicability of this approach for individual use cases the author proposes that several pilot studies should be undertaken using a document model oriented SensorThings API archetype model as the basis for concept definition and system implementation. The W3C maintains an up to date of potential use-cases on their Website that could inform additional studies[95].

7.4 Contributions Summary

The research has shown how a two-level modelling approach that is applied to geo-observational systems design can act as a key enabler to a Digital Earth as proposed by Gore and contributes to the Digital Earth research agenda as defined by Craglia et al. (2012).

Finally, this work defines a new research agenda for two-level modelling approaches to be applicable outside of the current domain (health) for which they were originally developed.

The following specific contributions have arisen from this work:

[95] https://www.w3.org/2015/spatial/wiki/Working_Use_Cases

Major Contribution: A robust translation methodology for adopting two-level modelling from the health domain to other domains (such as the geo-spatial domain) has been defined.

Minor Contribution: An augmented O&M data model was adapted, redefined and encoded with appropriate patterns to support two-level modelling. The approach taken here also supports the robust translation methodology for specific use-cases such as observing environmental phenomena using in situ sensor-based systems.

Minor Contribution: A limited library of geo-archetypes was developed to support further two level geo-spatial modelling including base air quality, ocean monitoring and SensorThings API (IoT) archetype definitions and to demonstrate this knowledge engineering aspect of two-level modelling in the ESS domain.

Major Contribution: A reference geo-spatial two-level modelling framework design was developed to support two-level modelling within geo-observational scenarios. The validation of the framework design resulted in a proof-of-concept set of base software components and tools (geo-templating and constrained geo-templating kernels). Through this work an appropriate translation approach of the two-level modelling methodology for the Earth Systems Science Domain has been defined.

An assessment of relevant geographic information-based (ISO & OGC) standards, and their suitability to leverage a two-level modelling approach has been performed and reported. This work has demonstrated how key features (e.g. recursive aggregation pattern, ontology bindings) of the two-level modelling approach, required for the approach to be successful can be embedded into existing geo-information models to transform them from a “model-of-reality” to a “model-of-documentation”.

This work has also shown how the transformation of existing information models within the geo-spatial domain can be achieved while also allowing systems to adhere to existing standardisation requirements within their domain. This has been specifically demonstrated for the ISO/OGC standard Observations & Measurements. Arising from this translation work a novel profile of the O&M standard has been produced. The work has demonstrated to the ESS community how the novel profile of O&M can facilitate enhanced flexibility and extensibility in the recording of semantically interoperable observational data.

An XML encoding of the novel O&M profile has been developed and thus provided to the Earth System Science and ESS informatics community.

This work has demonstrated for the first time how a two-level modelling approach can be coerced (*trip-ified*) onto a Linked Data model for the purpose of allowing knowledge acquisition to occur at the edge of a constrained geo-sensor network. This has been achieved by the development of a novel Operational Templates as a Service (OPTaaS) to support the fragmentation of semantically rich data instances, which although small in size, remain linked to a two-level distributed knowledge framework, even while residing on remote in situ observational platform. Therefore, the OPTaaS provides a novel mechanism to enable *born semantic* data at the edge of sensing networks.

A novel resource constrained, two-level model knowledge kernel design for embedded devices has been defined.

An evaluation of the novel geo-spatial two-level modelling approach using two use-cases has been undertaken, which also provides the ESS community with reference use-cases to build future larger scale pilot studies of the two-level modelling approach within ESS application domains.

7.5 Final Remarks (Implications)

There are many existing data models in existence within the Earth system sciences that can be analysed using the approach that has been described throughout this work, and over time, the wider community can define new reference models to enable the two-level approach to proliferate throughout the wider community. Care must be taken in relation to the context of use of standards such as O&M. For example, in certain circumstances, O&M concepts may be better employed in helping to realise content level archetypes, with upper ontologies such as DOLCE UltraLite informing the reference model (discussed in chapter 4). This was found to be the case while under-taking a concept mapping of the SensorThings API data model (see section 6.2.3). The wider implications of this needs further exploration.

Of course, these are not trivial tasks, and the work here offers only the 1st step in what would be a long and complex process involving many stakeholders. However, the benefits of adoption are clear from the experiences of the health domain. And it is recommended that this work should continue.

Adoption of two-level modelling needs to be consensus-based. This process is slow. Support from the community must progress gradually. Community consensus is an important element of the approach. Failure to achieve the necessary volume of participation can render the large investment needed to achieve the benefits of two-level modelling redundant. The experiences of the SMART-IWRM project show that ESS based domain specialists can be reluctant to engage fully with collaborative domain modelling (Kämpgen, 2014).

This has also been the experience of the author during this work. To gain further insight to the approach with domain experts two conference based participatory workshops were proposed and accepted related to this work. One at the IEEE/MTS Oceans 2018

conference in Charleston USA and one at the IEEE World Forum on Internet of Things (WFIoT), Limerick, 2019. The workshop in Charleston did not go ahead as it did not attract enough interest from participants, although a paper was also presented at the oral sessions which generated some very interesting discussions and interest (Stacey and Berry, 2018).

The workshop at the WFIoT conference in 2019 did go ahead, but there were too few participants to generate the required discussion to gain any real insights into the approach that could contribute to the evaluations in the previous chapter. Interestingly the workshop attracted technical specialists interested in the approach, instead of non-technical domain specialists.

In lieu of gaining wider domain expert insights through conference workshop participation, the archetype modelling output domain expert review exercise detailed in 6.3.3.1 has provided additional evidence to support the hypothesis. During the review session, it was evident when discussing the limitation of modelling methods available to the domain expert that they were inadvertently referring to insufficient mechanisms to capture Popper's world 3 objects. It was evident that the domain expert was articulating an understanding of what would be referred to as Popper's world 3 and its relationship to world 1 (see section 3.1.1) and commenting on the limitations of current approaches only providing mechanisms to detail abstract objects. During the review session, the domain expert identified several failings of current ongoing projects that have adopted the traditional single level modelling approach. It was noted that in their experience, systems are implemented to achieve syntactic interoperability only, by way of solving the naming heterogeneity issue (ie. data standards and terminologies).

The review highlighted that within real systems implementations, it is world 1 objects, or objects of the physical world that are typically captured within datasets i.e. sensor type,

or measurement. However, current implementations do not allow for the capturing of world 3 objects, or products of the human mind. There is essentially no mechanism to record this information within the datasets, and in their own experience this remains in the mind of the domain expert or within ad-hoc non standardised pdf documents. It was commented that the approach presented by the author may well solve the limitations of their current approach however, it was felt that many stakeholders, who are often decision makers in terms of financing implementations would not fully appreciate the need to record this level of information.

Future investigations of the approach as part of ongoing projects are now under discussion. What is evident is that any future work would need a wider engagement exercise to be completed before progressing to wider pilot projects, and tangible benefits to adoption of two-level approaches need to be demonstrated to attract interest.

Attracting the interest of clinicians by two-level modelling advocates has also proved difficult within the health domain. On reflection of the evolution of the approach in health, the recommendation by one of the main architects of the approach is that buy-in can only be achieved when more mature domain specific modelling tools and systems begin to emerge (Beale 2019, personal communication, August 15th, 2019). Therefore, one of the primary limitations from the outcomes of this work thus far is that the tools and software components remain rudimentary and too underdeveloped to draw real interest from non-technical domain specialists.

There are some recent positive developments, where Earth system science based informaticians are reporting success in engaging with non-technical Earth system science domain expert participants in ontology development using new approaches such as “semantics smackdown” sessions (Leadbetter et al., 2016). For this work and the two-level modelling approach, the overall complexity of the approach can be a barrier to

adoption. Articulating the gains of the approach are difficult without a significant demonstration application. Conversely, it is difficult to develop a useful demonstration system without buy-in from the community. As wider knowledge of the importance of semantics within information systems development grows, so too will the willingness of domain experts to engage in the process. The challenge and the future goal of the work started here is to ensure that the tools needed to engage the community properly are ready at the same point that the community is ready to engage them.

This work has begun the process of attracting the interest of Earth system science informaticians (Leadbetter, Buck and Stacey, 2015) (Diviacco and Leadbetter, 2017). This is arguably the first step to broader community engagement and possible acceptance. In any case, the problems that the approach - demonstrated here – ultimately aims to address will not be solved within the short term and will be the focus of many research agendas for years to come. As Al Gore stated in 1998:

"Clearly, the Digital Earth will not happen overnight"

and Krishnamurti in 1964:

"There is no end to the journey, and that is the mystery, the beauty of it"

Attempting to capture the true complexity of human knowledge and wisdom within digital systems using approaches such as two-level modelling is a difficult endeavour and one that will continue for as long as there are humans and digital systems. Two-level modelling is but one more step along that journey.

Bibliography

Abdel-Tawwab, M., Monier, M.N., Hoseinifar, S.H. and Faggio, C. (2019) Fish response to hypoxia stress: growth, physiological, and immunological biomarkers. Fish Physiology and Biochemistry, 45(3), pp.997-1013.

Adjih, C., Baccelli, E., Fleury, E., Harter, G., Mitton, N., Noel, T., Pissard-Gibollet, R., Saint-Marcel, F., Schreiner, G., Vandaele, J. and Watteyne, T. (2015) FIT IoT-LAB: A large scale open experimental IoT testbed. In 2015 IEEE 2nd World Forum on Internet of Things (WF-IoT) (pp. 459-464). IEEE.

Ahad, M.A.R., Antar, A.D. and Ahmed, M. (2020) Devices and Application Tools for Activity Recognition: Sensor Deployment and Primary Concerns. In IoT Sensor-Based Activity Recognition (pp. 77-94). Springer, Cham.

Allemang, D. and Hendler, J. (2011) Semantic web for the working ontologist: effective modeling in RDFS and OWL. Elsevier.

Ali, A. and Aslam, M. (2020) Operating Systems for the Internet of Things: A Survey. LGURJCSIT, 4(2), pp.71-93.

Andreev, S. and Koucheryavy, Y. (2012) Internet of things, smart spaces, and next generation networking. Springer, LNCS, vol. 7469, p. 464.

Antoine, C. Sandrine, V. and Jean-Valery, F. (2017) JERICO-NEXT. Report on Developments Dedicated to Monitor and Study Benthic Comportment and Processes.D3.10.

Antoniou, G. and Van Harmelen, F. (2004) Web ontology language: Owl. In Handbook on ontologies (pp. 67-92). Springer, Berlin, Heidelberg.

Apache (2003) X.M.L., Project. Java XMLBeans.

Apache Jena (2010) A free and open source Java framework for building Semantic Web and Linked Data applications, [online]. Available at http://Jena.Apache.Org (Accessed: 12 October 2017).

AQL, Archetype Query Language 1.0.3 (2015) Available: http://www.openehr.org/releases/QUERY/latest/docs/AQL/AQL.html

Atzori, L. Iera, A. and Morabito, G. (2010) The internet of things: A survey. Computer networks, vol. 54, no. 15, pp. 2787–2805.

Baccelli, E., Hahm, O., Günes, M., Wählisch, M. & Schmidt, T.C. (2013) RIOT OS:

Towards an OS for the Internet of Things, Computer Communications Workshops (INFOCOM WKSHPS), 2013 IEEE Conference on IEEE, pp. 79.

Ballou, D.P. and Pazer, H.L. (1985) Modeling data and process quality in multi-input, multi-output information systems. Management science, 31(2), pp.150-162.

Balmaseda, M.A., Mogensen, K. and Weaver, A.T. (2013) Evaluation of the ECMWF ocean reanalysis system ORAS4. Quarterly Journal of the Royal Meteorological Society, 139(674), pp.1132-1161.

Barik, R.K., Tripathi, A., Dubey, H., Lenka, R.K., Pratik, T., Sharma, S., Mankodiya, K., Kumar, V. and Das, H. (2018) Mistgis: optimizing geospatial data analysis using mist computing. In Progress in Computing, Analytics and Networking (pp. 733-742). Springer, Singapore.

Barnett, T.P., Pierce, D.W., AchutaRao, K.M., Gleckler, P.J., Santer, B.D., Gregory, J.M. and Washington, W.M. (2005) Penetration of human-induced warming into the world's oceans. Science, 309(5732), pp.284-287.

Bajaj, G., Agarwal, R., Singh, P., Georgantas, N. & Issarny, V. (2017) A study of existing Ontologies in the IoT-domain, arXiv preprint arXiv:1707.00112.

Beale, T. (2000) OCEAN GEHR- compliant Kernel. Application Programmer's Interface. Available: https://openehr.org/static/files/resources/gehr_api.pdf.

Beale, T., Heard, S., Kalra, D., Lloyd, D. (2006) openEHR Architecture Overview, The openEHR Foundation, London.

Beale, T. and Heard, S. (2007) Archetype Definition Language, The openEHR Foundation, London.

Beale, T. (2002) Archetypes: Constraint-based domain models for future-proof information systems, OOPSLA workshop on behavioural semantics.

Beale, T. (2003) Towards Design Principles for Health Information Systems, International Journal of Medical Informatics.

Beaulieu, S.E., Fox, P.A., Di Stefano, M. et al. (2016) Toward cyber infrastructure to facilitate collaboration and reproducibility for marine integrated ecosystem assessments Earth Sci Inform. doi:10.1007/s12145-016-0280-4

Berners-Lee, T., Cailliau, R., Groff, J. & Pollermann, B. (1992) World-Wide Web: the information universe, Internet Research, vol. 2, no. 1, pp. 52-58.

Bernstein, K., Tvede, I., Petersen, J. and Bredegaard, K. (2009) Can openEHR archetypes be used in a national context? The Danish archetype proof-of-concept project. In MIE (pp. 147-151).

Bernstein, A. and Kaufmann, E. (2006) Gino-a guided input natural language ontology editor. In International Semantic Web Conference (Vol. 2006, pp. 144-157).

Berry, D., Ven, J., Freriks, G., Moner, D. (2010) Use of OIDs and IIs in EN13606.

Berteaux, H. O. (1976), Buoy Engineering, John Wiley & Sons, 1976.

Bhatti, S., Carlson, J., Dai, H., Deng, J., Rose, J., Sheth, A., Shucker, B., Gruenwald, C., Torgerson, A. & Han, R. (2005) MANTIS OS: An embedded multithreaded operating system for wireless micro sensor platforms, Mobile Networks and Applications, vol. 10, no. 4, pp. 563-579.

Bisbal, J. and Berry, D. (2009) Archetype Alignment-A Two-level Driven Semantic Matching Approach to Interoperability in the Clinical Domain. In HEALTHINF (pp. 216-221).

Bizer, C. Heath, T. and Berners-Lee, T. (2009) Linked data-the story so far, Semantic Services, Interoperability and Web Applications: Emerging Concepts, pp. 205-227.

Bittner, T., Donnelly, M. and Smith, B., (2009) A spatio-temporal ontology for geographic information integration. International Journal of Geographical Information Science, 23(6), pp.765-798.

Blauw, A.N., Beninca, E., Laane, R.W., Greenwood, N. and Huisman, J. (2012) Dancing with the tides: fluctuations of coastal phytoplankton orchestrated by different oscillatory modes of the tidal cycle. PloS one, 7(11), p.e49319.

Blauw, A.N. (2015) Monitoring and prediction of phytoplankton dynamics in the North Sea. 9789462597723.

Blobel, B., Moner, D., Hildebrand, C., Robles, M. (2010) Standardized and flexible health data management with an archetype driven EHR system (EHRflex). In Seamless Care, Safe Care: The Challenges of Interoperability and Patient Safety in Health Care: Proceedings of the EFMI Special Topic Conference, June 2-4, Reykjavik, Iceland (Vol. 155, p. 212). IOS Press.

Blower, J.D., Masó, J., Díaz, D., Roberts, C.J., et al. (2015) Communicating thematic data quality with web map services. ISPRS International Journal of Geo-Information, vol. 4, no. 4, pp. 1965-1981.

Bodner, R.C. and Song, F. (1996) May. Knowledge-based approaches to query expansion in information retrieval. In Conference of the Canadian Society for Computational Studies of Intelligence (pp. 146-158). Springer, Berlin, Heidelberg.

Boegl, K., Adlassnig, K., Hayashi, Y., Rothenfluh, T.E. & Leitich, H. (2004) Knowledge acquisition in the fuzzy knowledge representation framework of a medical consultation system, Artificial Intelligence in Medicine, vol. 30, no. 1, pp. 1-26.

Boldt, D., Hasemann, H., Karnstedt, M., Kröeller, A. and Von Der Weth, C. (2015) Sparql for networks of embedded systems. In 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT) (Vol. 1, pp. 93-100). IEEE.

Bonnett, A. (2008) What is geography? Sage.

Boscá Tomás, D. (2016) Detailed Clinical Models & Their Relationship with Electronic Health Records (Doctoral dissertation).

Botts, M., Percivall, G., Reed, C. & Davidson, J. (2008) OGC® sensor web enablement: Overview and high level architecture in GeoSensor networks Springer, pp. 175-190.

Borges, B. D. (2008) Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities, Revista Negotium, (10), pp. 89-91, 2008.

Boulton, G. (2018) The challenges of a Big Data Earth, Big Earth Data, 2:1, 1-7, DOI: 10.1080/20964471.2017.1397411

Breunig, M., Bradley, P.E., Jahn, M., Kuper, P., Mazroob, N., Rösch, N., Al-Doori, M., Stefanakis, E. and Jadidi, M. (2020) Geospatial data management research: Progress and future directions. ISPRS International Journal of Geo-Information, 9(2), p.95.

Broekstra, J., Kampman, A. and Van Harmelen, F. (2002) Sesame: A generic architecture for storing and querying rdf and rdf schema. In International semantic web conference (pp. 54-68). Springer, Berlin, Heidelberg.

Brodeur, J., Coetzee, S., Danko, D., Garcia, S. and Hjelmager, J. (2019) Geographic information metadata—an outlook from the international standardization perspective. ISPRS International Journal of Geo-Information, 8(6), p.280.

Bröring, A., Echterhoff, J., Jirka, S., Simonis, I., Everding, T., Stasch, C., Liang, S. and

Lemmens, R. (2011a) New generation sensor web enablement. Sensors, 11(3), pp.2652-2699.

Bröring, A., Maué, P., Janowicz, K., Nüst, D. and Malewski, C. (2011b) Semantically-enabled sensor plug & play for the sensor web, Sensors, vol. 11, no. 8, pp. 7568-7605.

Bröring, A., Stasch, C. and Echterhoff, J. (2012) OGC sensor observation service interface standard. Open Geospatial Consortium Interface Standard, pp.12-006.

Bruegge, B. and Dutoit, A.H. (2009) Object--Oriented Software Engineering. Using UML, Patterns, and Java. Learning, 5(6), p.7.

Buck, J. and Leadbetter, A. (2015) Born semantic: Linking data from sensors to users and balancing hardware limitations with data standards, in EGU General Assembly Conference Abstracts, pp. 3781.

Cao, Q. & Abdelzaher, T. (2006) LiteOS: a lightweight operating system for C software development in sensor networks, Proceedings of the 4th international conference on Embedded networked sensor systemsACM, , pp. 361.

Car, N.J., Ip, A. and Druken, K. (2017) netCDF-LD SKOS: Demonstrating linked data vocabulary use within netCDF-compliant files. In Environmental Software Systems. Computer Science for Environmental Protection: 12th IFIP WG 5.11 International Symposium, ISESS 2017, Zadar, Croatia, May 10-12, 2017, Proceedings 12 (pp. 329-337). Springer International Publishing.

Carswell, J.D. and Yin, J. (2012) Mobile spatial interaction in the Future Internet of Things. In 2012 20th International Conference on Geoinformatics (pp. 1-6). IEEE.

Castellani, A.P., Gheda, M., Bui, N., Rossi, M. and Zorzi, M. (2011) Web Services for the Internet of Things through CoAP and EXI. In 2011 IEEE International Conference on Communications Workshops (ICC) (pp. 1-6). IEEE.

Castell, N., Viana, M., Minguillón, M.C., Guerreiro, C. and Querol, X. (2013) Real-world application of new sensor technologies for air quality monitoring. ETC/ACM Technical Paper.

Cetl, V., Nunes de Lima, V.,Tomas, R., Lutz, M., D'Eugenio, J., Nagy, A., Robbrecht, J. (2017) Summary Report on Status of implementation of the INSPIRE Directive in EU, EUR 28930 EN, Publications Office of the European Union, Luxembourg, 2017, ISBN 978-92-79-77058-6, doi:10.2760/143502, JRC109035.

Charpenay, V., Käbisch, S. and Kosch, H., (2017), May. uRDF Store: Towards

Extending the Semantic Web to Embedded Devices. In European Semantic Web Conference (pp. 76-80). Springer, Cham.

Chen, R., Garde, S., Beale, T., Nyström, M., Karlsson, D., Klein, G.O. and Åhlfeldt, H. (2008) An archetype-based testing framework. Studies in health technology and informatics, 136, p.401.

Chen, R., Klein, G.O., Sundvall, E., Karlsson, D. and Åhlfeldt, H. (2009) Archetype-based conversion of EHR content models: pilot experience with a regional EHR system. BMC medical informatics and decision making, 9(1), p.33.

Chen, R. & Klein, G. (2007) The openEHR Java reference implementation project, Studies in health technology and informatics, vol. 129, no. 1, pp. 58.

Chen, X. (2016) A two-level identity model to support interoperability of identity information in electronic health record systems. Doctoral Thesis, Dublin Institute of Technology. doi:10.21427/D7XG72.

Chodorow, K. (2013) MongoDB: the definitive guide: powerful and scalable data storage. O'Reilly Media, Inc.

CIMI (2020) Clinical Information Modeling Initiative, [online]. Available at https://cimi.hl7.org/, (Accessed: 28th August 2020).

Coetzee, S., Ivánová, I., Mitasova, H. and Brovelli, M.A. (2020) Open geospatial software and data: A review of the current state and a perspective into the future. ISPRS International Journal of Geo-Information, 9(2), p.90.

Copernicus (2017) Available: http://www.copernicus.eu/

Council, N.A. & Earth System Sciences Committee (1986) Earth System Science Overview: A Program for Global Change, National Aeronautics and Space Administration.

Clifford, D., Alegre, R., Bennett, V., Blower, et al. (2016) Capturing and sharing our collective expertise on climate data: the CHARMe project. Bulletin of the American Meteorological Society, vol. 97, no. 4, pp. 531-539.

Cointet, J.P. and Chavalarias, D. (2008) Science mapping with asymmetrical paradigmatic proximity. arXiv preprint arXiv:0803.2315.

Collina, M. (2013) node-coap library, [online]. Available at https://github.com/mcollina/node-coap (Accessed: 11 April 2017).

Columbus Consortium (2016) Marine Portals and Repositories and their role in Knowledge Transfer to support Blue Growth, 2016.

Compton, M., Barnaghi, P., Bermudez, L., García-Castro, R., Corcho, O., Cox, S., Graybeal, J., Hauswirth, M., Henson, C. & Herzog, A. (2012) The SSN ontology of the W3C semantic sensor network incubator group, Web Semantics: Science, Services and Agents on the World Wide Web, vol. 17, pp. 25-32.

Copernicus Marine In Situ Tac Data Management Team (2017) Copernicus in situ TAC - CMEMS System Requirements Document. http://doi.org/10.13155/40846, 2017.

Copernicus Marine (2018) Available: http://marine.copernicus.eu/situ-thematic-centre-ins-tac/. [Accessed: 02-Jul-2018].

Cova, T.J. and Goodchild, M.F. (2002) Extending geographical representation to include fields of spatial objects. International Journal of geographical information science, 16(6), pp.509-532.

Cox, S. (2006) Observations and measurements, Open Geospatial Consortium Best Practices Document. Open Geospatial Consortium.

Cox, S. (2015a) Ontology alignment - is PROV-O good enough? Open Geospatial Consortium GeoMatics Summit, Boulder, CO.

Cox, S. (2015b) Pitfalls in alignment of observation models resolved using PROV as an upper ontology. In AGU Fall Meeting Abstracts.

Cox, S. and Taylor, P. (2015) OGC Observations and Measurements JSON implementation, Open Geospatial Consortium, Tech. Rep. Discussion Paper 15-100r1. [online]. Available at http://www. opengis.net/doc/dp/om-json/.

Craglia, M., de Bie, K., Jackson D., et al. (2012) Digital Earth 2020: towards the vision for the next decade. Int J Digit Earth 5:4–21.

Crutzen, P.J. and Stoermer, E.F. (2000) The “Anthropocene.” Global Change Newsletter 41, 17–18. International Geosphere–Biosphere Programme (IGBP).

Crutzen, P.J. (2002) Geology of mankind. Nature, 415(6867), pp.23-23.

Crutzen, P.J. (2006) The “anthropocene”. In Earth system science in the anthropocene (pp. 13-18). Springer Berlin Heidelberg.

Dangermond, J. and Goodchild, M.F. (2019) Building geospatial infrastructure. Geo-spatial Information Science, pp.1-9.

Davis, E. (2005) Attribute Conventions for Data Discovery, [online]. Available at https://www.unidata.ucar.edu/software/thredds/v4.3/netcdf-java/formats/DataDiscoveryAttConvention.html. (Accessed: 02-Jul-2018).

De Abreu, D., Flores, A., Palma, G., Pestana, V., Pinero, J., Queipo, J., Sánchez, J. and Vidal, M.E. (2013) October. Choosing Between Graph Databases and RDF Engines for Consuming and Mining Linked Data. In Cold.

De, S., Christophe, B. & Moessner, K. (2014) Semantic enablers for dynamic digital–physical object associations in a federated node architecture for the Internet of Things, Ad Hoc Networks, vol. 18, pp. 102-120.

de la Hidalga, A.N., Hardisty, A., Martin, P., Magagna, B. and Zhao, Z. (2020) The ENVRI Reference Model. In Towards Interoperable Research Infrastructures for Environmental and Earth Sciences (pp. 61-81). Springer, Cham.

Delin, A. K., and Jackson P. J. (2001) The sensor web: a new instrument concept. Proceedings of SPIE's Symposium on Integrated Optics.

Dell'Aglio, D., Eiter, T., Heintz, F. and Le Phuoc, D. (2019) Special issue on stream reasoning. Semantic Web, 10(3), pp.453-455.

Deltares (2018) D-Water Quality Manual, [online]. Available at Index of /delft3d/manuals (deltares.nl). (Accessed: July 2018).

Dentler, K., ten Teije, A., Cornet, R. and de Keizer, N. (2012) Semantic integration of patient data and quality indicators based on openEHR archetypes. In Process support and knowledge representation in health care (pp. 85-97). Springer, Berlin, Heidelberg.

Dexter, P., Summerhayes, C.P., Pugh, D. and Holland, G. (2010) Ocean observations–the Global Ocean Observing System (GOOS). D. Pugh, G. Holland G (Eds.), Troubled waters: ocean science and governance, Cambridge University Press, Cambridge, pp.161-178.

Directive, I. (2007) Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE), Published in the official Journal on the 25th April.

Diviacco, P., De Cauwer, K., Leadbetter, A., Sorribas, J., Stojanov, Y., Busato, A. and Cova, A. (2015) Bridging semantically different paradigms in the field of marine acquisition event logging. Earth Science Informatics, 8(1), pp.135-146.

Diviacco, P. and Leadbetter, A. (2017) Balancing Formalization and Representation in Cross-Domain Data Management for Sustainable Development. Oceanographic and Marine Cross-Domain Data Management for Sustainable Development. IGI Global, 23-46.

Dockter, A.H., Murdoch, S., Faber, P., Niederwieser, D., Deboer, L.D. and Gröschke, R. (2017) The Gradle Build Tool.

DoD, U.S. (2007). Department of Defense dictionary of military and associated terms. Joint Publication, pp.1-02.

Dolin, R.H., Alschuler, L., Boyer, S., Beebe, C., Behlen, F.M., Biron, P.V. and Shabo, A. (2006). HL7 clinical document architecture, release 2. Journal of the American Medical Informatics Association, 13(1), pp.30-39.

Douglas, H. (Retrieved October 2017) Online Etymology Dictionary, [online]. Available at http://etymonline.com.

Dogac, A. et al. (2006) Artemis: Deploying semantically enriched Web services in the healthcare domain, Inf Syst, vol. 31, pp. 321-339, 2006.

Duckham, M. (2008). Ambient Spatial Intelligence: Decentralised spatial computing in geosensor networks. In Keynote, Workshop on Geosensor Networks, Hanover, Germany, February 20-22, 2008.

Duckham, M., Bennett, R., BISHOP, I., Fraser, C., Kealy, A., Leach, J., Ogleby, C., Rajabifard, A., Williamson, I. and Winter, S. (2010) Ambient spatial intelligence for sustainable cities. Australia, University of Melbourne, Parkville, Victoria.

Dunkels, A., Gronvall, B. & Voigt, T. (2004) Contiki-a lightweight and flexible operating system for tiny networked sensors, Local Computer Networks, 2004. 29th Annual IEEE International Conference on IEEE, pp. 455.

Duquennoy, S. (2017) Contiki-NG: The OS for Next Generation IoT Devices, [online]. Available at https://github.com/contiki-ng/contiki-ng.

Eaton, B., Gregory, J., Drach, B., Taylor, K., Hankin, S., Caron, J., Signell, R., Bentley, P., Rappa, G., Höck, H. and Pamment, A. (2003) NetCDF Climate and Forecast (CF) metadata conventions.

Edenhofer, O. (ed) (2014) Climate Change 2014: Mitigation of Climate Change: Summary for Policymakers; Working Group III Contribution to the Fifth Assessment Report AR5 of the Intergovernmental Panel on Climate Change. IPCC.

EO/Copernicus (2016) Europe's Copernicus programme Overview, [online]. Available at http://www.esa.int/Our_Activities/Observing_the_Earth/Copernicus/Overview3.

ECHO, Earth Observing System Clearing House (2005) ECHO Overview, [online]. Available at http://www.echo.eos.nasa.gov/overview/.

Egenhofer, M.J., (2002) Toward the semantic geospatial web. In Proceedings of the 10th ACM international symposium on Advances in geographic information systems (pp. 1-4).

El Kaed, C. and Boujonnier, M. (2017) Forte: A federated ontology and timeseries query engine. In 2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData) (pp. 983-990). IEEE.

Elenabaas, H. (2018) Changing of the guard? The rise of spatial data for reporting on the SDGs in the Global South, [online]. Available at http://www.thebrokeronline.eu/Blogs/Inclusive-Economy-Africa/Changing-of-the-guard-The-rise-of-spatial-data-for-reporting-on-the-SDGs-in-the-Global-South.

Elsts, A., Strazdins, G., Vihrov, A. & Selavo, L. (2012) Design and Implementation of MansOS: a Wireless Sensor Network Operating System. In Scientific Papers; University of Latvia: Riga, Latvia; Volume 787, pp. 79–105

EMODnet Ingestion Portal, [online]. Available at https://www.emodnet-ingestion.eu/. (Accessed: 02-Jul-2018).

EMODnet (2018) EMODnet compliance with the INSPIRE Directive: a matter of fact, [online]. Available at http://Emodnet.eu/emodnet-compliance-inspire-directive-matter-fact. (Accessed: 02-Jul-2018).

Estrin, D., Govindan, R., Heidemann, J. & Kumar, S. (1999) Next century challenges: Scalable coordination in sensor networks, Proceedings of the 5th annual ACM/IEEE international conference on Mobile computing and networkingACM, , pp. 263.

European Commission (2010) European Marine Observation and Data Network, [online]. Available http://www.emodnet-physics.eu/. (Accessed: April 2018).

European Commission (2012) Green Paper-Marine Knowledge 2020—from seabed mapping to ocean forecasting.

European Commission (2016) Marine INSPIRE Pilot, [online]. Available at

http://inspire-marine.jrc.ec.europa.eu/. (Accessed: 02-Jul-2018).

Eyres, H. (2017). A Short History of Earth Observation. In Seeing Our Planet Whole: A Cultural and Ethical View of Earth Observation (pp. 77-87). Springer International Publishing.

Fernandez, S., Marsa-Maestre, I., Velasco, J.R. & Alarcos, B. (2013) Ontology Alignment Architecture for Semantic Sensor Web Integration, Sensors, vol. 13, no. 9, pp. 12581-12604.

Fernandez-Breis, J. T. et al. (2007) Poseacle: An ontological infrastructure for managing clinical archetypes in semantic web environments, in Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems, pp. 2299.

Fielding, R.T. & Taylor, R.N. (2002) Principled design of the modern Web architecture, ACM Transactions on Internet Technology (TOIT), vol. 2, no. 2, pp. 115-150.

Fischer, A. et al. (2016) Initial AtlantOS Requirements Report, [online]. Available at 1.1 Initial AtlantOS Requirements Report.pdf (atlantos-h2020.eu).

Flemming, N.C. (1995) Making the case for GOOS. Sea Technology, Special Feature, pp.44-49.

Fowler, M. (2010) Domain-specific languages. Pearson Education.

Fraisl, D., Campbell, J., See, L., Wehn, U., Wardlaw, J., Gold, M., Moorthy, I., Arias, R., Piera, J., Oliver, J.L. and Masó, J. (2020) Mapping citizen science contributions to the UN sustainable development goals. Sustainability Science, pp.1-17.

Fredericks, J. and Botts, M. (2018) Promoting the capture of sensor data provenance: a role-based approach to enable data quality assessment, sensor management and interoperability. Open Geospatial Data, Software and Standards, 3(1), pp.1-8.

Friedman, D.P. and Wand, M. (1984) Reification: Reflection without metaphysics. In Proceedings of the 1984 ACM Symposium on LISP and functional programming (pp. 348-355).

Gahegan, M. & Pike, W. (2006) A situated knowledge representation of geographical information, Transactions in GIS, vol. 10, no. 5, pp. 727-749.

Garde, S., Hovenga, E. J., Granz, J., Foozonkhah, S. and Heard, S. (2007) Managing archetypes for sustainable and semantically interoperable electronic health records. eJHI electronic Journal of Health Informatics 2(2): 10.

Garde, S., Chen, R., Leslie, H., Beale, T., McNicoll, I. and Heard, S. (2009) Archetype-based knowledge management for semantic interoperability of electronic health records. In MIE (pp. 1007-1011).

Gay, D., Levis, P., Von Behren, R., Welsh, M., Brewer, E. & Culler, D. (2003) The nesC language: A holistic approach to networked embedded systems, Acm Sigplan NoticesACM, pp. 1.

GEMET (2012) General Multilingual Environmental Thesaurus. [online]. Available at http://www.eionet.europa.eu/gemet/.

GEO (2016) Group on Earth Observations, [online]. Available at https://www.earthobservations.org/index.php.

Geographical Sciences Committee (2005) Learning to think spatially. National Academies Press.

GeoServer (2019) About GeoServer, [online]. Available at http://geoserver.org/. (Accessed June 2019)

Gerritsen, H., De Vries, H. and Philippart, M. (1995) The Dutch continental shelf model. Coastal and estuarine studies, pp.425-425.

Gilbert, S. and Lynch, N. (2002) Brewer's conjecture and the feasibility of consistent, available, partition-tolerant web services. ACM Sigact News, 33(2), pp.51-59.

Gillam, L., Tariq, M. and Ahmad, K. (2005) Terminology and the construction of ontology. Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication, 11(1), pp.55-81.

Glaves, H., Schaap, D., Arko, R. and Proctor, R. (2014) Ocean Data Interoperability Platform (ODIP): supporting the development of a common global framework for marine data management through international collaboration. In EGU General Assembly Conference Abstracts (Vol. 16).

Goldberg, D., Olivares, M., Li, Z. and Klein, A.G. (2014) Maps & GIS data libraries in the era of big data and cloud computing. Journal of Map & Geography Libraries, 10(1), pp.100-122.

Gomes, V.C., Queiroz, G.R. and Ferreira, K.R. (2020) An overview of platforms for big earth observation data management and analysis. Remote Sensing, 12(8), p.1253.

Goodchild, M.F. (1991) Just the facts, Political Geography Quarterly, vol. 10, no. 4, pp. 335-337.

Goodchild, M.F. (2006) GIScience ten years after Ground Truth, Transactions in GIS, vol. 10, no. 5, pp. 687-692.

Goodchild, M.F. (2007) Citizens as sensors: the world of volunteered geography. GeoJournal, 69(4), pp.211-221.

Goodchild, M.F. (2010) Twenty years of progress: GIScience in 2010, Journal of spatial information science, vol. 2010, no. 1, pp. 3-20.

Goodchild, M.F. (2020) How well do we really know the world? Uncertainty in GIScience. Journal of Spatial Information Science, 2020(20), pp.97-102.

Goodwin, C. & Russomanno, D.J. (2006) An ontology-based sensor network prototype environment, Proceedings of the Fifth International Conference on Information Processing in Sensor Networks, pp. 1.

GOOS (2019) The Global Ocean Observing System 2030 Strategy. IOC, Paris, 2019, IOC Brochure 2019-5 (IOC/BRO/2019/5 rev.)

Gore, A. (1998) The digital Earth: understanding our planet in the 21st century, Aust Surv 43:89–91.

Gottschalk, K., Graham, S., Kreger, H. and Snell, J. (2002) Introduction to web services architecture. IBM systems Journal, 41(2), pp.170-177.

Gregory, J. (2003) The CF metadata standard, CLIVAR Exchanges, vol. 8, (4), pp. 4.

Grimson J. et al. (1996) Synapses - Federated Healthcare Record Server. Procs. Of MIE 96, Copenhagen, J Brender et al. (Eds), IOS Press, 695-699.

Grimson, J., Grimson, W., Berry, D., et al. (1998) A CORBA-based integration of distributed electronic healthcare records using the synapses approach. IEEE Transactions on information technology in biomedicine, vol. 2, no. 3, pp. 124-138.

Grimson, J., Grimson, W. and Hasselbring, W. (2000) The SI challenge in health care. Communications of the ACM, 43(6), pp.48-55.

Grossner, K.E., Goodchild, M.F. and Clarke, K.C. (2008) Defining a digital earth system. Transactions in GIS, 12(1), pp.145-160.

Gruber, T.R. (1993) A translation approach to portable ontology specifications. Knowledge acquisition, 5(2), pp.199-220.

Gudgin, M., Hadley, M., Mendelsohn, N., Moreau, J., Nielsen, H.F., Karmarkar, A. &

Lafon, Y. (2003) SOAP Version 1.2, W3C recommendation, vol. 24.

Guo, H., Goodchild, M.F. and Annoni, A., (eds.) (2020) Manual of digital Earth (p. 852). Springer Nature.

Hamilton, C. and Grinevald, J. (2015) Was the Anthropocene anticipated? The Anthropocene Review, 2(1), pp.59-72.

Han, C., Kumar, R., Shea, R., Kohler, E. & Srivastava, M. (2005) A dynamic operating system for sensor nodes, Proceedings of the 3rd international conference on Mobile systems, applications, and services ACM, pp. 163.

Hart, J. and Martinez, K. (2020) Sensor Networks and Geohazards in Reference Module in Earth Systems and Environmental Sciences, Elsevier.

Hart, J. and Martinez, K. (2006) Environmental Sensor Networks: A revolution in the earth system science? Earth-Science Reviews, vol. 78, no. 3, pp. 177-191.

Harvey, F. and Chrisman, N. (1998) Boundary objects and the social construction of GIS technology. Environment and planning A, 30(9), pp.1683-1694.

Hasemann, H., Kröller, A. and Pagel, M. (2012) RDF Provisioning for the Internet of Things. In 2012 3rd IEEE International Conference on the Internet of Things (pp. 143-150). IEEE.

Hasemann, H., Kröller, A. and Pagel, M. (2014) The wiselib tuplestore: a modular RDF database for the internet of things. arXiv preprint arXiv:1402.7228.

Hastie, T.J. (2017) Generalized additive models. In Statistical models in S (pp. 249-307). Routledge.

Hastie, T.J. and Tibshirani, R.J. (1990) Generalized additive models (Vol. 43). CRC press.

Hayes, J., O'Connor, E., Cleary, J., Kolar, H., McCarthy, R., Tynan, R., O'Hare, G.M., Smeaton, A.F., O'Connor, N.E. & Diamond, D. (2009) Views from the coalface: chemo-sensors, sensor networks and the semantic sensor web, Corcho, O., Hauswirth, M., Koubarakis, M.(eds.). Proceedings of the 1st SemSensWeb2009 Workshop on the Semantic Sensor Web Collocated with ESWC 2009 Heraklion, Crete, Greece.

Heard, T. and Beale, T. (1996) The Good Electronic Health Record (GeHR) (Australia).

Heidt, H., Puig-Suari, J., Moore, A., Nakasuka, S. and Twiggs, R. (2000). CubeSat: A

new generation of picosatellite for education and industry low-cost space experimentation.

Henver, A, March, S.T., Park, J. and Ram, S. (2004) Design science in information systems research. MIS quarterly, 28(1), pp.75-105.

Hill, B. W. (2015) Evaluation of Efficient XML Interchange (EXI) for Large Datasets and as an Alternative to Binary JSON Encodings.

Hill, J., Szewczyk, R., Woo, A., Hollar, S., Culler, D. & Pister, K. (2000) System architecture directions for networked sensors, ACM SIGOPS operating systems reviewACM, pp. 93.

Holt, M. (2003) Towards NOOS—the eurogoos nw shelf task team 1996–2002. In Elsevier Oceanography Series (Vol. 69, pp. 461-465). Elsevier.

Honeywill, C., Paterson, D.M. and Hagerthey, S.E, (2002) Determination of microphytobenthic biomass using pulse-amplitude modulated minimum fluorescence. European Journal of Phycology, 37(4), pp.485-492.

Hordoir, R., Axell, L., Höglund, A., Dieterich, C., Fransner, F., Groger, M., Liu, Y., Pemberton, P., Schimanke, S., Andersson, H. and Ljungemyr, P. (2019) Nemo-Nordic 1.0: a NEMO-based ocean model for the Baltic and North seas-research and operational applications. Geoscientific Model Development, 12(1), pp.363-386.

Hoy, D., Hardiker, N. R., McNicoll, I. T., Westwell, P. and Bryans, A. (2009) Collaborative development of clinical templates as a national resource. International Journal of Medical Informatics 78 Suppl 1: S95-100.

Horrocks, I. (2008) Ontologies and the semantic web, Communications of the ACM, vol. 51, no. 12, pp. 58-67.

Horsburgh, J.S., Tarboton, D.G., Maidment, D.R. and Zaslavsky, I. (2008) A relational model for environmental and water resources data. Water Resources Research, 44(5).

Horsburgh, J.S., Caraballo, J., Ramírez, M., Aufdenkampe, A.K., Arscott, D.B. and Damiano, S.G. (2019) Low-cost, open-source, and low-power: But what to do with the data? Frontiers in Earth Science, 7(67), p.1.

Horsburgh, J.S., Aufdenkampe, A.K., Mayorga, E., Lehnert, K.A., Hsu, L., Song, L., Jones, A.S., Damiano, S.G., Tarboton, D.G., Valentine, D. and Zaslavsky, I. (2016) Observations Data Model 2: A community information model for spatially discrete Earth observations. Environmental Modelling & Software, 79, pp.55-74.

Houghton, J.T., Jenkins, C.J. & Ephraums, J.J. (eds) (1990) Climate change: the IPCC scientific assessment. Mass, Cambridge.

Howarth, C., Parsons, L. and Thew, H. (2020) Effectively communicating climate science beyond academia: harnessing the heterogeneity of climate knowledge. One Earth, 2(4), pp.320-324.

Hsu, L., Mayorga, E., Horsburgh, J., Carter, M., Lehnert, K. and Brantley, S. (2017) Enhancing Interoperability and Capabilities of Earth Science Data using the Observations Data Model 2 (ODM2). Data Science Journal, 16.

Hu, T., Yang, J., Li, X. and Gong, P. (2016) Mapping urban land use by using landsat images and open social data. Remote Sensing, 8(2), p.151.

IETF (2015) Int Area Wiki - Internet-of-Things Directorate. IOTDirWiki. IETF, n.d. Web. http://trac.tools.ietf.org/area/int/trac/wiki/IOTDirWik

Iorga, M., Feldman, L., Barton, R., Martin, M., Goren, N. and Mahmoudi, C. (2017) The NIST definition of fog computing (No. NIST Special Publication (SP) 800-191 (Draft)). National Institute of Standards and Technology.

Ingram, D. (1995). The good European health record. Health in the New Communication Age, pp.66-74.

INSPIRE (2007) INSPIRE Directive. EC of the European Parliament and of the Council of, 14.

INSPIRE (2013a) D2.8.II/III.7 INSPIRE Data Specification on Environmental Monitoring Facilities –Technical Guidelines, [online]. Available at https://inspire.ec.europa.eu/file/1607/download?token=620R7l7F

INSPIRE (2013b) Technical Guidance for the implementation of INSPIRE Download Services version 3.0, 2013.

INSPIRE (2016) D2.9 INSPIRE Guidelines for the use of Observations & Measurements and Sensor Web Enablement-related standards in INSPUIRE Annex II and II data specification development. ONLINE, Available: https://inspire.ec.europa.eu/file/1638/download?token=EtGplOtQ

Irwin, A.J. and Finkel, Z.V. (2008) Mining a sea of data: Deducing the environmental controls of ocean chlorophyll. PloS one, 3(11), p.e3836.

ISO/TC037 (2000) Terminology work -- Vocabulary -- Part 1: Theory and application 1087-1:2000.

ISO/TC211 (2011) Geographic information -- Observations and measurements

ISO/TC215 (2006) Health Informatics-Electronic health record communication-Part1: Reference model.

ISO/TC215 (2008) Health Informatics-Electronic health record communication-Part2: Archetype interchange specification.

ISO/TC215 (2009a) Health informatics-Electronic health record communication-Part 3: Reference archetypes and term lists.

ISO/TC215 (2009b) Health Informatics-Electronic Health Record Communication-Part 4: Security.

ISO/TC211 (2013) 19157 Geographic Information – Data Quality.

Janowicz, K., Haller, A., Cox, S.J., Le Phuoc, D. and Lefrançois, M. (2019) SOSA: A lightweight ontology for sensors, observations, samples, and actuators. Journal of Web Semantics, 56, pp.1-10.

Jiang, Y., Li, J. and Guo, Z. (2010) Design and implementation of a prototype system of ocean sensor web. In: Proceedings of the IET international conference on wireless sensor network (IET-WSN'10). p. 21–6.

Jirka, S. and Stasch, C. (2018) Preface to the special issue, Geospatial Sensor Web-Concepts, Technologies and Applications.

Johnson, M., Healy, M., van de Ven, P., Hayes, M.J., Nelson, J., Newe, T. & Lewis, E. (2009) A comparative review of wireless sensor network mote technologies, Sensors, 2009 IEEE, pp. 1439.

Joint Research Council (2017) Summary Report on Status of Implementation of the INSPIRE Directive in EU. European Union, doi: 10.27.60/143502

Käbisch, S., Peintner, D. and Anicic, D. (2015) Standardized and efficient RDF encoding for constrained embedded networks. In European Semantic Web Conference (pp. 437-452). Springer, Cham.

Kalman, R.E. (1960) A new approach to linear filtering and prediction problems. Journal of basic Engineering, 82(1), pp.35-45.

Kalra, D., Anthony, A., O'Connor A., Patterson, D. et al. (2001) Design and Implementation of a Federated Health Record Server. pages 1–13. Medical Records Institute for the Centre for Advancement of Electronic Records Ltd.

Kalra, D., Beale, T. & Heard, S. (2005) The openEHR foundation, Studies in health technology and informatics, vol. 115, pp. 153-173.

Kämpgen, B., Riepl, D. & Klinger, J. (2014) SMART Research using Linked Data-Sharing Research Data for Integrated Water Resources Management in the Lower Jordan Valley. in SePublica.

Kedron, P., Li, W., Fotheringham, S. and Goodchild, M. (2021) Reproducibility and replicability: opportunities and challenges for geospatial research. International Journal of Geographical Information Science, 35(3), pp.427-445.

Khalsa, S.J.S. (2020) Developing Standards for Earth Observation Data Products. In IOP Conference Series: Earth and Environmental Science (Vol. 509, No. 1, p. 012030). IOP Publishing.

Kilic, O. Bicer, V. and Dogac, A. (2005) Mapping archetypes to OWL, Middle East Tech.Univ., Ankara, Turkey, Tech.Rep.TR-2005-3, [online]. Available at Http://www.Srdc.Metu.Edu.tr/webpage/publications/2005/MappingArchetypestoOWLTechnical.Pdf, (Accessed: 11 April 2017).

Ko, J., Eriksson, J., Tsiftes, N., Dawson-Haggerty, S., Terzis, A., Dunkels, A. & Culler, D. (2011) Contikirpl and tinyrpl: Happy together, Workshop on Extending the Internet to Low Power and Lossy Networks (IP SN).

Kraak, M.J., Sliwinski, A. and Wytzisk, A. (2005) What happens at 52N? An Open source approach to education and research. In Proceedings of the Joint ICA Commissions Seminar, Internet-Based Cartographic Teaching and Learning: Atlases, Map Use, and Visual Analytics.

Klien, E., Annoni, A. & Marchetti, P.G. (2009) The GIGAS project–an action in support to GEOSS, INSPIRE, and GMES.

Klien, E., Lutz, M. and Kuhn, W. (2006) Ontology-based discovery of geographic information services—An application in disaster management. Computers, environment and urban systems, 30(1), pp.102-123.

Klyne, G. & Carroll, J.J. (2006) Resource description framework (RDF): Concepts and abstract syntax.

Kotsev, A., Schade, S., Craglia, M., Gerboles, M., Spinelle, L. and Signorini, M. (2016) Next generation air quality platform: Openness and interoperability for the internet of things. Sensors, 16(3), p.403.

Kovatsch, M., Lanter, M. and Shelby, Z. (2014) Californium: Scalable cloud services for the internet of things with coap. In 2014 International Conference on the Internet of Things (IOT) (pp. 1-6). IEEE.

Kuhn, W. (2005) Introduction to spatial data infrastructures. Presentation held on March 14th, p.2005.

Larizgoitia, I., Muguira, L. & Vazquez, J.I. (2010) Architecture for WSN nodes integration in context aware systems using semantic messages. In Ad Hoc Networks Springer, pp. 731-746.

Lawrence, B.N., Lowry, R., Miller, P., Snaith, H. & Woolf, A. (2009) Information in environmental data grids, Philosophical transactions. Series A, Mathematical, physical, and engineering sciences, vol. 367, no. 1890, pp. 1003-1014.

Leach, P., Mealling, M. and Salz, R., (2005) RFC 4122: A universally unique identifier (UUID) URN namespace. Proposed Standard.

Leadbetter, A. and Fredericks, J. (2014) We have "born digital" - now what about "born semantic"? European Geophysical Union General Assembly.

Leadbetter A. et al. (2016) Linked Ocean Data 2.0, Oceanographic and Marine Cross-Domain Data Management for Sustainable Development, pp. 69.

Leadbetter, A., Meaney, W., Tray, E. et al. (2020) A Modular Approach to Cataloguing Marine Science Data. Earth Sci Inform. https://doi.org/10.1007/s12145-020-00445-w

Leadbetter, A., Lowry, R. and Clements, D.O. (2012) The NERC Vocabulary Server: Version 2.0. In Geophysical Research Abstracts (Vol. 14).

Leadbetter, A., Smyth, D., Fuller, R., O'Grady, E. and Shepherd, A. (2016) Where Big Data meets Linked Data: Applying standard data models to environmental data streams, 2016 IEEE International Conference on Big Data.

Leadbetter, A.M., Shepherd, A., Arko, R., Chandler, C., Chen, Y., Dockery, N., Ferreira, R., Fu, L., Thomas, R., West, P. and Zednik, S. (2016) Experiences of a "semantics smackdown". Earth Science Informatics, 9(3), pp.355-363.

Leadbetter, A., and Vodden, P.N. (2016) Semantic linking of complex properties, monitoring processes and facilities in web-based representations of the environment. International Journal of Digital Earth, 9(3), pp.300-324.

Le-Phuoc, D., Quoc, H.N.M., Quoc, H.N., Nhat, T.T. and Hauswirth, M. (2016) The

Graph of Things: A step towards the Live Knowledge Graph of connected things. Journal of Web Semantics, 37, pp.25-35.

Le-Tuan, A., Hayes, C., Wylot, M. and Le-Phuoc, D. (2018) RDF4Led: an RDF engine for lightweight edge devices. In Proceedings of the 8th International Conference on the Internet of Things (p. 2). ACM.

Le-Tuan, A., Hayes, C., Hauswirth, M. and Le-Phuoc, D. (2020) Pushing the Scalability of RDF Engines on IoT Edge Devices. Sensors, 20(10), p.2788.

Leslie, H. (2008) International developments in openEHR archetypes and templates. Health Information Management Journal, 37(1), pp.38-39.

Levis, P., Madden, S., Polastre, J., Szewczyk, R., Whitehouse, K., Woo, A., Gay, D., Hill, J., Welsh, M. & Brewer, E. (2005) TinyOS: An operating system for sensor networks" in Ambient intelligence Springer, pp. 115-148.

Levis, P. (2012) Experiences from a Decade of TinyOS Development., OSDI, pp. 207.

Lezcano Matías, L. (2012) Combining ontologies and rules with clinical archetypes (Doctoral dissertation, Universidad de Alcalá).

Lezcano, L., Santos, L. & García-Barriocanal, E. (2013) Semantic Integration of Sensor Data and Disaster Management Systems: The Emergency Archetype Approach, International Journal of Distributed Sensor Networks, vol. 2013.

Lezcano, L. Sicilia, M. and Rodríguez-Solano, C. (2011) Integrating reasoning and clinical archetypes using OWL ontologies and SWRL rules, J. Biomed. Inform., vol. 44, pp. 343-353.

Liang, S., Huang, C.Y. and Khalafbeigi, T. (2016) OGC SensorThings API Part 1: Sensing, Version 1.0.

Livingstone, D.N. (1988) Science, magic and religion: a contextual reassessment of geography in the sixteenth and seventeenth centuries. History of science, 26(3), pp.269-294.

Lopez, P.L., Wanders, N., Schellekens, J., Renzullo, L.J., Sutanudjaja, E.H. and Bierkens, M.F. (2016) Improved large-scale hydrological modelling through the assimilation of streamflow and downscaled satellite soil moisture observations. Hydrology and Earth System Sciences, 20(7), pp.3059-3076.

Los, F.J., Villars, M.T. and Van der Tol, M.W.M. (2008) A 3-dimensional primary production model (BLOOM/GEM) and its applications to the (southern) North Sea

(coupled physical–chemical–ecological model). Journal of Marine Systems, 74(1-2), pp.259-294.

Loseto, G., Ieva, S., Gramegna, F., Ruta, M., Scioscia, F. and Di Sciascio, E. (2016) October. Linked Data (in low-resource) Platforms: a mapping for Constrained Application Protocol. In International Semantic Web Conference (pp. 131-139). Springer, Cham.

Ludovici, A., Moreno, P. & Calveras, A. (2013) TinyCoAP: a novel constrained application protocol (CoAP) implementation for embedding RESTful web services in wireless sensor networks based on TinyOS, Journal of Sensor and Actuator Networks, vol. 2, no. 2, pp. 288-315.

Mackenzie, F.T. & Mackenzie, J.A. (2010) Our changing planet: an introduction to earth system science and global environmental change, 4th edn, Prentice Hall New Jersey.

Maldonado, J.A., Moner, D., Boscá, D., Fernández-Breis, J.T., Angulo, C. & Robles, M. (2009) LinkEHR-Ed: A multi-reference model archetype editor based on formal semantics, International journal of medical informatics, vol. 78, no. 8, pp. 559-570.

Malone, T.C. and Nowlin, W.D. (2006) GOOS regional alliances and the development of the coastal module of GOOS. In 2006 IEEE US/EU Baltic International Symposium (pp. 1-16). IEEE.

Markensteijn, G.C. (2017) Performing data assimilation experiments with hydrodynamic models: A Java Sea case.

Martinez-costa, C. Menarguez-tortosa, M. and Fernandez-Breis, J.T. (2010) An approach for the semantic interoperability of ISO EN 13606 and OpenEHR archetypes. Journal of Biomedical Informatics, 43(5), pp. 736-746.

Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, A. and Schneider, L. (2003) Dolce: a descriptive ontology for linguistic and cognitive engineering. WonderWeb Project, Deliverable D17 v2, 1, pp.75-105.

Masó, J. and Serral, I., Pons X. (2011) GEOVIQUA: a FP7 scientific project to promote spatial data quality usability: metadata, search and visualization. In Proceedings 7th International Symposium on Spatial Data Quality, October 12-14 2011, Coimbra, Portugal.

Mayewski, P.A., Rohling, E.E., Stager, J.C., Karlén, W., Maasch, K.A., Meeker, L.D., Meyerson, E.A., Gasse, F., van Kreveld, S., Holmgren, K. and Lee-Thorp, J. (2004) Holocene climate variability. Quaternary research, 62(3), pp.243-255.

McGuinness, D.L. & Van Harmelen, F. (2004) OWL web ontology language overview, W3C recommendation, vol. 10, no. 10, pp. 2004.

Michelsen, L., Pedersen, S.S., Tilma, H.B. and Andersen, S.K. (2005) Comparing different approaches to two-level modelling of electronic health records. Studies in health technology and informatics, 116, pp.113-118.

Millard, K., Smits, P., Abramic, A., Calewaert, J. B., Shepherd, I. (2015) Marine Pilot - EMODnet and INSPIRE: Benefits of Closer Collaboration and a Framework for Action.

Milsum, J.H. (1968) The technosphere, the biosphere, the sociosphere Their systems modelling and optimization. IEEE Spectrum, 5(6), pp.76-82.

Min, L., Tian, Q., Lu, X. and Duan, H. (2018) Modeling EHR with the openEHR approach: an exploratory study in China. BMC medical informatics and decision making, 18(1), pp.1-15.

Moner, D., Maldonado, J.A. and Robles, M. (2018) Archetype modeling methodology. Journal of biomedical informatics, 79, pp.71-81.

Morawska, L., Thai, P.K., Liu, X., Asumadu-Sakyi, A., Ayoko, G., Bartonova, A., Bedini, A., Chai, F., Christensen, B., Dunbabin, M. and Gao, J. (2018) Applications of low-cost sensing technologies for air quality monitoring and exposure assessment: How far have they gone?. Environment international, 116, pp.286-299.

Muñoz, P., Trigo, J.D., Martínez, I., Muñoz, A., Escayola, J. and García, J. (2011) The ISO/EN 13606 standard for the interoperable exchange of electronic health records. Journal of Healthcare Engineering, 2.

NASA (2019) NASA CF Conventions, [online]. Available at https://earthdata.nasa.gov/user-resources/standards-and-references/climate-and-forecast-cf-metadata-conventions. (Accessed: 02-Jul-2018).

Nativi, S. and Bemmelen, J. (2016) GEO DAB and GEOSS Portal, in AGU Abstract Fall Meeting, 2016.

Nittel, S. (2009) A survey of geosensor networks: Advances in dynamic environmental monitoring. Sensors, 9(7), pp.5664-5678.

Naur, P (ed.) (1960) Revised Report on the Algorithmic Language ALGOL 60., Communications of the ACM, Vol. 3 No.5, pp. 299-314, May.

Noy, N.F. and McGuinness, D.L. (2001) Ontology development 101: A guide to creating your first ontology. Stanford knowledge systems laboratory technical report KSL-01-05, [online]. Available at http://www.ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness.pdf.

Novellino, A., D'Angelo, P., Benedetti, G., Manzella, G., Gorringe, P., Schaap, D., Pouliquen, S. and Rickards, L. (2015) European marine observation data network—EMODnet physics. In OCEANS 2015-Genova (pp. 1-6). IEEE.

Nurseitov, N., Paulson, M., Reynolds, R. and Izurieta, C. (2009) Comparison of JSON and XML data interchange formats: a case study. Caine, 9, pp.157-162.

Obrst, L. (2003) Ontologies for semantically interoperable systems. In Proceedings of the twelfth international conference on Information and knowledge management (pp. 366-369).

OceanSites (2015) OceanSITES Data Format Reference Manual, [online]. Available at http://www.oceansites.org/docs/oceansites_data_format_reference_manual.pdf. (Accessed: 02-Jul-2018).

ODIP (2018) Marine Profiles of the OGC Sensor Web Enablement Standards, [online]. Available at https://odip.github.io/MarineProfilesForSWE/. (Accessed: 02-Jul-2018).

Ogden, C.K. and Richards, I.A. (1923) The meaning of meaning: A study of the influence of thought and of the science of symbolism.

O'Hare, G.M., Diamond, D., Lau, K.T., Hayes, J., Muldoon, C., O'Grady, M.J., Tynan, R., Rancourt, G., Kolar, H.R. & McCarthy, R.J. (2009) The adaptive environment: Delivering the vision of in situ real-time environmental monitoring, IBM Journal of Research and Development, vol. 53, no. 3, pp. 2: 1-2: 11.

Olenin, S. et al. (2010) Marine strategy framework directive, Task Group, vol. 2, 2010.

Oniki, T.A., Coyle, J.F., Parker, C.G. and Huff, S.M. (2014) Lessons learned in detailed clinical modeling at Intermountain Healthcare. Journal of the American Medical Informatics Association, 21(6), pp.1076-1081.

Orsini, G., Bade, D. and Lamersdorf, W. (2015) Computing at the mobile edge: Designing elastic android applications for computation offloading. In 2015 8th IFIP Wireless and Mobile Networking Conference (WMNC) (pp. 112-119). IEEE.

Osen, L. O., Wang, H., Hjelmervik, K. B., Schøyen, H. (2017) Organizing Data from Industrial Internet of Things for Maritime Operations. In Oceans' 17 MTS/IEEE Conference.

Ott, W.R. (1978) Environmental indices: theory and practice.

Parkinson, C.L., Ward, A. and King, M.D. (2006) Earth science reference handbook: a guide to NASA's earth science program and earth observing satellite missions. National Aeronautics and Space Administration, p.277.

Pastres, R., Ciavatta, S. and Solidoro, C. (2003) The Extended Kalman Filter (EKF) as a tool for the assimilation of high frequency water quality data. Ecological modelling, 170(2-3), pp.227-235.

Pearlman, J.S., Bushnell, M., Coppola, L., Buttigieg, P.L., Pearlman, F., Simpson, P., Barbier, M., Karstensen, J., Muller-Karger, F.E., Munoz-Mas, C. and Pissierssens, P. (2019) Evolving and sustaining ocean best practices and standards for the next decade. Frontiers in Marine Science, 6, p.277.

Pease, A., Niles, I. and Li, J. (2002) The suggested upper merged ontology: A large ontology for the semantic web and its applications. In Working notes of the AAAI-2002 workshop on ontologies and the semantic web (Vol. 28, pp. 7-10).

Peirce, C.S. (1935) Collected papers. Vol. VI. Scientific metaphysics (Ed. by C. Hartshorne and P. Weiss.).

Pfister, R., Ullman, R., Wichmann, K. & Perkins, D.C. (2001) ECHO Responds to NASA's Earth Science User Community.

Pham, D.T., Verron, J. and Roubaud, M.C. (1998) A singular evolutive extended Kalman filter for data assimilation in oceanography. Journal of Marine systems, 16(3-4), pp.323-340.

Pickles, J. (1995) Ground truth: The social implications of geographic information systems, Guilford Press.

Pitman, A. (2005) On the role of geography in earth system science, Geoforum, vol. 36, no. 2, pp. 137-148.

Polanyi, M. (1941) The growth of thought in society," Economica, vol. 8, pp. 428-456.

Pontin, B. (2020) Fourth assessment report of the intergovernmental panel on climate change: Key developments in the science of global warming. Environmental Law and Management, 18(6).

Popper, K.R. (1948) Objective Knowledge: An Evolutionary Approach. Ox-43.

Popper, K. (1978) Three worlds: The Tanner lecture on human values: Delivered at the University of Michigan. The Tanner Lectures, Humanities Center, University of Utah, [online]. Available at https://tannerlectures.utah.edu/_documents/a-to-z/p/popper80.pdf (Accessed: May 28th 2019).

Porn, A.M., Peres, L.M. and Del Fabro, M.D. (2015) A Process for the Representation of openEHR ADL Archetypes in OWL Ontologies. In MedInfo (pp. 827-831).

Pouliquen, S. (2011) Recommendations for in situ data Real Time Quality Control.

Pottie, G. (2003) Casting the Wireless Sensor Net, MIT Technology Review, July 2003.

Pouliquen, S. (2011) Recommendations for in-situ data Real Time Quality Control.

Probst, F., Gordon, A. & Dornelas, I. (2006) OGC Discussion Paper: Ontology-based Representation of the OGC Observations and Measurements Model, Institute for Geoinformatics (IFGI).

Pruteanu, A., Iyer, V. and Dulman, S. (2011) Churndetect: a gossip-based churn estimator for large-scale dynamic networks. In European Conference on Parallel Processing (pp. 289-301). Springer, Berlin, Heidelberg.

Queiroz, D.V., Gomes, R.D. and Benavente-Peces, C. (2017) Performance Evaluation of Default Active Message Layer (AM) and TKN15. 4 Protocol Stack in TinyOS 2.1.2. In SENSORNETS (pp. 69-79).

Raskin, R.G. & Pan, M.J. (2005) Knowledge representation in the semantic web for Earth and environmental terminology (SWEET), Computers & Geosciences, vol. 31, no. 9, pp. 1119-1125.

Rea, S., Pathak, J., Savova, G., Oniki, T.A., et al. (2012) Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: the SHARPn project. Journal of biomedical informatics, 45(4), pp.763-771.

Rector, A.L. (1999) Clinical terminology: why is it so hard? Methods of information in medicine, 38(04/05), pp.239-252.

Retscher, C., De Mazière, M., Meijer, Y., Vik, A.F., Boyd, I., Niemeijer, S., Koopman, R.M., Bojkov, B. (2011) The Generic Earth Observation Metadata Standard (GEOMS).

Reusing, T. (2012) Comparison of operating systems tinyos and contiki, Sens.Nodes-Operation, Netw.Appli.(SN), vol. 7.

Rew, R. and Davis, G. (1990) NetCDF: an interface for scientific data access. IEEE computer graphics and applications, 10(4), pp.76-82.

Reyes, S. (2014) owl2json library, [online]. Available at http://github.com/stain/owl2jsonld/. (Accessed:12th October 2017).

Ridler, M.E., van Velzen, N., Hummel, S., Sandholt, I., Falk, A.K., Heemink, A. and Madsen, H. (2014) Data assimilation framework: Linking an open data assimilation library (OpenDA) to a widely adopted model interface (OpenMI). Environmental modelling & software, 57, pp.76-89.

Riepl, D. (2014) Knowledge-based decision support for integrated water resources management with an application for Wadi Shueib, Jordan. KIT Scientific Publishing.

Rocher, G. and Brown, J. (2009) GORM. In The Definitive Guide to Grails (pp. 249-288). Apress.

Robinson, I., Webber, J. and Eifrem, E. (2013) Graph databases. O'Reilly Media, Inc.

Rowley, J. (2007) The wisdom hierarchy: representations of the DIKW hierarchy. Journal of information science, 33(2), pp.163-180.

Rueda, C., Galbraith, N., Morris, R.A., Bermudez, L.E., Arko, R.A. & Graybeal, J. (2010) The MMI Device Ontology: Enabling Sensor Integration, AGU Fall Meeting Abstracts, pp. 08.

Sargent, P. (1999) Feature Identities, Descriptors, and Handles. In International Conference on Interoperating Geographic Information Systems (pp. 41-53). Springer, Berlin, Heidelberg.

Shvets, A. (2018) Dive Into Design Patterns, ebook, [online] Available at https://refactoring.guru/design-patterns/composite. (Accessed: September 2020).

Seymour, T., Frantsvog, D. and Kumar, S. (2011) History of search engines. International Journal of Management & Information Systems (IJMIS), 15(4), pp.47-58.

Schaap, D.M. and Lowry, R.K. (2010) SeaDataNet–Pan-European infrastructure for marine and ocean data management: unified access to distributed datasets. International Journal of Digital Earth, 3(S1), pp.50-69.

Schade, S., Granell, C., Vancauwenberghe, G., Keßler, C., Vandenbroucke, D., Masser, I. and Gould, M. (2020) Geospatial information infrastructures. In Manual of Digital Earth (pp. 161-190). Springer, Singapore.

Schalles, J.F., Gitelson, A.A., Yacobi, Y.Z. and Kroenke, A.E. (1998) Estimation of chlorophyll a from time series measurements of high spectral resolution reflectance in an eutrophic lake. Journal of Phycology, 34(2), pp.383-390.

Schuurman, N. (2000) Trouble in the heartland: GIS and its critics in the 1990s, Progress in Human Geography, vol. 24, no. 4, pp. 569-590.

Sharma, D.K., Solbrig, H.R., Tao, C., Weng, C., Chute, C.G. and Jiang, G. (2017) Building a semantic web-based metadata repository for facilitating detailed clinical modeling in cancer genome studies. Journal of biomedical semantics, 8(1), p.19.

Shelby, Z. and Bormann, C. (2011) 6LoWPAN: The wireless embedded Internet (Vol. 43). John Wiley & Sons.

Shelby, Z. et al. (2012) Constrained Application Protocol (CoAP), draft-ietf-core-coap-13, Orlando: The Internet Engineering Task Force–IETF, Dec.

Shelby, Z. (2012) Constrained RESTful environments (CoRE) link format.

Sheth, A.P. and Larson, J.A. (1990) Federated database systems for managing distributed, heterogeneous, and autonomous databases. ACM Computing Surveys (CSUR), 22(3), pp.183-236.

Rahimifard, S., and Trollman, H. (2018) UN Sustainable Development Goals: an engineering perspective, International Journal of Sustainable Engineering, 11:1, 1-3, DOI:10.1080/19397038.2018.1434985

Sherwood, R., Chien, S. (2007) Sensor Web Technologies: A New Paradigm for Operations. In Proceedings of the 7th International Symposium on Reducing the Cost of Spacecraft Ground Systems and Operations (The Moscow, The Russia, June 11-15, 2007). RCSGSO 2007, AIAA.

Sheth, A.P. and Larson, J.A. (1990) Federated database systems for managing distributed, heterogeneous, and autonomous databases. ACM Computing Surveys (CSUR), 22(3), pp.183-236.

Showstack, R. (2014) Sentinel satellites initiate new era in earth observation. Eos, Transactions American Geophysical Union, 95(26), pp.239-240.

Schreiner, A.T. (1993) Object oriented programming with ANSI-C.

Simonis, I. (2019) OGC Standardization: From Early Ideas to Adopted Standards. In IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium (pp. 4511-4514). IEEE.

Singleton, A. and Arribas-Bel, D. (2019). Geographic data science. Geographical Analysis.

Sinton, D. (1978) The inherent structure of information as a constraint to analysis: Mapped thematic data as a case study, Harvard papers on geographic information systems, vol. 6, pp. 1-17.

Sirin, E., Parsia, B., Grau, B.C., Kalyanpur, A. & Katz, Y. (2007) Pellet: A practical owl-dl reasoner, Web Semantics: science, services and agents on the World Wide Web, vol. 5, no. 2, pp. 51-53.

Smith, B., Kumar, A. and Bittner, T. (2005) Basic formal ontology for bioinformatics.

Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K. B., Tignor, M., Miller, H. L. (eds) (2007) Climate Change : The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge.

Staab, S., Angele, J., Decker, S., Erdmann, M., Hotho, A., Maedche, A., Schnurr, H.P., Studer, R. and Sure, Y. (2000) AI for the web-ontology-based community web portals. In AAAI/IAAI (pp. 1034-1039).

Stacey, P. and Berry, D. (2019a) Extending Two-Level Information Modelling to the Internet of Things, In Proc. of IEEE World Forum on IoT Limerick, Ireland, 2019.

Stacey, P. and Berry, D. (2019b) Beyond Standards Compliant Internet of Things Data-streams. Tutorial presented at the IEEE World Forum on IoT, Limerick, Ireland, 2019.

Stacey, P., Berry, D. (2015) Applying Two-Level Modelling to Remote Sensor Systems Design to Enable Future Knowledge Generation. In IEEE YP Conference in Remote Sensing, Barcelona, Abstracts.

Star, S.L. and Griesemer, J.R. (1989) Institutional ecology, translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. Social studies of science, 19(3), pp.387-420.

Stirbu, V. (2008) Towards a restful plug and play experience in the web of things, Semantic computing, IEEE international conference onIEEE, pp. 512.

Stoppani, A. (1873) Corso di geologia del professore Antonio Stoppani: Geologia stratigrafica (Vol. 2). G. Bernardoni e G. Brigola.

Strachey, C. (1966) Towards a formal semantics, in Formal Lan9uage Description Languages, r. B. Steel, ed., North Holland, pp. 198-220.

Sun, K., Zhu, Y., Pan, P., Hou, Z., Wang, D., Li, W. and Song, J. (2019) Geospatial data ontology: the semantic foundation of geospatial data integration and sharing, Big Earth Data, 3:3, 269-296, DOI: 10.1080/20964471.2019.1661662

Sundvall, E., Qamar, R., Nyström, M., Forss, M., Petersson, H., Karlsson, D., Åhlfeldt, H. & Rector, A. (2008) Integration of tools for binding archetypes to SNOMED CT, BMC Medical Informatics and Decision Making, vol. 8, no. 1, pp. 1.

Sundvall, E. et al. (2013) Applying representational state transfer (REST) architecture to archetype-based electronic health record systems, BMC Med. Inform. Decis. Mak., vol. 13, pp. 57-6947-13-57, May 9.

Tavra, M., Jajac, N. and Cetl, V. (2017) Marine Spatial Data Infrastructure Development Framework: Croatia Case Study. ISPRS International Journal of Geo-Information, 6(4), p.117.

Taylor, P.J. (1990) Editorial comment GKS, Political Geography Quarterly, vol. 9, no. 3, pp. 211-212.

Taylor, K., Haller, A., Lefrançois, M., Cox, S.J., Janowicz, K., García-Castro, R., Le Phuoc, D., Lieberman, J., Atkinson, R. and Stadler, C. (2019) The Semantic Sensor Network Ontology, Revamped. In JT@ ISWC.

Thatcher, J., Bergmann, L., Ricker, B., Rose-Redwood, R., O'Sullivan, D., Barnes, T.J., Barnesmoore, L.R., Beltz Imaoka, L., Burns, R., Cinnamon, J. and Dalton, C.M. (2016). Revisiting critical GIS. Environment and Planning A, 48(5), pp.815-824.

Thierauf, R.J. (1999) Knowledge management systems for business. Greenwood Publishing Group.

TI (2017) FRAM – Ultra-Low-Power Embedded Memory, Texas Instruments, [online]. Available at http://www.ti.com/lsds/ti/microcontrollers-16-bit-32-bit/msp/ultra-low-power/msp430frxx-fram/what-is-fram.page. (Accessed: 12th October 2017).

TinyOS Alliance (2017) Tiny-OS main development branch, [online]. Available at

https://github.com/tinyos/tinyos-main. (Accessed: 12th October 2017).

Tsiftes, N. and Dunkels, A. (2011) A database in every sensor. In Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems (pp. 316-332). ACM.

UML, O. (2001). Unified modeling language. Object Management Group.

UNESCO (2018) IOC Oceans | United Nations Educational, Scientific and Cultural Organization, UNESCO, [online]. Available at http://www.unesco.org/new/en/natural-sciences/ioc-oceans. (Accessed: 02-Jul-2018).

Usui, N., Ishizaki, S., Fujii, Y., Tsujino, H., Yasuda, T. and Kamachi, M. (2006) Meteorological Research Institute multivariate ocean variational estimation (MOVE) system: Some early results. Advances in Space Research, 37(4), pp.806-822.

Van Leeuwen, P.J., Vetra-Carvalho, S., Nerger, L., Heemink, A., van Velzen, N., Verlaan, M., Altaf, M.U., Beckers, J.M., Barth, A., Brasseur, P. and Brankart, J.M. (2011) SANGOMA: Stochastic Assimilation for the Next Generation Ocean Model Applications EU FP7 SPACE-2011-1 project 283580.

Verdone, R., Dardari, D., Mazzini, G. & Conti, A. (2010) Wireless sensor and actuator networks: technologies, analysis and design, Academic Press.

Verlaan, M., van Velzen, N., Hummel, S. and Gerritsen, H. (2010) OpenDA, a generic toolbox for data-assimilation in numerical modelling. In 15th Biennial Conference of the Joint Numerical Sea Modelling Group, Delft, The Netherlands.

Von Schuckmann, K., Le Traon, P.Y., Alvarez-Fanjul, E., Axell, L., Balmaseda, M., Breivik, L.A., Brewin, R.J., Bricaud, C., Drevillon, M., Drillet, Y. and Dubois, C. (2016). The Copernicus marine environment monitoring service ocean state report. Journal of Operational Oceanography, 9(sup2), pp.s235-s320.

Verrier, S., Le Traon, P.Y. and Remy, E. (2017) Assessing the impact of multiple altimeter missions and Argo in a global eddy-permitting data assimilation system. Ocean Science, 13(6), p.1077.

W3C World Wide Web Consortium (2017a) Web of Things (WoT) Current Practices, Unofficial Draft, [online]. Available at https://www.w3.org/WoT/IG/wiki/Thing_Description. (Accessed: 11 April 2017).

W3C World Wide Web Consortium (2017b) Spatial Data on the Web Best Practices,

Working Group Note 28 September 2017, [online]. Available at https://www.w3.org/TR/sdw-bp/ (Accessed: 19 May 2018).

Wang, X. and Chen, W. (2020) Knowledge Graph Data Management: Models, Methods, and Systems. In International Conference on Web Information Systems Engineering (pp. 3-12). Springer, Singapore.

Wang, L., Min, L., Wang, R., Lu, X. and Duan, H. (2015) Archetype relational mapping-a practical openEHR persistence solution. BMC medical informatics and decision making, 15(1), p.88.

Wanner, H., Beer, J., Bütikofer, J., Crowley, T.J., Cubasch, U., Flückiger, J., Goosse, H., Grosjean, M., Joos, F., Kaplan, J.O. and Küttel, M. (2008) Mid-to Late Holocene climate change: an overview. Quaternary Science Reviews, 27(19), pp.1791-1828.

Watts, N., Amann, M., Arnell, N., Ayeb-Karlsson, S., Belesova, K., Boykoff, M., Byass, P., Cai, W., Campbell-Lendrum, D., Capstick, S. and Chambers, J. (2019) The 2019 report of The Lancet Countdown on health and climate change: ensuring that the health of a child born today is not defined by a changing climate. The Lancet, 394(10211), pp.1836-1878.

Wehde, H. et al.,(2016) CMEMS Quality Information Document. Available: http://marine.copernicus.eu/documents/QUID/CMEMS-INS-QUID013-030-036.pdf.

Wiegand, J. (2004) Eclipse: A platform for integrating development tools. IBM Systems Journal, 43(2), pp.371-383.

Weiser, M., (1991) The Computer for the 21 st Century. Scientific american, 265(3), pp.94-105.

Westerbeeke, H., Secretariaat, G. & Geneve, Z. (2006) Group on Earth Observations (GEO), RUIMTEVAART, vol. 55, no. 4, pp. 10.

Williams, M.A.J., Dunkerley, D.L., De Deckker, P., Kershaw, A.P. and Stokes, T.J. (1997) Quaternary environments. Science Press.

Williams, R.S., Jr., and Ferrigno, J.G. (eds.) (2012) Plate Figure 4 in State of the Earth's cryosphere at the beginning of the 21st century–Glaciers, global snow cover, floating ice, and permafrost and periglacial environments: U.S. Geological Survey Professional Paper 1386–A.

Wollersheim D., Sari A., Rahayu W. (2009) Archetype-Based Electronic Health

Records: A Literature Review and Evaluation of Their Applicability to Health Data Interoperability and Access. Health Information Management Journal, Vol 38, Issue 2, pp. 7 - 17.

Wolf, L., & Hötzl, H. (2011) SMART-IWRM: Integrated water resources management in the lower Jordan Rift Vaalley: project report phase 1. Karlsruhe: KIT Scientific Publishing

Wölger, S., Siorpaes, K., Bürger, T., Simperl, E., Thaler, S. and Hofer, C. (2011) A survey on data interlinking methods. Technical report, Semantic Technology Institute, Innsbruck.

Woo, M., Neider, J., Davis, T. and Shreiner, D. (1999) OpenGL programming guide: the official guide to learning OpenGL, version 1.2. Addison-Wesley Longman Publishing Co., Inc..

World Wide Web Consortium (2014) JSON-LD 1.0: a JSON-based serialization for linked data.

Woods, J., et al. (1996) The strategy for EuroGOOS, EuroGOOS Publication,vol. 1, 1996.

Woolf, A., Lawrence, B., Lowry, R., van Dam, K.K., Cramer, R., Gutierrez, M., Kondapalli, S., Latham, S., O'Neill, K. and Stephens, A. (2005) Climate Science Modelling Language: standards-based markup for metocean data. In Proceedings of 85th meeting of American Meteorological Society.

Xu, G., Shi, Y., Sun, X. and Shen, W. (2019) Internet of things in marine environment monitoring: A review. Sensors, 19(7), p.1711.

Yuan, M. Mark, D M. Egenhofer, M J. and Peuquet, D J. (2005) A Research Agenda for Geographic Information Science, Extensions to geographic representation. In McMaster R B and Usery E I (eds) Boca Raton, FL, CRC Press: 129–56.

Yin, Y., Alves, O. and Oke, P.R. (2011) An ensemble ocean data assimilation system for seasonal prediction. Monthly Weather Review, 139(3), pp.786-808.

Zagzebski, L. (2017) What is knowledge? The Blackwell guide to epistemology, pp.92-116.

Zalasiewicz, J., Williams, M., Smith, A., Barry, T.L., Coe, A.L., Bown, P.R., Brenchley, P., Cantrill, D., Gale, A., Gibbard, P. and Gregory, F.J. (2008) Are we now living in the Anthropocene? Gsa Today, 18(2), pp.4-8.

Zhao, Z. (2020) Towards Interoperable Research Infrastructures for Environmental and Earth Sciences: A Reference Model Guided Approach for Common Challenges. Springer Nature.

Zemmouchi-Ghomari, L. and Ghomari, A.R. (2012) Ontology versus terminology, from the perspective of ontologists. IJWS, 1(4), pp.315-331.

Zender, C.S., Vicente, P. and Wang, W. (2012) NCO: Simpler and faster model evaluation by NASA satellite data via unified file-level netCDF and HDF-EOS data post-processing tools. In AGU Fall Meeting Abstracts.

ZigBee Alliance (2006) Specification, Z. ZigBee Document 053474r06, Version, 1.

Zárate, M., Rosales, P., Braun, G., Lewis, M., Fillottrani, P.R. and Delrieux, C. (2019) OceanGraph: Some Initial Steps Toward a Oceanographic Knowledge Graph. In Iberoamerican Knowledge Graphs and Semantic Web Conference (pp. 33-40). Springer, Cham.

Appendices

Appendix A

XML Schema of Augmented O&M model

```
<?xml version="1.0" encoding="utf-8" ?>
<!-- Augmented OGC Observations & Measurements RM (Reference Model)  XML schema -->
<!-- Authored by TeaPOT July 2018 -->
<!-- Usage: RM for Geo-Spatial O&M -->

<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema" version="v1.0.0"
                targetNamespace="http://tpot.dit.ie"  xmlns="http://tpot.dit.ie">

        <xs:include schemaLocation ="OM-identity_component.xsd" />
        <xs:include schemaLocation ="OM-dataTypes.xsd" />

        <!-- BASED ON GRIM FLEXIBLE IDENTITY_COMPONENT,this is the documentation level-->
        <xs:complexType name="IDENTITY_ABSTRACT" abstract="true">
                <xs:extension name="LOCATABLE">
                <xs:sequence>
                        <xs:element name="name" type="xs:string" />
                        <xs:element name="archetype_id" type="xs:string" minOccurs="0"
                                maxOccurs="1" />
                </xs:sequence>
        </xs:complexType>

        <xs:complexType name="ABSTRACT_OBS">
                <xs:extension name="LOCATABLE">
                <xs:sequence>
                        <xs:element name="name" type="xs:string" />
                        <xs:element name="archetype_id" type="xs:string" minOccurs="0"
                                maxOccurs="1" />
                </xs:sequence>
        </xs:complexType>

        <!-- IDENTITY is a Composition Archetype Class of which we can generate storage
                        level Concepts-->
        <xs:element name="geo_identity" type="Geo_Data_Document" />

        <xs:complexType name="Geo_Data_Document">
                <xs:complexContent>
                        <xs:extension base="IDENTITY_ABSTRACT">
                                <xs:sequence>
                                        <xs:element name="name" type="xs:string" />
                                        <xs:element name="archetype_id" type="xs:string"
                                                minOccurs="0" maxOccurs="1" />
                                        <xs:element name="geoDataComposition"
                                                type="IDENTITY_ABSTRACT" minOccurs="0"
                                                maxOccurs="unbounded" />
                                        <xs:element name="details" type="DETAILS_COMPOSITE"
                                                />
                                </xs:sequence>
                        </xs:extension>
                </xs:complexContent>
        </xs:complexType>

        <xs:complexType name="OBSERVATION_SET">
                <xs:complexContent>
                        <xs:extension base="ABSTRACT_OBS">
                                <xs:sequence>
                                        <xs:element name="details" type="DETAILS_COMPOSITE"
                                                />
                                        <xs:element name="observation" type="ABSTRACT_OBS"
                                                minOccurs="0" maxOccurs="unbounded"/>
                                        <xs:element name="relationship"
                                                type="ObservedProperty" minOccurs="1"
                                                maxOccurs="1" />
                                        <xs:element name="relationship"
                                                type="FeatureOfInterest" minOccurs="1"
                                                maxOccurs="1" />
                                </xs:sequence>
                        </xs:extension>
                </xs:complexContent>
        </xs:complexType>
```

```
<xs:complexType name="OBSERVATION">
	<xs:complexContent>
		<xs:extension base="ABSTRACT_OBS">
			<xs:sequence>
				<xs:element name="details" type="DETAILS_COMPOSITE"
					minOccurs="0" maxOccurs="1"/>
				<xs:element name="featureofinterest"
					type="FeatureOfInterest" minOccurs="1"
					maxOccurs="1" />
				<xs:element name="obsproperty"
					type="ObservedProperty" minOccurs="1"
					maxOccurs="1" />
				<xs:element name="results_cluster" type="ANY_TYPE"
					minOccurs="1" maxOccurs="unbounded"/>
				<xs:element name="resultTime" type="xs:string"/>
				<xs:element name="procedure" type="OM_PROCESS"/>
			</xs:sequence>
		</xs:extension>
	</xs:complexContent>
</xs:complexType>

<xs:complexType name="GeoData_COMPOSITION">
	<xs:complexContent>
		<xs:extension base="IDENTITY_ABSTRACT">
			<xs:sequence>
				<xs:element name="name" type="xs:string" />
				<xs:element name="archetype_id" type="xs:string"
					minOccurs="0" maxOccurs="1" />
				<xs:element name="observationSet"
					type="ABSTRACT_OBS" minOccurs="0"
					maxOccurs="unbounded" />
				<xs:element name="details" type="DETAILS_COMPOSITE"
					/>
			</xs:sequence>
		</xs:extension>
	</xs:complexContent>
</xs:complexType>

<xs:complexType name="NAMED_VALUE" abstract="true">

</xs:complexType>

<xs:complexType name="DETAILS_COMPOSITE">
	<xs:complexContent>
		<xs:extension base="NAMED_VALUE">
			<xs:sequence>
				<xs:element name="element" type="NAMED_VALUE"
					minOccurs="1" maxOccurs="unbounded" />
			</xs:sequence>
		</xs:extension>
	</xs:complexContent>
</xs:complexType>

<xs:complexType name="DETAILS_ELEMENT">
	<xs:complexContent>
		<xs:extension base="NAMED_VALUE">
			<xs:sequence>
				<xs:element name="data_value" type="DATA_VALUE"
					minOccurs="1" maxOccurs="unbounded" />
			</xs:sequence>
		</xs:extension>
	</xs:complexContent>
</xs:complexType>
<xs:complexType name="ObservedProperty">
		<xs:sequence>
			<xs:element name="details" type="NAMED_VALUE" />
		</xs:sequence>
</xs:complexType>

<xs:complexType name="OM_PROCESS">
		<xs:sequence>
			<xs:element name="null" type="NAMED_VALUE" />
		</xs:sequence>
</xs:complexType>
```

```
<xs:complexType name="FeatureOfInterest">
		<xs:sequence>
			<xs:element name="details" type="NAMED_VALUE" />
		</xs:sequence>
</xs:complexType>


<xs:complexType name="TS">
	<xs:complexContent>
		<xs:extension base="DATA_VALUE">
			<xs:sequence>
				<xs:element name="time" type="xs:date" />
			</xs:sequence>
		</xs:extension>
	</xs:complexContent>
</xs:complexType>


<xs:complexType name="OM_STRING">
	<xs:complexContent>
		<xs:extension base="DATA_VALUE">
			<xs:sequence>
				<xs:element name="value" type="xs:string" />
			</xs:sequence>
		</xs:extension>
	</xs:complexContent>
</xs:complexType>

<xs:complexType name="OM_INTEGER">
	<xs:complexContent>
		<xs:extension base="DATA_VALUE">
			<xs:sequence>
				<xs:element name="value" type="xs:integer" />
			</xs:sequence>
		</xs:extension>
	</xs:complexContent>
</xs:complexType>

<xs:complexType name="OM_DECIMAL">
	<xs:complexContent>
		<xs:extension base="DATA_VALUE">
			<xs:sequence>
				<xs:element name="value" type="xs:decimal" />
			</xs:sequence>
		</xs:extension>
	</xs:complexContent>
</xs:complexType>

<xs:complexType name="OM_FLOAT">
	<xs:complexContent>
		<xs:extension base="DATA_VALUE">
			<xs:sequence>
				<xs:element name="value" type="xs:float" />
			</xs:sequence>
		</xs:extension>
	</xs:complexContent>
</xs:complexType>

<xs:complexType name="OM_DOUBLE">
	<xs:complexContent>
		<xs:extension base="DATA_VALUE">
			<xs:sequence>
				<xs:element name="value" type="xs:double" />
			</xs:sequence>
		</xs:extension>
	</xs:complexContent>
</xs:complexType>


<xs:complexType name="DATA_VALUE" abstract="true">
	<xs:sequence>
		<xs:element name="null" type="ANY_TYPE" minOccurs="0"
			maxOccurs="unbounded" />
	</xs:sequence>
</xs:complexType>

<xs:complexType name="ANY_TYPE" abstract="true">
```

```
        </xs:complexType>

        <xs:complexType name="Result">
                <xs:complexContent>
                        <xs:extension base="ANY_TYPE">
                                <xs:sequence>
                                        <xs:element name="data" type="DATA_VALUE"
                                                minOccurs="1" maxOccurs="unbounded" />
                                </xs:sequence>
                        </xs:extension>
                </xs:complexContent>
        </xs:complexType>

        <xs:complexType name="Results">
                        <xs:complexContent>
                                <xs:extension base="ANY_TYPE">
                                        <xs:sequence>
                                                <xs:element name="result_element"
                                                        type="ANY_TYPE" minOccurs="1"
                                                        maxOccurs="unbounded" />
                                        </xs:sequence>
                                </xs:extension>
                </xs:complexContent>
        </xs:complexType>
</xs:schema>
```

Listing A.1 XML Schema of Augmented O&M Model

Appendix B

Geo Archetype Library (DKM)

https://github.com/pstacey/geo-archetype-library

Appendix C

Evaluation 1 AirQuailty Observing Files

SensorThings API Thing Archetype Model

```
archetype (adl_version=1.4)
    TPOT-OM-GeoData_COMPOSITION.Thing.v1
concept
    [at0000]
language
    original_language = <[ISO_639-1::en]>
description
    original_author = <
        ["date"] = <"2019-01-10">
        ["name"] = <"Paul Stacey">
        ["organisation"] = <"TU Dublin">
        ["email"] = <"paul.stacey@tudublin.ie">
    >
    lifecycle_state = <"Draft">
    details = <
        ["en"] = <
            language = <[ISO_639-1::en]>
        >
    >
definition
    GeoData_COMPOSITION[at0000] occurrences matches {1..1} matches {  -- Thing
        details existence matches {1..1} matches {
            DETAILS_COMPOSITE[at0014] occurrences matches {1..1} matches {  --
                                                                DETAILS_COMPOSITE
                element existence matches {1..1} cardinality matches {1..*; unordered;
                                                                unique} matches {
                    DETAILS_ELEMENT[at0018] occurrences matches {1..1} matches {  -- name
                        data_value existence matches {1..1} cardinality matches {1..*;
                                                        unordered; unique} matches {
                            OM_STRING[at0022] occurrences matches {0..*} matches {  --
                                                                OM_STRING
                                value existence matches {1..1} matches {/.*/}
                            }
                        }
                    }
                    DETAILS_ELEMENT[at0019] occurrences matches {1..1} matches {  --
                                                                description
                        data_value existence matches {1..1} cardinality matches {1..*;
                                                        unordered; unique} matches {
                            OM_STRING[at0023] occurrences matches {0..*} matches {  --
                                                                OM_STRING
                                value existence matches {1..1} matches {/.*/}
                            }
                        }
                    }
                    DETAILS_ELEMENT[at0020] occurrences matches {0..1} matches {  --
                                                                properties
                        data_value existence matches {1..1} cardinality matches {1..*;
                                                        unordered; unique} matches {
                            OM_STRING[at0024] occurrences matches {0..*} matches {*}  --
                                                                JSON_Object
                        }
                    }
                    DETAILS_COMPOSITE[at0021] occurrences matches {0..*} matches {  --
                                                                Location
                        element existence matches {1..1} cardinality matches {1..*;
                                                        unordered; unique} matches {
                            DETAILS_ELEMENT[at0001] occurrences matches {1..1} matches {
                                                                -- name
                                data_value existence matches {1..1} cardinality matches
                                                        {1..*; unordered; unique} matches {
                                    OM_STRING[at0004] occurrences matches {0..*} matches
                                                                {  -- OM_STRING
                                        value existence matches {1..1} matches {/.*/}
                                    }
                                }
```

```
                            }
                            DETAILS_ELEMENT[at0002] occurrences matches {0..*} matches
                                                                {*}  -- description
                            DETAILS_ELEMENT[at0003] occurrences matches {1..1} matches {
                                                                    -- location
                                data_value existence matches {1..1} cardinality matches
                                                        {1..*; unordered; unique} matches {
                                    OM_STRING[at0005] occurrences matches {0..*} matches
                                                                    {  -- OM_STRING
                                        value existence matches {1..1} matches {
                                            [ac0001]
                                        }
                                    }
                                }
                            }
                        }
                    }
                }
            }
            observationSet existence matches {0..1} cardinality matches {0..*; unordered;
                                                                unique} matches {
                OBSERVATION_SET[at0010] occurrences matches {1..*} matches {  -- Slot to
                                                                    Datastream
                    observation existence matches {1..1} cardinality matches {0..*; unordered;
                                                                unique} matches {
                        OBSERVATION[at0006] occurrences matches {0..*} matches {  --
                                                                    Observation
                            archetype_id existence matches {0..1} matches {*}
                            details existence matches {0..1} matches {*}
                            featureofinterest existence matches {1..1} matches {*}
                            name existence matches {1..1} matches {*}
                            obsproperty existence matches {1..1} matches {
                                ObservedProperty[at0008] occurrences matches {1..1} matches
                                                                    {*}  -- ObservedProperty
                            }
                            om_process existence matches {1..1} matches {
                                allow_archetype OM_PROCESS[at0009] occurrences matches {1..1}
                                                                matches {  -- Sensor
                                    include
                                        archetype_id/value matches {/TPOT-OM-
                                                                    OM_PROCESS\.Sensor\.v1/}
                                }
                            }
                            procedure existence matches {1..1} matches {*}
                            results_cluster existence matches {1..1} cardinality matches
                                                        {1..*; unordered; unique} matches {
                                Results[at0038] occurrences matches {0..*} matches {  --
                                                                        Results
                                    result_element existence matches {1..1} cardinality
                                                matches {1..*; unordered; unique} matches {
                                        Result[at0039] occurrences matches {0..*} matches {
                                                                    -- Result
                                            data existence matches {1..1} cardinality matches
                                                        {1..*; unordered; unique} matches {
                                                DV_TIME[at0040] occurrences matches {0..*}
                                                                matches {  -- DV_TIME
                                                    value existence matches {1..1} matches
{/([01]\d|2[0-3])([0-5]\d([0-5]\d([,.]\d+)?)?)?(Z|([+\-]((0\d)|(1[0-
2]))(00|30)?))?|([01]\d|2[0-3])(:[0-5]\d(:[0-5]\d([,.]\d+)?)?)?(Z|([+\-]((0\d)|(1[0-
2]))(:(00|30))?))?/}
                                                }
                                            }
                                        }
                                    }
                                }
                            }
                        }
                    }
                }
                details existence matches {1..1} matches {
                    DETAILS_COMPOSITE[at0007] occurrences matches {1..1} matches {  --
                        element existence matches {1..1} cardinality matches {1..*;
                                                        unordered; unique} matches {
                            DETAILS_ELEMENT[at0011] occurrences matches {1..1} matches {
                                                                        -- name
                                data_value existence matches {1..1} cardinality matches
                                                        {1..*; unordered; unique} matches {
```

```
                OM_STRING[at0026] occurrences matches {0..*} matches
                                                        {  -- OM_STRING
                    value existence matches {1..1} matches {"empty"}
                }
            }
        }
        DETAILS_ELEMENT[at0012] occurrences matches {1..1} matches {
                                                         -- description
            data_value existence matches {1..1} cardinality matches
                                    {1..*; unordered; unique} matches {
                OM_STRING[at0027] occurrences matches {0..*} matches
                                                        {  -- OM_STRING
                    value existence matches {1..1} matches {/.*/}
                }
            }
        }
        DETAILS_ELEMENT[at0013] occurrences matches {1..1} matches {
                                                    -- unitOfMeasurement
            data_value existence matches {1..1} cardinality matches
                                     {1..*; unordered; unique} matches {
                OM_STRING[at0028] occurrences matches {0..*} matches
                                                        {  -- OM_STRING
                    value existence matches {1..1} matches {
                        [ac0002]
                    }
                }
            }
        }
        DETAILS_ELEMENT[at0015] occurrences matches {1..1} matches {
                                                  -- observationType
            data_value existence matches {1..1} cardinality matches
                                     {1..*; unordered; unique} matches {
                OM_STRING[at0029] occurrences matches {0..*} matches
                                                        {  -- OM_STRING
                    value existence matches {1..1} matches {
                        [ac0003]
                    }
                }
            }
        }
        DETAILS_ELEMENT[at0016] occurrences matches {0..1} matches {
                                                        -- observedArea
            data_value existence matches {1..1} cardinality matches
                                     {1..*; unordered; unique} matches {
                OM_STRING[at0030] occurrences matches {0..*} matches
                                                        {  -- OM_STRING
                    value existence matches {1..1} matches {/.*/}
                }
            }
        }
        DETAILS_ELEMENT[at0017] occurrences matches {0..1} matches {
                                                    -- phenomenonTime
            data_value existence matches {1..1} cardinality matches
                                    {1..*; unordered; unique} matches {
                DV_TIME[at0031] occurrences matches {0..*} matches {
                                                            -- TM_Time
                    accuracy existence matches {0..1} matches {
                        DV_DURATION[at0033] occurrences matches
                                                {0..1} matches {  -- DV_DURATION
                            value existence matches {1..1} matches
{/P(\d+Y)?(\d+M)?(\d+W)?(\d+D)?(T(\d+H)?(\d+M)?(\d+(\.\d+)?S)?)?/}
                        }
                    }
                    value   existence   matches   {1..1}   matches
{/([01]\d|2[0-3])([0-5]\d([0-5]\d([,.]\d+)?)?)?(Z|([+\-]((0\d)|(1[0-
2]))(00|30)?))?|([01]\d|2[0-3])(:[0-5]\d(:[0-5]\d([,.]\d+)?)?)?(Z|([+\-]((0\d)|(1[0-
2]))(:(00|30))?))?/}
                    magnitude_status existence matches {0..1} matches
                                                                {/.*/}
                    normal_range existence matches {0..1} matches {
                        DV_INTERVAL[at0034] occurrences matches
                                                {0..1} matches {  -- DV_INTERVAL
                            lower existence matches {0..1} matches
                                                                    {*}
                            lower_included existence matches {0..1}
                                                            matches {*}
                            lower_unbounded existence matches {1..1}
```

```
                                                                    matches {*}
                                                upper existence matches {0..1} matches
                                                                                {*}
                                                upper_included existence matches {0..1}
                                                                    matches {*}
                                                upper_unbounded existence matches {1..1}
                                                                    matches {*}
                                            }
                                        }
                                    }
                                }
                            }
                            DETAILS_ELEMENT[at0025] occurrences matches {0..1} matches {
                                                                        -- resultTime
                                data_value existence matches {1..1} cardinality matches
                                                        {1..*; unordered; unique} matches {
                                    DV_TIME[at0032] occurrences matches {0..*} matches {
                                                                        -- TM_Period
                                        value    existence    matches    {1..1}    matches
{/([01]\d|2[0-3])([0-5]\d([0-5]\d([,.]\d+)?)?)?(Z|([+\-]((0\d)|(1[0-
2]))(00|30)?))?|([01]\d|2[0-3])(:[0-5]\d(:[0-5]\d([,.]\d+)?)?)?(Z|([+\-]((0\d)|(1[0-
2]))(:(00|30))?))?/}
                                        accuracy existence matches {0..1} matches {*}
                                        magnitude_status existence matches {0..1} matches
                                                                                {*}
                                        normal_range existence matches {0..1} matches {*}
                                        normal_status existence matches {0..1} matches
                                                                                {*}
                                    }
                                }
                            }
                        }
                    }
                    archetype_id existence matches {0..1} matches {*}
                    name existence matches {1..1} matches {*}
                }
            }
        }

ontology
    term_definitions = <
        ["en"] = <
            items = <
                ["at0000"] = <
                    text = <"Thing">
                    description =
                      <"http://www.opengis.net/spec/iot_sensing/1.0/req/thing">
                >
                ["at0010"] = <
                    text = <"Slot to Datastream">
                    description = <"A Datastream groups a collection of Observations
                      measuring the same ObservedProperty and produced by the same Sensor.
                      http://docs.opengeospatial.org/is/15-078r6/15-078r6.html#28">
                    comment = <"This node was originaly a slot node, it was solved to
                              {TPOT-OM-OBSERVATION_SET.Datastream.v1}">
                >
                ["at0014"] = <
                    text = <"DETAILS_COMPOSITE">
                    description = <"">
                >
                ["at0018"] = <
                    text = <"name">
                    description = <"A property provides a label for Thing entity, commonly
                      a descriptive name.">
                >
                ["at0019"] = <
                    text = <"description">
                    description = <"This is a short description of the corresponding Thing
                      entity.">
                >
                ["at0020"] = <
                    text = <"properties">
                    description = <"A JSON Object containing user-annotated properties as
                      key-value pairs.">
                >
                ["at0021"] = <
```

```
    text = <"Location">
     description = <"The Location entity locates the Thing. Multiple Things
     MAY be located at the same Location. A Thing MAY not have a Location.
     A Thing SHOULD have only one Location.

     However, in some complex use cases, a Thing MAY have more than one
     Location representations. In some case, the Thing MAY have more than
     one Locations. These Locations SHALL have different encodingTypes and
     the encodingTypes SHOULD be in different spaces (e.g., one
     encodingType in Geometrical space and one encodingType in Topological
     space).">
    comment = <"This node was originaly a slot node, it was solved to
       {TPOT-OM-DETAILS_COMPOSITE.Location.v1}">
>
["at0022"] = <
    text = <"OM_STRING">
    description = <"">
>
["at0023"] = <
    text = <"OM_STRING">
    description = <"">
>
["at0024"] = <
    text = <"JSON_Object">
    description = <"">
>
["at0001"] = <
    text = <"name">
    description = <"A property provides a label for Location entity,
       commonly a descriptive name.">
>
["at0002"] = <
    text = <"description">
    description = <"The description about the Location.">
>
["at0003"] = <
    text = <"location">
    description = <"">
>
["at0004"] = <
    text = <"OM_STRING">
    description = <"">
>
["at0005"] = <
    text = <"OM_STRING">
    description = <"">
>
["at0006"] = <
    text = <"Observation">
    description = <"A Datastream has zero-to-many Observations. One
       Observation SHALL occur in one-and-only-one Datastream.">
>
["at0007"] = <
    text = <"">
    description = <" ">
>
["at0008"] = <
    text = <"ObservedProperty">
    description = <"The Observations of a Datastream SHALL observe the
      same ObservedProperty. The Observations of different Datastreams
      MAY observe the same ObservedProperty.">
>
["at0009"] = <
    text = <"Sensor">
    description = <"The Observations in a Datastream are performed by one-
      and-only-one Sensor. One Sensor MAY produce zero-to-many
      Observations in different Datastreams.">
>
["at0011"] = <
    text = <"name">
     description = <"A property provides a label for Datastream entity,
       commonly a descriptive name.">
>
["at0012"] = <
    text = <"description">
    description = <"The description of the Datastream entity.">
>
```

```
["at0013"] = <
    text = <"unitOfMeasurement">
    description = <"A JSON Object containing three key-value pairs. The
        name property presents the full name of the unitOfMeasurement; the
        symbol property shows the textual form of the unit symbol; and the
        definition contains the URI defining the unitOfMeasurement.

        The values of these properties SHOULD follow the Unified Code for
        Unit of Measure (UCUM).">
>
["at0015"] = <
    text = <"observationType">
    description = <"The type of Observation (with unique result type),
        which is used by the service to encode observations.">
>
["at0016"] = <
    text = <"observedArea">
    description = <"The spatial bounding box of the spatial extent of all
         FeaturesOfInterest that belong to the Observations associated
         with this Datastream.">
>
["at0017"] = <
    text = <"phenomenonTime">
    description = <"The temporal interval of the phenomenon times of all
     observations belonging to this Datastream.">
>
["at0025"] = <
    text = <"resultTime">
    description = <"The temporal interval of the result times of all
         observations belonging to this Datastream.">
>
["at0026"] = <
    text = <"OM_STRING">
    description = <"">
>
["at0027"] = <
    text = <"OM_STRING">
    description = <"">
>
["at0028"] = <
    text = <"OM_STRING">
    description = <"">
>
["at0029"] = <
    text = <"OM_STRING">
    description = <"">
>
["at0030"] = <
    text = <"OM_STRING">
    description = <"">
>
["at0031"] = <
    text = <"TM_Time">
    description = <"ISO 8601 Time Interval">
>
["at0032"] = <
    text = <"TM_Period">
    description = <"">
>
["at0033"] = <
    text = <"DV_DURATION">
    description = <"">
>
["at0034"] = <
    text = <"DV_INTERVAL">
    description = <"">
>
["at0038"] = <
    text = <"Results">
    description = <"">
>
["at0039"] = <
    text = <"Result">
    description = <"">
>
["at0040"] = <
    text = <"DV_TIME">
```

```
                    description = <"">
                >
            >
        >
    >
    constraint_definitions = <
        ["en"] = <
            items = <
                ["ac0002"] = <
                    text = <"JSON Object">
                    description = <"When a Datastream does not have a unit of measurement
                        (e.g., a OM_TruthObservation type), the corresponding
                        unitOfMeasurement properties SHALL have null values.">
                >
                ["ac0003"] = <
                    text = <"The observationType defines the result types for specialized
                        observations [OGC 10-004r3 and ISO 19156:2011 Table 3]. The
                        description below shows some of the valueCodes that maps the UML
                        classes in O&M v2.0 [OGC 10-004r3 and ISO 19156:2011] to
                        observationType names and observation result types.">
                    description = <"OM_CategoryObservation :
                         http://www.opengis.net/def/observationType/OGC-
                         OM/2.0/OM_CategoryObservation: URI
                         OM_CountObservation:
                         http://www.opengis.net/def/observationType/OGC-
                         OM/2.0/OM_CountObservation : integer
                         OM_Measurement: http://www.opengis.net/def/observationType/OGC-
                         OM/2.0/OM_Measurement : double
                         OM_Observation: http://www.opengis.net/def/observationType/OGC-
                         OM/2.0/OM_Observation : Any
                         OM_TruthObservation:
                         http://www.opengis.net/def/observationType/OGC-
                         OM/2.0/OM_TruthObservation : boolean">
                >
            >
        >
    >
```

Listing C.1 SensorThings API ADL based Archetype Model

Air Quailty OPT File

```
<?xml version="1.0" encoding="UTF-8"?>
<!--Operational template XML automatically generated by LinkEHR editor 20201113-->
<template xmlns="http://schemas.openehr.org/v1"
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
        xmlns:xsd="http://www.w3.org/2001/XMLSchema">
  <language>
    <terminology_id>
      <value>ISO_639-1</value>
    </terminology_id>
    <code_string>en</code_string>
  </language>
  <description>
    <original_author id="description" />
    <original_author id="text" />
    <original_author id="date">2019-01-06</original_author>
    <original_author id="name">Paul Stacey</original_author>
    <original_author id="organisation">TU Dublin</original_author>
    <original_author id="email">paul.stacey@tudublin.ie</original_author>
    <lifecycle_state>Draft</lifecycle_state>
    <other_details id="description" />
    <other_details id="text" />
    <other_details id="lastExportDate">11/13/2020 14:26:18</other_details>
    <details>
      <language>
        <terminology_id>
          <value>ISO_639-1</value>
        </terminology_id>
        <code_string>en</code_string>
      </language>
      <purpose> INSIGHT Limerick – Air Quality is a hypothetical project developed as part
         of evaluating a tranlation approach of two-level modelling from the health domain
```

```
        to the geo-spatial domain. The am of the project is to provide fine grained detail
        of the air quality at key urban locations & spaces, and to inform decision making
        about environmental practices within the Limerick region. Using this OPT Air
        quality data will be published under a Data-as-a-Service framework based on the
        SensorThings API. Allowing all citizens to access and contribute to the service.

        An air quality sensing platform will be deployed by the city council consisting
        of sensors to observe the following properties: Temperature; Precipitation; Wind
        Speed; Wind Direction; Luminosity; Noise; Particles; CO (Carbon Monoxide); NO2
        (Nitrogen Dioxide).
      </purpose>
  </details>
</description>
<template_id>
  <value>LimerickCityAQ_Report</value>
</template_id>
<concept>at0000</concept>
<definition>
  <rm_type_name>GEO_DATA_DOCUMENT</rm_type_name>
  <occurrences>
    <lower_included>true</lower_included>
    <upper_included>true</upper_included>
    <lower_unbounded>false</lower_unbounded>
    <upper_unbounded>false</upper_unbounded>
    <lower>1</lower>
    <upper>1</upper>
  </occurrences>
  <node_id>at0000</node_id>
  <attributes xsi:type="C_MULTIPLE_ATTRIBUTE">
    <rm_attribute_name>geoDataComposition</rm_attribute_name>
    <existence>
      <lower_included>true</lower_included>
      <upper_included>true</upper_included>
      <lower_unbounded>false</lower_unbounded>
      <upper_unbounded>false</upper_unbounded>
      <lower>0</lower>
      <upper>1</upper>
    </existence>
    <children xsi:type="C_COMPLEX_OBJECT">
      <rm_type_name>GEO_DATA_DOCUMENT</rm_type_name>
      <occurrences>
        <lower_included>true</lower_included>
        <lower_unbounded>false</lower_unbounded>
        <upper_unbounded>true</upper_unbounded>
        <lower>0</lower>
      </occurrences>
      <node_id>at0001</node_id>
      <attributes xsi:type="C_MULTIPLE_ATTRIBUTE">
        <rm_attribute_name>geoDataComposition</rm_attribute_name>
        <existence>
          <lower_included>true</lower_included>
          <upper_included>true</upper_included>
          <lower_unbounded>false</lower_unbounded>
          <upper_unbounded>false</upper_unbounded>
          <lower>0</lower>
          <upper>1</upper>
        </existence>
        <children xsi:type="C_COMPLEX_OBJECT">
          <rm_type_name>GEO_DATA_COMPOSITION</rm_type_name>
          <occurrences>
            <lower_included>true</lower_included>
            <lower_unbounded>false</lower_unbounded>
            <upper_unbounded>true</upper_unbounded>
            <lower>0</lower>
          </occurrences>
          <node_id>at0003</node_id>
          <attributes xsi:type="C_SINGLE_ATTRIBUTE">
            <rm_attribute_name>details</rm_attribute_name>
            <existence>
              <lower_included>true</lower_included>
              <upper_included>true</upper_included>
              <lower_unbounded>false</lower_unbounded>
              <upper_unbounded>false</upper_unbounded>
              <lower>1</lower>
              <upper>1</upper>
            </existence>
            <children xsi:type="C_COMPLEX_OBJECT">
```

```
<rm_type_name>DETAILS_COMPOSITE</rm_type_name>
<occurrences>
  <lower_included>true</lower_included>
  <upper_included>true</upper_included>
  <lower_unbounded>false</lower_unbounded>
  <upper_unbounded>false</upper_unbounded>
  <lower>1</lower>
  <upper>1</upper>
</occurrences>
<node_id>at0004</node_id>
<attributes xsi:type="C_MULTIPLE_ATTRIBUTE">
  <rm_attribute_name>element</rm_attribute_name>
  <existence>
    <lower_included>true</lower_included>
    <upper_included>true</upper_included>
    <lower_unbounded>false</lower_unbounded>
    <upper_unbounded>false</upper_unbounded>
    <lower>1</lower>
    <upper>1</upper>
  </existence>
  <children xsi:type="C_COMPLEX_OBJECT">
    <rm_type_name>DETAILS_COMPOSITE</rm_type_name>
    <occurrences>
      <lower_included>true</lower_included>
      <lower_unbounded>false</lower_unbounded>
      <upper unbounded>true</upper_unbounded>
      <lower>0</lower>
    </occurrences>
    <node_id>at0015</node_id>
  </children>
  <children xsi:type="C_COMPLEX_OBJECT">
    <rm type name>DETAILS_ELEMENT</rm_type_name>
    <occurrences>
      <lower_included>true</lower_included>
      <upper_included>true</upper_included>
      <lower_unbounded>false</lower_unbounded>
      <upper_unbounded>false</upper_unbounded>
      <lower>1</lower>
      <upper>1</upper>
    </occurrences>
    <node_id>at0016</node_id>
  </children>
  <children xsi:type="C_COMPLEX_OBJECT">
    <rm_type_name>DETAILS_ELEMENT</rm_type_name>
    <occurrences>
      <lower_included>true</lower_included>
      <upper_included>true</upper_included>
      <lower_unbounded>false</lower_unbounded>
      <upper_unbounded>false</upper_unbounded>
      <lower>1</lower>
      <upper>1</upper>
    </occurrences>
    <node_id>at0017</node_id>
  </children>
  <children xsi:type="C_COMPLEX_OBJECT">
    <rm_type_name>DETAILS_ELEMENT</rm_type_name>
    <occurrences>
      <lower_included>true</lower_included>
      <upper_included>true</upper_included>
      <lower_unbounded>false</lower_unbounded>
      <upper_unbounded>false</upper_unbounded>
      <lower>0</lower>
      <upper>1</upper>
    </occurrences>
    <node_id>at0018</node_id>
  </children>
  <cardinality>
    <is_ordered>false</is_ordered>
    <is_unique>true</is_unique>
    <interval>
      <lower_included>true</lower_included>
      <lower_unbounded>false</lower_unbounded>
      <upper_unbounded>true</upper_unbounded>
      <lower>1</lower>
    </interval>
  </cardinality>
</attributes>
```

```
    </children>
  </attributes>
  <attributes xsi:type="C_MULTIPLE_ATTRIBUTE">
    <rm_attribute_name>observationSet</rm_attribute_name>
    <existence>
      <lower_included>true</lower_included>
      <upper_included>true</upper_included>
      <lower_unbounded>false</lower_unbounded>
      <upper_unbounded>false</upper_unbounded>
      <lower>0</lower>
      <upper>1</upper>
    </existence>
    <children xsi:type="C_COMPLEX_OBJECT">
      <rm_type_name>OBSERVATION_SET</rm_type_name>
      <occurrences>
        <lower_included>true</lower_included>
        <lower_unbounded>false</lower_unbounded>
        <upper_unbounded>true</upper_unbounded>
        <lower>0</lower>
      </occurrences>
      <node id>at0005</node_id>
    </children>
    <children xsi:type="C_COMPLEX_OBJECT">
      <rm_type_name>OBSERVATION_SET</rm_type_name>
      <occurrences>
        <lower included>true</lower included>
        <lower_unbounded>false</lower_unbounded>
        <upper_unbounded>true</upper_unbounded>
        <lower>0</lower>
      </occurrences>
      <node_id>at0006</node_id>
    </children>
    <children xsi:type="C_COMPLEX_OBJECT">
      <rm_type_name>OBSERVATION_SET</rm_type_name>
      <occurrences>
        <lower_included>true</lower_included>
        <lower_unbounded>false</lower_unbounded>
        <upper_unbounded>true</upper_unbounded>
        <lower>0</lower>
      </occurrences>
      <node_id>at0008</node_id>
    </children>
    <children xsi:type="C_COMPLEX_OBJECT">
      <rm_type_name>OBSERVATION_SET</rm_type_name>
      <occurrences>
        <lower_included>true</lower_included>
        <lower_unbounded>false</lower_unbounded>
        <upper_unbounded>true</upper_unbounded>
        <lower>0</lower>
      </occurrences>
      <node_id>at0009</node_id>
    </children>
    <children xsi:type="C_COMPLEX_OBJECT">
      <rm_type_name>OBSERVATION_SET</rm_type_name>
      <occurrences>
        <lower_included>true</lower_included>
        <lower_unbounded>false</lower_unbounded>
        <upper_unbounded>true</upper_unbounded>
        <lower>0</lower>
      </occurrences>
      <node_id>at0010</node_id>
    </children>
    <children xsi:type="C_COMPLEX_OBJECT">
      <rm_type_name>OBSERVATION_SET</rm_type_name>
      <occurrences>
        <lower_included>true</lower_included>
        <lower_unbounded>false</lower_unbounded>
        <upper_unbounded>true</upper_unbounded>
        <lower>0</lower>
      </occurrences>
      <node_id>at0011</node_id>
    </children>
    <children xsi:type="C_COMPLEX_OBJECT">
      <rm_type_name>OBSERVATION_SET</rm_type_name>
      <occurrences>
        <lower included>true</lower included>
        <lower_unbounded>false</lower_unbounded>
```

```
            <upper_unbounded>true</upper_unbounded>
            <lower>0</lower>
          </occurrences>
          <node_id>at0012</node_id>
        </children>
        <children xsi:type="C_COMPLEX_OBJECT">
          <rm_type_name>OBSERVATION_SET</rm_type_name>
          <occurrences>
            <lower_included>true</lower_included>
            <lower_unbounded>false</lower_unbounded>
            <upper_unbounded>true</upper_unbounded>
            <lower>0</lower>
          </occurrences>
          <node_id>at0013</node_id>
        </children>
        <children xsi:type="C_COMPLEX_OBJECT">
          <rm_type_name>OBSERVATION_SET</rm_type_name>
          <occurrences>
            <lower_included>true</lower_included>
            <lower_unbounded>false</lower_unbounded>
            <upper unbounded>true</upper_unbounded>
            <lower>0</lower>
          </occurrences>
          <node_id>at0014</node_id>
        </children>
        <cardinality>
          <is_ordered>false</is_ordered>
          <is_unique>true</is_unique>
          <interval>
            <lower_included>true</lower_included>
            <lower_unbounded>false</lower_unbounded>
            <upper unbounded>true</upper_unbounded>
            <lower>0</lower>
          </interval>
        </cardinality>
      </attributes>
    </children>
    <cardinality>
      <is_ordered>false</is_ordered>
      <is_unique>true</is_unique>
      <interval>
        <lower_included>true</lower_included>
        <lower_unbounded>false</lower_unbounded>
        <upper_unbounded>true</upper_unbounded>
        <lower>0</lower>
      </interval>
    </cardinality>
  </attributes>
</children>
<children xsi:type="C_COMPLEX_OBJECT">
  <rm_type_name>GEO_DATA_DOCUMENT</rm_type_name>
  <occurrences>
    <lower_included>true</lower_included>
    <lower_unbounded>false</lower_unbounded>
    <upper_unbounded>true</upper_unbounded>
    <lower>0</lower>
  </occurrences>
  <node_id>at0007</node_id>
</children>
<cardinality>
  <is_ordered>false</is_ordered>
  <is_unique>true</is_unique>
  <interval>
    <lower_included>true</lower_included>
    <lower_unbounded>false</lower_unbounded>
    <upper_unbounded>true</upper_unbounded>
    <lower>0</lower>
  </interval>
</cardinality>
</attributes>
<attributes xsi:type="C_SINGLE_ATTRIBUTE">
  <rm_attribute_name>archetype_id</rm_attribute_name>
  <existence>
    <lower_included>true</lower_included>
    <upper_included>true</upper_included>
    <lower unbounded>false</lower unbounded>
    <upper_unbounded>false</upper_unbounded>
```

```
      <lower>0</lower>
      <upper>1</upper>
    </existence>
  </attributes>
  <attributes xsi:type="C_SINGLE_ATTRIBUTE">
    <rm_attribute_name>details</rm_attribute_name>
    <existence>
      <lower_included>true</lower_included>
      <upper_included>true</upper_included>
      <lower_unbounded>false</lower_unbounded>
      <upper_unbounded>false</upper_unbounded>
      <lower>1</lower>
      <upper>1</upper>
    </existence>
  </attributes>
  <attributes xsi:type="C_SINGLE_ATTRIBUTE">
    <rm_attribute_name>name</rm_attribute_name>
    <existence>
      <lower_included>true</lower_included>
      <upper_included>true</upper_included>
      <lower_unbounded>false</lower_unbounded>
      <upper_unbounded>false</upper_unbounded>
      <lower>1</lower>
      <upper>1</upper>
    </existence>
  </attributes>
  <archetype_id>
    <value>TPOT-OM-Geo_Data_Document.LimerickCityAQ_Report.v1</value>
  </archetype_id>
  <template_id>
    <value>LimerickCityAQ_Report</value>
  </template_id>
  <term_definitions code="at0000">
    <items id="description">LimerickCityAQ_Report</items>
    <items id="text">LimerickCityAQ_Report</items>
  </term_definitions>
  <term_definitions code="at0001">
    <items id="description" />
    <items id="text">AQ_SensorDataRecord</items>
  </term_definitions>
  <term_definitions code="at0007">
    <items id="description" />
    <items id="text">AQ_IndexRecord</items>
  </term_definitions>
  <term_definitions code="at0003">
    <items id="description" />
    <items id="text">AQ_Station</items>
  </term_definitions>
  <term_definitions code="at0004">
    <items id="description">*</items>
    <items id="text">*DETAILS_COMPOSITE</items>
  </term_definitions>
  <term_definitions code="at0005">
    <items id="description" />
    <items id="text">Particles</items>
  </term_definitions>
  <term_definitions code="at0006">
    <items id="description" />
    <items id="text">Precipitation</items>
  </term_definitions>
  <term_definitions code="at0008">
    <items id="description" />
    <items id="text">Luminosity</items>
  </term_definitions>
  <term_definitions code="at0009">
    <items id="description" />
    <items id="text">Noise</items>
  </term_definitions>
  <term_definitions code="at0010">
    <items id="description" />
    <items id="text">CO</items>
  </term_definitions>
  <term_definitions code="at0011">
    <items id="description" />
    <items id="text">NO2</items>
  </term_definitions>
  <term_definitions code="at0012">
```

```
      <items id="description" />
      <items id="text">Temperature</items>
    </term_definitions>
    <term_definitions code="at0013">
      <items id="description" />
      <items id="text">WindDirection</items>
    </term_definitions>
    <term_definitions code="at0014">
      <items id="description" />
      <items id="text">WindSpeed</items>
    </term_definitions>
    <term_definitions code="at0015">
      <items id="description">The Location entity locates the Thing. Multiple Things MAY
          be located at the same Location. A Thing MAY not have a Location. A Thing SHOULD
          have only one Location.However, in some complex use cases, a Thing MAY have
          more than one Location representations. In such case, the Thing MAY have more
          than one Locations. These Locations SHALL have different encodingTypes and the
          encodingTypes SHOULD be in different spaces (e.g., one encodingType in
          Geometrical space and one encodingType in Topological space).</items>
      <items id="text">Location</items>
    </term_definitions>
    <term_definitions code="at0016">
      <items id="description">*A property provides a label for Thing entity, commonly a
         descriptive name.</items>
      <items id="text">*name</items>
    </term_definitions>
    <term_definitions code="at0017">
      <items id="description">*This is a short description of the corresponding Thing
         entity.</items>
      <items id="text">*description</items>
    </term_definitions>
    <term_definitions code="at0018">
      <items id="description">*A JSON Object containing user-annotated properties as key-
         value pairs.</items>
      <items id="text">*properties</items>
    </term_definitions>
  </definition>
</template>
```

Listing C.2 Air Quality OPT File

Evaluation 2 Ocean Observing Files

oceanSITES netCDF and Oceanotron Archetype Model Specialisation

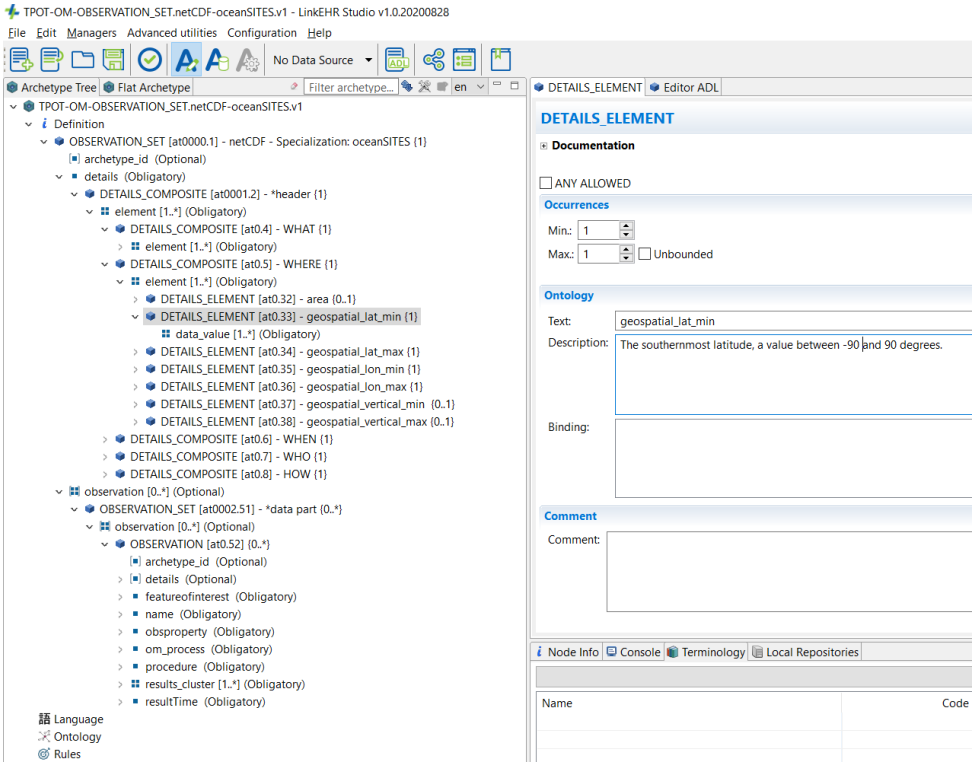


Figure C.1 Developing the *oceanSITES* Archetype

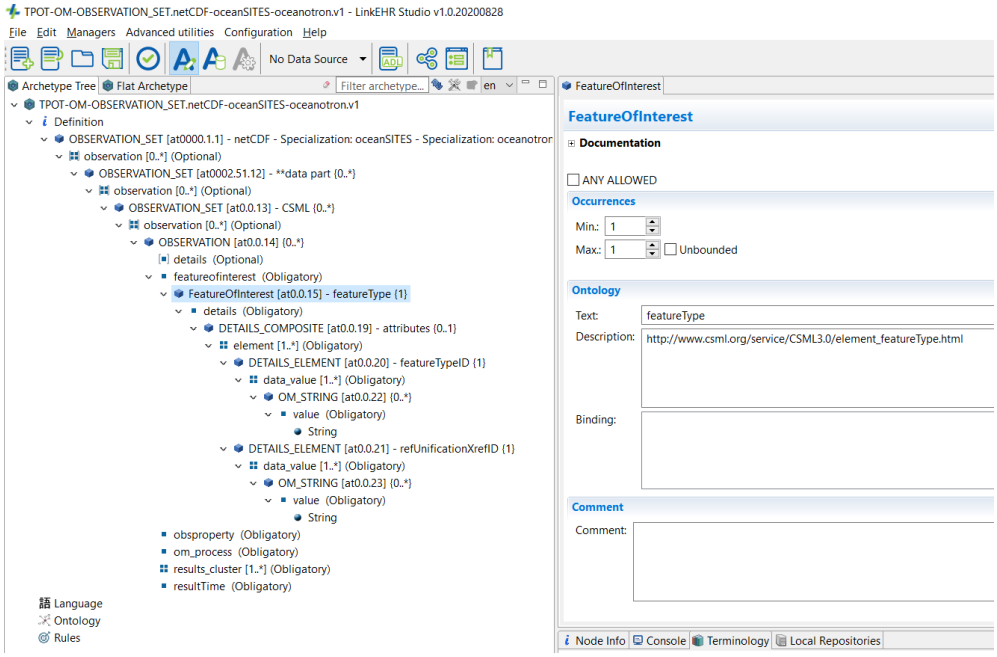


Figure C.2 Developing the *oceanotron* Archetype

oceanSITES specialisation of Platform Archetype Model

```
archetype (adl_version=1.4)
    TPOT-OM-GeoData_COMPOSITION.platform-oceanSITES.v1
specialize
    TPOT-OM-GeoData_COMPOSITION.platform.v1
concept
    [at0000.1]
language
    original_language = <[ISO_639-1::en]>
description
    original_author = <
        ["name"] = <"Paul Stacey">
        ["organisation"] = <"TU Dublin">
    >
    lifecycle_state = <"Draft">
    details = <
        ["en"] = <
            language = <[ISO_639-1::en]>
        >
    >
definition
    GeoData_COMPOSITION[at0000.1] occurrences matches {1..1} matches {  -- platform -
                                                          Specialization: oceanSITES
        archetype_id existence matches {0..1} matches {*}
        details existence matches {1..1} matches {
            DETAILS_COMPOSITE[at0001.2] occurrences matches {1..1} matches {  --
                                                           *platform_details
                element existence matches {1..1} cardinality matches {1..*; unordered;
                                                          unique} matches {
                    DETAILS_ELEMENT[at0002.3] occurrences matches {1..1} matches {  --
                                                             *platform_type
                        data_value existence matches {1..1} cardinality matches {1..*;
                                                          unordered; unique} matches {
                            OM_STRING[at0.6] occurrences matches {0..*} matches {  --
                                value existence matches {1..1} matches {/.*/}
                            }
                        }
                    }
                    DETAILS_ELEMENT[at0004.4] occurrences matches {1..1} matches {*}  --
                                                                  *location
                    DETAILS_ELEMENT[at0005.5] occurrences matches {1..1} matches {  --
                                                            platform_category
                        data_value existence matches {1..1} cardinality matches {1..1;
                                                           unordered; unique} matches {
                            OM_STRING[at0.8] occurrences matches {0..*} matches {  --
                                value existence matches {1..1} matches {"Air-Sea Flux
                                     Site","TransportSite", "Physical", "Meteorological",
                                      "Biogeochemical", "Geophysical"}
                            }
                        }
                    }
                    DETAILS_ELEMENT[at0.7] occurrences matches {0..1} matches {  --
                                                             wmo_message_format
                        data_value existence matches {1..1} cardinality matches {1..*;
                                                           unordered; unique} matches {
                            OM_STRING[at0.10] occurrences matches {0..*} matches {  --
                                value existence matches {1..1} matches
                                                          {"FM13","FM18","FM64","FM65"}
                            }
                        }
                    }
                    DETAILS_ELEMENT[at0.9] occurrences matches {0..1} matches {*}  --
                                                           wind_direction_conventions
                    DETAILS_ELEMENT[at0.11] occurrences matches {0..1} matches {  --
                                                    platform_message_reporting_frequency
                        data_value existence matches {1..1} cardinality matches {1..*;
                                                          unordered; unique} matches {
                            OM_STRING[at0.12] occurrences matches {0..*} matches {  --
                                value existence matches {1..1} matches {/.*/}
                            }
                        }
                    }
                }
            }
```

```
        }
    }

ontology
    term_definitions = <
        ["en"] = <
            items = <
                ["at0000.1"] = <
                    text = <"platform - Specialization: oceanSITES">
                    description = <"Keyword identifies a specific vehicle, object,
                      structure or organism capable of bearing instruments or tools for
                      the collection of physical, chemical, geological or biological
                      samples.
                      http://vocab.nerc.ac.uk/collection/L19/current/SDNKG04/
                      SDN:L19::SDNKG04 (SeaDataNet) - Specialization: oceanSITES

                      An OceanSITES platform is an independently deployable package of
                      instruments and sensors forming part of site. It may be fixed to
                      the ocean floor, may float or may be self-propelled">
                >
                ["at0001.2"] = <
                    text = <"*platform_details">
                    description = <"*">
                >
                ["at0002.3"] = <
                    text = <"*platform_type">
                    description = <"*https://mmisw.org/ont/ioos/platform">
                >
                ["at0004.4"] = <
                    text = <"*location">
                    description = <"*">
                >
                ["at0005.5"] = <
                    text = <"platform_category">
                    description = <"Air-Sea Flux Site, Transport Site, Physical,
                      Meteorological, Biogeochemical, Geophysical
                      Ref:                                                              :
                      http://www.odip.org/documents/odip/downloads/19/oceansites_user_ma
                      nual_version1.2.pdf section 3.1">
                >
                ["at0.6"] = <
                    text = <"">
                    description = <"">
                >
                ["at0.8"] = <
                    text = <"">
                    description = <"">
                >
                ["at0.9"] = <
                    text = <"wind_direction_conventions">
                    description = <"WMO standard uses wind-from-direction, indicate if the
                      real-time wind direction received by GDAC/DAC is a wind-to-direction
                      before GTS dissemination">
                >
                ["at0.7"] = <
                    text = <"wmo_message_format">
                    description = <"WMO standard formats: FM13, FM18, FM64, or FM65. PIs
                      may request desired WMO formats and GDAC will determine the final
                      formats to be used
                      http://www.odip.org/documents/odip/downloads/19/oceansites_user_ma
                      nual_version1.2.pdf section 3.1">
                >
                ["at0.10"] = <
                    text = <"">
                    description = <"">
                >
                ["at0.11"] = <
                    text = <"platform_message_reporting_frequency">
                    description = <"The frequency of message reporting from buoy to DAC,
                      such as daily, hourly, or every 10min etc.
                      ref:
                      http://www.odip.org/documents/odip/downloads/19/oceansites_user_ma
                      nual_version1.2.pdf section 3.1">
                >   ["at0.12"] = <text = <""> description = <""> > > > >
```

Listing C.3 OceanSITES ADL Archetype Model

Warp (TH1) NMMP SmartBuoy Dataset

Authors note: This dataset is one of three used during Evaluation 2, chapter 6. The dataset was obtained from the EMODnet-Physics data portal. The data were retrieved in netCDF format. NetCDF data files were converted to JSON using the netCDF operator tool suite NCO toolkit (Zender et al., 2012) for ease of parsing and assessment. One of the platform's datasets is reproduced below. The dataset has been converted to CSV format, with several days removed for document formatting purposes.

Platform Details: WARP CEFAS-62010720 http://wavenet.cefas.co.uk/Smartbuoy/Map / platform code 6201072 Warp-TH1-6201072 /

http://www.emodnet-physics.eu/Map/platinfo/piroosplot.aspx?platformid=11836

```
#Global attributes;Value
# data_type;="OceanSITES time-series data"
# format_version;="1.2"
# platform_code;="6201072"
# platform_name;="Warp (TH1) NMMP SmartBuoy"
# date_update;="2019-04-15T07:05:37Z"
# institution;="Centre for Environment - Fisheries and Aquaculture Science"
# institution_edmo_code;="28"
# site_code;="NO"
# wmo_platform_code;="6201072"
# source;="mooring"
# source_platform_category_code;="48"
# history;="2019-04-15T07:05:37Z : Creation"
# data_mode;="R"
# quality_control_indicator;="1"
# quality_index;="A"
# references;="http://www.oceansites.org,http://www.myocean.org"
# comment;="None"
# Conventions;="CF-1.6 OceanSITES-Manual-1.2 Copernicus-InSituTAC-SRD-1.4 Copernicus-InSituTAC-ParametersList-3.1.0"
# title;="NWS - NRT in situ Observations"
# summary;="Oceanographic data from North West Shelf"
# naming_authority;="OceanSITES"
# id;="NO_TS_MO_6201072_201608"
# cdm_data_type;="Time-series"
# area;="North West Shelf"
# geospatial_lat_min;="51.5255"
```

```
# geospatial_lat_max;="51.5255"
# geospatial_lon_min;="1.028"
# geospatial_lon_max;="1.028"
# geospatial_vertical_min;="0"
# geospatial_vertical_max;="1"
# family_label;="mooring"
# family_code;="MO"
# time_coverage_start;="2016-08-01T00:00:00Z"
# time_coverage_end;="2016-08-31T23:59:59Z"
# institution_references;="http://www.cefas.co.uk/ "
# contact;="cmems-service@bsh.de"
# author;="cmems-service"
# data_assembly_center;="German National Oceanographic Data Centre"
# pi_name;="sarah.turner@cefas.co.uk"
# distribution_statement;="These data follow Copernicus standards; they are public and free of charge. User assumes all risk for use of data. User
must display citation in any publication or product using data. User must contact PI prior to any commercial use of data."
# citation;="These data were collected and made freely available by the Copernicus project and the programs that contribute to it"
# update_interval;="hourly"
# qc_manual;="OceanSITES User's Manual v1.2"
# last_date_observation;="2016-08-17T03:59:08Z"
# last_latitude_observation;="51.5255"
# last_longitude_observation;="1.028"
# netcdf_version;="netCDF-4 classic model"


TIME;DEPTH;LATITUDE;LGH4;LGH4_QC;LONGITUDE;OSAT;OSAT_QC;POSITION_QC;PSAL;PSAL_QC;TEMP;TEMP_QC;TIME_QC;TUR6;TUR6_QC
02/08/2016 19:59:08:000;0;51,5255012512207;0,287;1;1,02799999713898;0;-127;1;0;-127;0;-127;1;0;-127
02/08/2016 19:59:08:000;1;51,5255012512207;0;-127;1,02799999713898;101,237;1;1;34,225;1;19,625;1;1;2,197;1
02/08/2016 21:59:08:000;0;51,5255012512207;0,141;1;1,02799999713898;0;-127;1;0;-127;0;-127;1;0;-127
02/08/2016 21:59:08:000;1;51,5255012512207;0;-127;1,02799999713898;99,475;1;1;34,339;1;19,474;1;1;2,753;1
02/08/2016 23:59:08:000;0;51,5255012512207;0,141;1;1,02799999713898;0;-127;1;0;-127;0;-127;1;0;-127
02/08/2016 23:59:08:000;1;51,5255012512207;0;-127;1,02799999713898;98,637;1;1;34,399;1;19,173;1;1;5,183;1
```

//Observational data from 3rd of August to 16th of August 2016 removed for brevity by Author

```
17/08/2016 01:59:08:000;0;51,5255012512207;0,14;1;1,02799999713898;0;-127;1;0;-127;0;-127;1;0;-127
17/08/2016 01:59:08:000;1;51,5255012512207;0;-127;1,02799999713898;101,833;1;1;34,454;1;19,271;1;1;9,842;1
17/08/2016 03:59:08:000;0;51,5255012512207;0,14;1;1,02799999713898;0;-127;1;0;-127;0;-127;1;0;-127
17/08/2016 03:59:08:000;1;51,5255012512207;0;-127;1,02799999713898;101,557;1;1;34,317;1;19,304;1;1;7,099;1
```

Generalised Additive Model Parameter Summary

Table C.1 GAM Model Parameter Summary.

LinearGAM

Distribution:	NormalDist	Effective DoF:	1.1859
Link Function:	IdentityLink	Log Likelihood:	-17.1797
Number of Samples:	15	AIC:	38.7312
		AICc:	39.9101
		GCV:	0.5046
		Scale:	0.4333
		Pseudo R-Squared:	0.0142

Feature Function	Lambda	Rank	EDoF	P > x	Sig. Code
s(0)	[0.6]	20		1.11e-16	***
s(1)	[0.6]	20		1.11e-16	***
s(2)	[0.6]	20		1.11e-16	***
intercept		1		4.71e-10	***

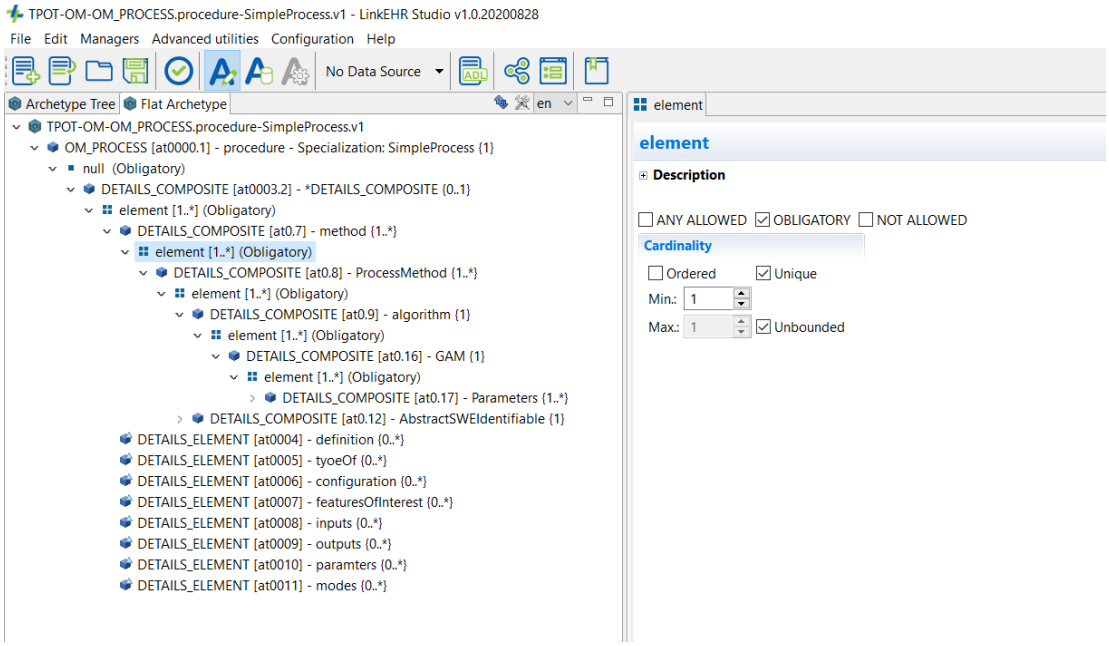


Figure C.3 Developing a GAM based *SimpleProcess* archetype

Appendix D

RESTful Approaches for Constrained Devices

The IPv6 protocol stack for low-power and lossy networks (LLNs) consist of the traditional IPv6 protocols but subsequently augmented with the IETFs RPL routing protocol, and the 6loWPAN adaption layer. 6LoWPAN is a standard specified by the IETF (RFC4944) that provides IP networking on top of IEEE 802.15.4 compliant devices. Where 802.15.4 defines media-access controller (MAC) and the physical circuits (PHY) layers, 6LoWPAN is layered above the MAC. An adaption layer is defined to bridge interoperability issues between IPv6 and 802.15.4 (Zigbee) networks. There are a number IPv6 challenges in sensor networks such as implementation complexity, header compression and routing. The IETF 6loWPAN adaption layer (Figure D.1) and IETF RPL Protocol provide solutions that are suitable for constrained sensor networks.

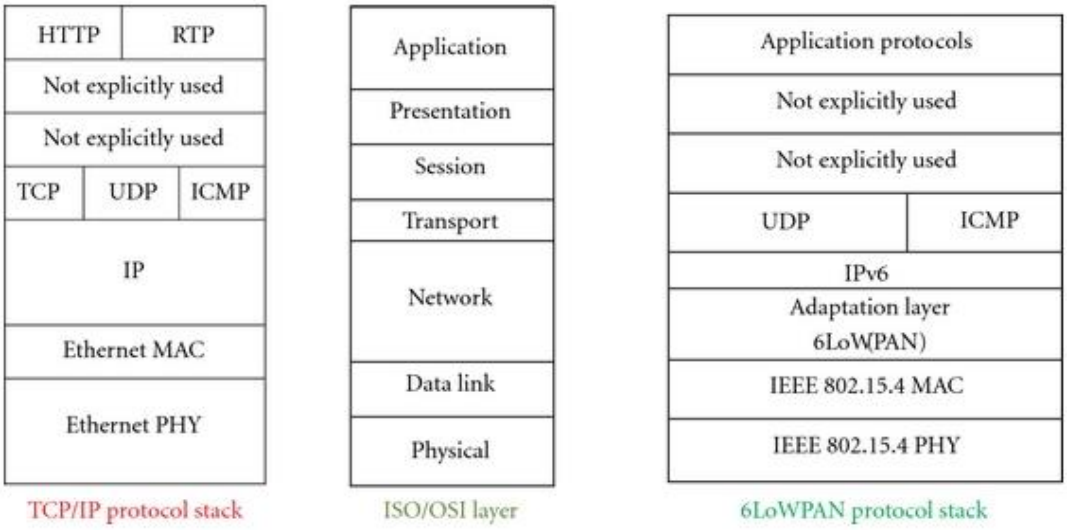


Figure D.1 6LoWPAN protocol stack mapped to OSI and TCP/IP stacks

The IETF has also developed a specialised routing protocol for low-power and lossy networks over IPv6 called RPL (Routing Protocol for Low-Power and Lossy Networks). RPL has been defined for a many-to-one traffic environment: where many nodes route

data back to one point (a gateway/border/sink node). However, any-to-any routing is also possible. Both Contiki and TinyOS provide independent implementations of the IETFs RPL: ContikiRPL and TinyRPL. Ko (2011) gives a good overview of implementation experiences with both ContikiRPL and TinyRPL, and interoperability experiences with both running in the same sensor network (Figure D.2).

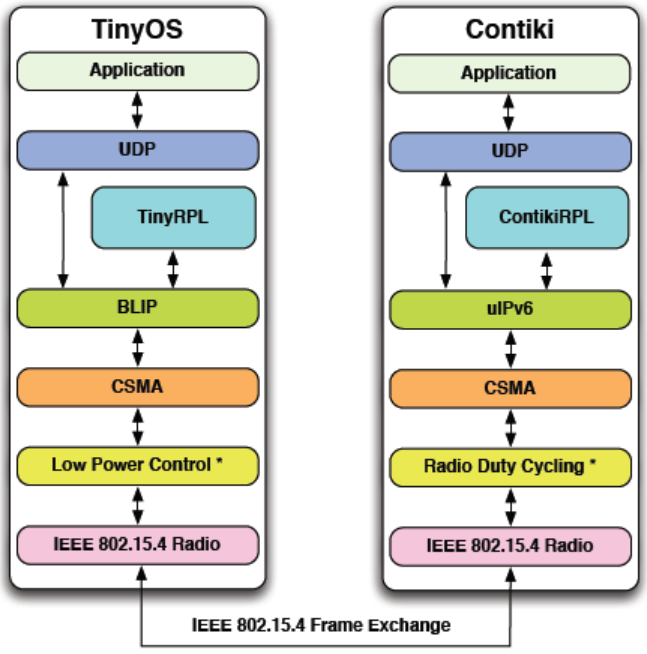


Figure D.2 ContikiRPL and TinyRPL interoperability (Ko, 2011)

6LoWPAN only provides IP connectivity but no interoperability at higher layers. Web services in constrained devices have been proposed as a solution. Web services can follow several architectural styles, for example REST (Fielding and Taylor 2002) and SOAP (Gudgin et al. 2003). Analysis of techniques has shown that following RESTful implementation principles results in a lower overhead than SOAP. (Stirbu, 2008) has shown how RESTful techniques can be applied to sensor networks and nodes.

To support web services running on platforms with very limited resources the IETF formed the Constrained RESTful Environments group (CoRE) (Shelby, 2012). CoRE has been tasked with developing a framework for deploying web services to constrained

environments such as sensor nodes. In the CoRE framework a network of nodes called Devices interact. Devices are responsible for one or more Resources which could be a representation of sensors, actuators, combinations of values or other information. Devices in the network can send messages to each other to request, query and publish data. As part of the overall effort to enable these types of applications to be built, the Constrained Application Protocol (CoAP) (Shelby et al., 2012) has been defined.

CoAP

The Constrained Application Protocol (CoAP) is a specialised web transfer protocol for use with constrained nodes and networks. Several studies have shown improved performance of CoAP over HTTP in terms of ROM usage and response time. CoAP provides a request/response interaction model between application endpoints, supports built-in discovery of services and resources. CoAP is designed to easily interface with HTTP for integration with the Web with very low overhead, and simplicity for constrained environments.

To-date there have been many implementations of CoAP and libraries exist for many of the WSN based operating systems. For TinyOS, TinyOS Blip is available as a CoAP external library. Ludovici et al. (2013) describes a novel CoAP implementation for TinyOS (TinyCoAP). TinyCoAP differs from TinyOS Blip as it is developed as a native library for TinyOS. TinyCoAP claims to be a better option over Blip as its native implementation means the code will be optimised. One of the problems with Blip is that it is built around a dynamic memory allocation model.

Publications & Communications

Journals

Stacey, P. and Berry, D. (2018) Towards a Digital Earth: Using Archetypes to enable Knowledge Interoperability within Geo Observational Sensor Systems Design. Journal of Earth Science Informatics (2018). doi:10.1007/s12145-018-0340-z.

Conferences

Stacey, P. and Berry, D. (2019a). Extending Two-Level Information Modelling to the Internet of Things, In Proc. of IEEE World Forum on IoT Limerick, Ireland, 2019.

Stacey, P. and Berry, D. (2019b). Beyond Standards Compliant Internet of Things Data-streams. [*Tutorial*] Presented at the IEEE World Forum on IoT, Limerick, Ireland, 2019.

Stacey, P. and Berry, D. (2018) Interoperable Ocean Observing using Archetypes: A use-case based evaluation, In Proc. of IEEE OCEANS 2018. IEEE/MTS OCEANS Conference, Charleston, USA, October 2018.

Stacey, P. and Berry, D. (2017) A Community-Consensus Approach to Knowledge Interoperability within Heterogeneous Earth System Science based Observational Systems, [*abstracts*] European Meteorological Society Annual Conference, Dublin, September 2017.

Stacey, P. and Berry, D. (2017) Design and Implementation of an Archetype Based Interoperable Knowledge Eco-System for Data Buoys. In Proc. of IEEE/MTS Oceans conference, Aberdeen, Scotland, June 2017.

Stacey, P. and Berry, D. (2015) Applying Two-Level Modelling to Remote Sensor Systems Design to Enable Future Knowledge Generation, [*poster*] IEEE YP Conference in Remote Sensing 2015, December 3-4, Barcelona, Spain, 2015.

Other/Related

Leadbetter, A., Buck, J., Stacey, P. (2015) Practical Solutions to Implementing Born Semantic Data Systems, [*abstracts*] American Geophysical Union Fall Meeting, 2015, December 14-18, San Francisco.