Dissertations

School of Computer Science

2015-03-16

Investigating Correlation between Availability of Free Time Facilities for Pupils and Their Educational Achievements while Exploiting Linked Data and Volunteered Geographic Information

Peter Kovar
Technological University Dublin

Recommended Citation

Kovar, P. (2015) Investigating Correlation between Availability of Free Time Facilities for Pupils and Their Educational Achievements while Exploiting Linked Data and Volunteered Geographic, Masters Dissertation. Technological University Dublin, 2015.

Investigating Correlation between Availability of Free Time Facilities for Pupils and Their Educational Achievements while Exploiting Linked Data and Volunteered Geographic Information

Peter Kovár

A dissertation submitted in partial fulfilment of the requirements of
Dublin Institute of Technology for the degree of
M.Sc. in Computing (Data Analytics)

March 2015

I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Data Analytics), 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 test of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Dublin Institute of Technology and has not been submitted in whole or part for an award in any other Institute or University.

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

Signed: ____________________________________

Date: **6 March 2015**

ABSTRACT

The advent of Semantic Web as an extension of the current World Wide Web (Web) brings methods and technologies to structure the data, interlink it with other related data and capture its semantics. This holds for new data to be published as well as for the data already on the Web. The interlinking of data led to definition of a linked data paradigm and the Semantic Web allows for creating a Global Linked Data Space. The linked data sets can combine data coming from heterogeneous sources. Using ontologies the meaning in data from different areas of interest can be compared and also new knowledge inferred.

Current literature recognises potential the linked data represents and actively seeks novel uses of it. For Data Analytics using linked data means ability to bring more contextual information and ultimately gain more meaningful insights and improve accuracy of analysis results.

To this end educational attainment data of pupils was analysed for correlation relationship with availability of after-school and free time facilities in corresponding areas as part of this project. Additionally linked data from selected geographical source was assessed for its benefits when included in data analysis.

Key words: Linked data, Geospatial linked data, Semantic Web, Volunteered Geographic Information, Educational Attainment, Educational Performance

ACKNOWLEDGMENTS

I would like to thank my supervisor Dr. Pierpaolo Dondio for his guidance and help throughout the dissertation.

Thank you Naneth and Nathanko for your endless support and encouragement throughout the whole process as it could not have been completed without you two.

TABLE OF CONTENTS

TABLE OF FIGURES

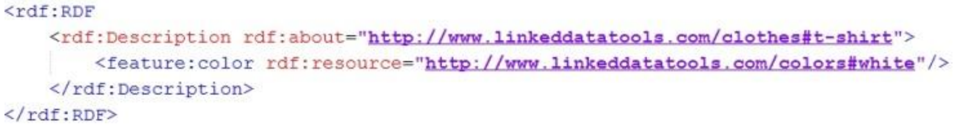
TABLE OF TABLES

1. INTRODUCTION

1.1 *Project Background*

The World Wide Web (Web) has been providing a global platform for publishing and interlinking documents of any kind for many years. With the amount of data on the Web increasing exponentially search engines tend to return more and more irrelevant results. The advent of Semantic Web as an extension of the current one brings methods and technologies to structure the data, interlink it with other related data and capture its semantics, i.e. meaning, in a way that can be processed programmatically.

Among components making the Semantic Web working are Resource Description Framework (RDF) and ontologies. RDF is a data model for describing Web resources in form of triples - <subject, predicate, object>. The subject, the resource itself, can have attribute or relationship described by the predicate with some other resource, the object (Chen *et al.* 2012). The triples are structured using custom XML (Extensible Markup Language) tags.



```
<rdf:RDF
    <rdf:Description rdf:about="http://www.linkeddatatools.com/clothes#t-shirt">
        <feature:color rdf:resource="http://www.linkeddatatools.com/colors#white"/>
    </rdf:Description>
</rdf:RDF>
```

Figure 1.1: RDF triple example[1]

The data model underlying the RDF is a directed labelled graph of data.

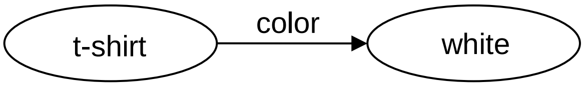


Figure 1.2: RDF triple as a graph of data

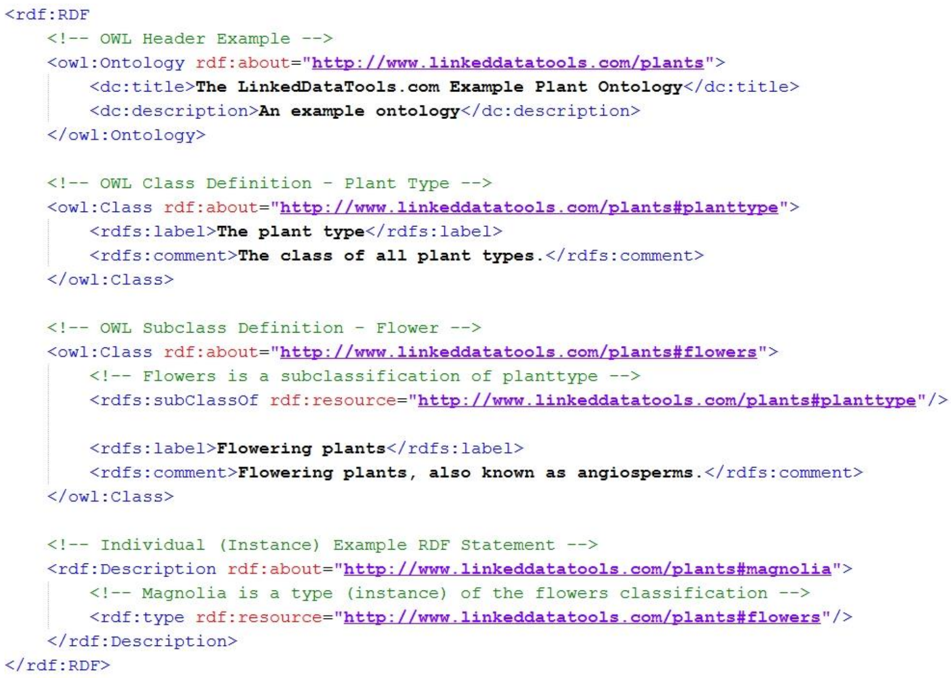
Although a RDF triple describes a relationship between Web resources there is actually no meaning behind the data. For this reasons ontologies are required. Ontologies are used to define terms belonging to an area of interest. They specify terms that can be used in a particular context including relationships and constrains of the terms. Ontologies serve to integrate the heterogeneous data on the Semantic Web. In case of ambiguities of terms, e.g. zip code vs. postal code both representing the same concept, combining

[1] Example taken from Semantic Web tutorial available from http://www.linkeddatatools.com/introducing-rdf

ontologies can help to resolve them, i.e. establish that they have the same meaning, but also to discover new relationships (Berners-Lee *et al.* 2001).

Ontology with regards to philosophy studies theories about things that are. In terms of Artificial Intelligence '… an ontology is a document or file that formally defines the relations among terms' (Berners-Lee *et al.* 2001). On the Semantic Web ontologies serve to relate and thus allow for linking data from various sources and ultimately from different disciplines, areas of interest.

OWL (Ontology Web Language) extends RDF as its better expressiveness allows for explicitly representing meaning of terms in ontologies and relationships between the terms (McGuinness and Van Harmelen 2004). In contrast to RDF which describes only simple semantics the OWL ads description of properties and classes, relationships between classes and instances of classes. To constrain values of both classes and instances axioms are used in OWL. The axioms can be used for inferring knowledge and they can eliminate combining of incompatible data (Chen *et al.* 2012). There have been already many ontologies defined and made available and OWL allows them to be reused and extended to meet particular needs.



```
<rdf:RDF
    <!-- OWL Header Example -->
    <owl:Ontology rdf:about="http://www.linkeddatatools.com/plants">
        <dc:title>The LinkedDataTools.com Example Plant Ontology</dc:title>
        <dc:description>An example ontology</dc:description>
    </owl:Ontology>

    <!-- OWL Class Definition - Plant Type -->
    <owl:Class rdf:about="http://www.linkeddatatools.com/plants#planttype">
        <rdfs:label>The plant type</rdfs:label>
        <rdfs:comment>The class of all plant types.</rdfs:comment>
    </owl:Class>

    <!-- OWL Subclass Definition - Flower -->
    <owl:Class rdf:about="http://www.linkeddatatools.com/plants#flowers">
        <!-- Flowers is a subclassification of planttype -->
        <rdfs:subClassOf rdf:resource="http://www.linkeddatatools.com/plants#planttype"/>

        <rdfs:label>Flowering plants</rdfs:label>
        <rdfs:comment>Flowering plants, also known as angiosperms.</rdfs:comment>
    </owl:Class>

    <!-- Individual (Instance) Example RDF Statement -->
    <rdf:Description rdf:about="http://www.linkeddatatools.com/plants#magnolia">
        <!-- Magnolia is a type (instance) of the flowers classification -->
        <rdf:type rdf:resource="http://www.linkeddatatools.com/plants#flowers"/>
    </rdf:Description>
</rdf:RDF>
```

Figure 1.3: OWL ontology example[2]

[2] Example taken from Semantic Web tutorial available from http://www.linkeddatatools.com/introducing-rdfs-owl

Figure 1.3 shows an example of a simple ontology for ‘plants’ defined using OWL. The ontology defines a base class ‘plant type’ with subclass ‘flowers’. To define hierarchies like this RDFS (RDF Schema) is used which is another extension of RDF. The last part of the code in rdf:Description tag defines an actual RDF triple. The subject of the triple is ‘magnolia’ with predicate ‘type’ and object ‘flowers’. As the object URI points to class ‘flowers’ this triple creates an instance, in OWL called individual, of the class ‘flowers’. In practice the RDF triple would be defined outside the ontology in its own rdf:RDF tag.

Continuing with the plants example, once additional subclasses, e.g. shrubs, crops etc., are defined in the OWL ontology and then each instantiated using RDF triples a Linked data set about plants will be created. The ontology will then allow for discovering meaningful relationships among plants, e.g. magnolia is a flower and flowers are type of a plant etc., and also making inferences. Including semantic data created using these technologies in the Metadata of Web sites makes the data of the sites ‘understandable’ for and processable by the machines.

Linked data since its principles were defined by Berners-Lee (2006) for the first time gradually became a well-recognised format for publishing structured data on the Web and continues to gain its popularity also among state institutions while they are making by them held data sets freely available under open data initiatives. As more and more data sets are published in this format so it grows its importance in the data analysis. Most of the research has been done in terms of transforming data sets into the linked format, defining and aligning ontologies, optimising performance of querying the linked data and others. On the other hand there seems to be only little research concerned with practical applications of analyses supported by linked data published (e.g. Kämpgen 2011, Kalampokis *et al.* 2013).

1.1 Project Description

With growing number of published open data sets and increasing popularity of linked data format new opportunities for analysing data across heterogeneous data sets are emerging thus promising potential for new knowledge discovery or verification of established facts. One of such opportunities can be combination of demographic data from official statistics with Volunteered Geographic Information to discover interesting patterns. As demographic data describes attributes of population which is spread across locations or areas, the space in general, it seems reasonable to analyse this data in a spatial context.

This project will combine freely available data from official and crowd-sourced data sets in both linked and tabular formats to analyse correlation between pupils' educational achievements in England as provided by the Department for Education in the UK and availability of free time facilities in individual areas.

For each entity from the educational performance data facilities of various types will be queried against OpenStreetMap project geographical data to find those in a close proximity. The facilities will be counted for overall count as well as grouped per type to find combinations with reasonable correlation relationships. A scoring mechanism will be applied to discriminate individual facilities based on relevancy of their type and actual distance.

To support the project the community centres and primary schools across UK will be contacted to take part in an online survey to gain understanding of after-school and other free time activity opportunities for children in their corresponding local areas, challenges their communities face regarding children and also to help building the scoring mechanism described above.

1.2 Research Aims and Objectives

The aim of this project is to support current research in the area of Semantic Web in its claims of linked data usability and potentials to facilitate information discovery and inference while also including additional dimension with crowd-sourced data in form of Volunteered Geographic Information in the analyses.

This will be done by investigating correlation between availability of free time facilities for pupils and their educational achievements while combining data from Department for Education in UK with data collected as part of the open-source project OpenStreetMap and its linked version available via LinkedGeoData portal.

It will be evaluated in terms of ability to combine information from multiple data sets using already created links between them on the Semantic Web. Additionally the evaluation will be done by assessing the correlation found using standard statistical methods.

The research objectives are as follows:

- Review linked data research focusing on geospatial linked data and analysis,
- Review research about Volunteered Geographic Information,
- Review social research in terms of educational performance and free time facilities,
- Acquire data from the selected sources and reformat it to suit analytical purposes,
- Survey community centres in the UK about after-school and other free time activity opportunities for children in their local area,
- Undertake analysis of correlation between availability of free time facilities for pupils and their educational achievements,
- Assess how linked data contributes to the analytical process.

1.3 Research Methods

The research will be focused on three areas, linked data, Volunteered Geographic Information and social domain.

The linked data area of research will cover some technologies of the Semantic Web, analytical application of linked data and geospatial linked data and sources.

The Volunteered Geographic Information research will be reviewed in terms of challenges in using this type of data and available projects suitable as data source for purposes of this project.

The social research will be focused on effects of free time activities and facilities on pupils' educational achievements.

1.4 Scope and Limitations

The scope of the project is to acquire educational performance data as well as relevant geographical data for the area of England, UK based on the entities from the educational data.

Not only will the overall geographical area be scoped but also spatial queries for neighbouring objects so as to limit the number of the objects retrieved from the map data source. A radius query for as little as 2 or 3 km can yield amount of data which can impose data processing overhead going beyond the scope of this project.

The educational performance data will be limited only to Key Stage 2 level of primary education and the year 2013.

1.5 Organisation of Dissertation

The rest of the dissertation is organised into seven chapters as described below.

Chapters 2, 3, 4 each provide literature review in areas of linked data, Volunteered Geographic Information and social research respectively.

Chapter 5 outlines the design of the experiment along with scope, data sources description and methodologies of evaluation.

Chapter 6 describes the implementation of the experiment including individual steps taken and their results.

Chapter 7 evaluates outcomes of the experiment as well as the experiment itself and discusses survey conducted as part of the experiment.

Chapter 8 concludes the dissertation with brief summary of the research, experiment and evaluation describing contribution to body of knowledge and future work.

2. LINKED DATA

2.1 Introduction

This chapter provides short introduction into linked data in the context of Semantic Web and reviews literature focusing on practical applications of linked data, especially of the open datasets, and its use for analytical purposes. Separate section is devoted to geospatial linked data as spatial context is an important dimension to any data describing entities placed in the world. Potential data sources from the linked data domain are identified for purposes of this project.

2.2 Linked Data

The well-known World Wide Web (Web) has been present for over 20 years. As a system of interlinked documents it was primarily designed to be readable by humans. Over the years it became desirable to enable machines to process the increasing amount of data on the Web meaningfully. Berners-Lee et al. (2001) introduced the term Semantic Web providing a vision how the Web should evolve. They described the Semantic Web as an extension of the current Web where data is given a 'well-defined meaning' and structured into collections. Such data can then be interlinked allowing computer programs to interpret its meaning and perform reasoning.

The interlinking of data led to definition of a linked data paradigm which refers to data described in a structured way based on its meaning and pointing to other related data. The relations between data are defined by ontologies. Chen et al. (2012) note that the novelty of the Semantic Web is to build 'Global Linked Data Space'. They also explain four principles suggested by Tim Berners-Lee, the inventor of Web and one of the figures behind the idea of the Semantic Web, as how to build the linked data and thus to make this global data space to grow:

- Use URIs (Uniform Resource Identifiers) to identify things,
- Use HTTP URIs so that these things can be referred to and de-referenced by both people and intelligent agents,

- Publish machine-understandable information about the things when their URIs are de-referenced, using standard Semantic Web languages,
- Include links to other related URI-identified things in other places on the Web to improve discovery of related information.

Following these principles allows for uniquely identifying entities on a global scale, retrieving the data over the internet and freedom in creating links between diverse data sets.

Since the presentation of the Semantic Web idea and later the rules for the linked data there has been a continuous effort in diverse communities to encode data used by them as RDF triples and interlinking them with other data sets. World Wide Web Consortium (W3C) plays an important role in these endeavours by defining necessary standards for the Semantic Web.

Many state institutions across the globe started making available their sets of publically collected data as linked data sets which initiated a Linked Open Data (LOD) projects. Among the projects is LOD2[3] which is supported by the European Commission and represented by the leading Linked Open Data technologies researchers, companies and service providers from EU states and also from Korea. The project amongst others aims to develop tools, methodologies, algorithms and standards for exposing, managing, interlinking, searching, authoring etc. large amounts of information on the Semantic Web.

One of the significant applications of linked data is project DBpedia[4]. Bizer et al. (2009) describe it as 'a community effort to extract structured information from Wikipedia and to make this information accessible on the Web' (p.154). The DBpedia project extracts structured information from Wikipedia creating a knowledge base covering various domains with description of several millions of entities and with links to other Web pages as well as RDF links to other Web data sources. As other data publishers started setting RDF links to this knowledge base the DBpedia became one of the central interlinking hubs on the Semantic Web.

[3] http://lod2.eu

[4] http://dbpedia.org

The popularity of linked data grows among institutions and organisations from diverse areas of interest. Zapilko et al. (2011) note that linked data not only supports but also encourages publishing of data and even governments worldwide recognise its potential as it helps them to be more transparent and engage citizen participation. It is advantageous for both data providers and users. Providers can add value to their data by creating links to relevant data in other datasets. For developers it opens up many opportunities in terms of building applications and tools exploiting the data which ultimately benefits users. Although the linked data has still to be transformed for use with standard analysing tools the ongoing standardisation will allow for automating such processes. Regarding analysis of diverse data sets they further argue that it is important for data from these sources to be 'in the same data format' as well as '… on the same level of aggregation' (p.5). Because that is usually not the case pre-processing steps are required prior to combining and analysing the data.

Data analytics is concerned with extracting information from raw data, finding meaningful patterns and getting useful insights. It is used in academia, by state institutions and in industry among others to explore new trends and features, improving policies as well as increasing performance or obtaining competitive advantage. In this context it is unquestionable that data published by linked data principles provides valuable source to be exploited by data analytics.

Many governmental institutions around the world already started publishing by them held data under the open governmental initiatives but as Kalampokis et al. (2013) argue the availability of such data has still not been used to its full potential. They explain the true value lies in combining data from multiple statistical data sets as it has potential to unveil previously unprecedented insights. Their work gives an example of performing data analysis on linked open government data, though the data was transformed to RDF as part of the project, and suggests potential values for users. From the high level perspective the publishing of data is perceived as yielding advantages through novel reuse of data but the actual low level data access and processing from multiple sources are very complex and complicated.

The Semantic Web proves to be an optimal platform for providing data in linked format on global scale. The current trends though are to publish data in RDF and establishing valid links between data sets while there is not much focus on information inference (Hausenblas 2009) which is of great value regarding data analytics efforts.

Shadbolt et al. (2012) experienced many issues when integrating UK open government data into the Semantic Web based on linked data principles. Meta-data at national level was much more comprehensive than at local levels but in all cases not very understandable for layman. They recommend modifying the meta-data in future to better facilitate public reuse of data. They also recognised challenges in identifying so called ‘joint points’ when interlinking heterogeneous data sets. Geographies are among the strong candidates for this purpose but pose a problem in UK context where Parliamentary constituencies do not align well with administrative geographies. In their approach they inferred geographical containment through establishing if a constituency was fully within a county and thus were able to provide user with relevant data while also using data from other sources on the Semantic Web (see Figure 2.1).

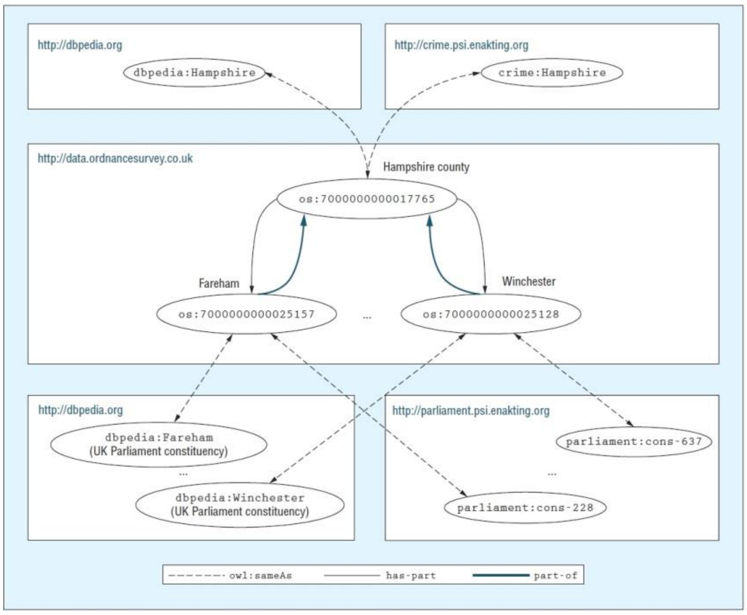


Figure 2.1: Relations for inference of spatial containment in RDF

To demonstrate and enable people without necessary knowledge about linked data to perform comparison of statistical data in linked format coming from various sources on the Semantic Web Capadisli et al. (2013) developed Web service supported by a simple user interface capable of data retrieval using federated queries in SPARQL query language for RDF. The user selects independent and dependent variables along with required time period and the system returns regression analysis including one or more scatter plots. The variables for the analysis are macro observations from data sets like World Bank, Transparency International etc. representing particular attributes of individual countries. To cater for multi-dimensionality of statistical data they employed RDF Data Cube vocabulary which is currently only a W3C recommendation candidate but has already been used in other projects due to its compatibility with Statistical Data and Metadata exchange standard popular among organisations.

2.3 Geospatial Linked Data

Data with some spatial or geographical attributes requires special attention as it presents an additional overhead to the processing but nevertheless most of the data represents entities placed in space and time which makes the geospatial attributes vital parts of analysis. Regarding data published by governmental organisations in many cases it relates to certain place or area. Janowicz et al. (2012) argue that many sets of linked data include some spatiotemporal identifier or point to data sets with such identifiers. They also highlight the fact that semantics and ontologies for geospatial data have been researched for over 20 years and the community behind the research was amongst the first to adopt the Semantic Web. Spatial data had always come from different sources due to the diverse methods of its collection. That made its processing and analysis from multiple sources a big challenge. The Semantic web thus promises a platform to link the geospatial data and allow for more meaningful analysis results.

The differentiating factor in terms of Semantic Web is a need for specific, geospatial, ontologies. Egenhofer (2002) argues that approaches of the Semantic Web do not explicitly capture some basic geospatial properties like entities and relationships required by spatial queries. He also notes that 'the enormous variety of encodings of geospatial semantics makes it particularly challenging to process requests for geospatial information' (p.1). In his paper he introduces the term Semantic Geospatial Web and suggests necessity of multiple spatial and terminological ontologies development for it to allow geospatial query results evaluation and thus more precise data retrieval.

Various governmental organizations while publishing by them managed data sets as RDF triples also include geospatial attributes where appropriate thus enriching the Linked geospatial data space. Other important sources of spatial data are GeoNames[5]

[5] http://www.geonames.org

and OpenStreetMap[6]. The Geonames is a geographical database containing over 10 million geographical names in various languages that is published as Linked data free of charge. GeoNames has already become an important interlinking hub on the Semantic Web. The OpenStreetMap (OSM) is a project built by community of mappers and providing open data about geographical features from around world. It serves primarily to render different maps in form of map tile images but there are initiatives to make the OSM data available as Linked geospatial data.

Auer et al. (2009) present project LinkedGeoData[7] in which they transformed OSM data based on the RDF model, interlinked the data with other spatial data sets and made the data accessible for machines according to the linked data paradigm as well as for humans by means of a faceted geo-data browser. The browser was developed as part of the project.

They argue that OSM data, which adheres to relatively simple model for representing spatial features, contains significant amount of information in so called OSM tags that is not presented on the maps. The tags are key : value pairs storing additional information (e.g. type of roads, opening hours, etc.) along with the spatial features. The OSM data transformation into linked geospatial data and connections to other data sets on the Semantic Web thus enable exploitation of this valuable spatial data set.

LinkedGeoData is potentially very interesting data set with all the accompanying technologies to query it. It should allow for exploiting of information stored in the OSM tags while bringing in data about the individual OSM objects from other linked data sets as long as these objects have their external links correctly established.

2.4 Conclusions

This chapter introduced linked data and looked at some projects and applications where it was successfully deployed. Although most of the research is still focused on transformations of data into RDF there is growing interest in practical usage of the data and potential use cases for data analytics.

[6] http://www.openstreetmap.org
[7] http://linkedgeodata.org

DBpedia and LinkedGeoData were recognised as potentially valuable for this project. The former is important for its role of a central hub on the Semantic Web and the latter for providing free and open geographical data in the linked form.

The LinkedGeoData makes Volunteered Geographic Information from the project OpenStreetMap available and thus interlinked into the Semantic Web. Both, the Volunteered Geographic Information and the OpenStreetMap, are discussed in detail in the next chapter.

3. VOLUNTEERED GEOGRAPHIC INFORMATION AND OPENSTREETMAP

3.1 Introduction

Phenomenon of Volunteered Geographic Information (VGI) is presented in this chapter along with the challenges it brings about and its increasing importance in supplementing geographic information generated by professional agencies. Project OpenStreetMap (OSM) is introduced as an example of deploying VGI on a global scale and being one of the main data sources for purposes of this project.

3.2 Volunteered Geographic Information

Volunteered Geographic Information was termed by Goodchild (2007) and corresponds to the location related information provided voluntarily by members of public like in case of OSM. He discusses how VGI can enrich conventional sources of information and its importance in time when national mapping agencies cannot afford thorough mapping of the entire countries by means of qualified field mappers as well as limitation of satellite imagery to describe certain local conditions and phenomena.

Voluntary contribution of spatial data became very popular due to its increasing accessibility with the introduction of WEB 2.0 and continuous widespread of GPS and mobile phone network technologies. Coleman *et al.* (2009) by identifying also Wiki technologies analyse motivations of volunteer contributors and a shift of geographical data type production from the professional bodies to communities of such contributors. Among other motivations altruism and willingness to present own areas without expecting any direct gains are discussed. At the same time they recognise problems with reliability and credibility of contributions that need to be taken into account when dealing with this type of information.

Karam and Melchiori (2013) highlight two problems, ‘variable quality’ and ‘description conflicts’, especially related to points of interests (POI) in context of volunteer contributions. They note moderator-based approach to approval process for new entries and updates being adopted by e.g. OSM which takes into account local knowledge possessed by community of users. As part of their research a framework was developed to improve correctness of geo-spatial data in area of linked data by ranking user corrections. Another approach is proposed by Mülligann et al. (2011) where to support contributors as well as editors in assigning correct tags to POIs in a consistent manner point pattern analysis is combined with semantic similarity measure.

Inconsistencies in VGI are unavoidable due to the sheer amount of contributors and better than disregarding them they should be embraced. Whether introduced by new contributions, edits or a contributor’s habits changing over time it is necessary to recognise them. It is a trade-of between limited amounts of data being generated following strict rules and opportunity of generating mass amounts of data from diverse origins. The use of VGI needs to be considered on individual case base and any generalisation rather be avoided (Mooney *et al.* 2012). Knowing the limitations helps to exploit this valuable source of data to its full extent.

3.3 OpenStreetMap

OpenStreetMap project[8] is one of the use cases of VGI on a global scale. It is a free and editable map allowing volunteers who join the OSM community to contribute and develop its content under an open license[9]. Haklay and Weber (2008) recognise OSM as the most extensive and effective among comparable VGI projects. Although an arguable statement the global coverage and the amount of contributions make OSM a very valuable and significant source of geographic information.

Some research projects assess quality of OSM data by comparing it to data from national geographic agencies or against other competing providers. Haklay (2010) compares OSM coverage for England against the Ordnance Survey UK (OS) data. He identifies places where due to the commitment of small group of contributors the OSM dataset achieved accuracy comparable with that of OS. As problematic are recognised places where data was inaccurate or missing altogether due to the low interest of mapping them.

Ciepłuch *et al.* (2010) conducted comparison of OSM data for Ireland with major competing providers from commercial domain, namely Google maps and Bing maps. Focusing on five particular areas their findings show that inconsistencies are present in

[8] http://www.openstreetmap.org
[9] http://wiki.openstreetmap.org/wiki/About

commercial products as well and none of the three data sets shows consistent accuracy. Regarding OSM they discovered loose connection between number of mappers for an area and the resulting accuracy. The most salient shortcomings are in terms of areas of low interest to be mapped which gives an advantage to the commercial products.

In more recent analysis focused on objects with 15 and more edits in Ireland and UK Mooney and Corcoran (2012) investigated OSM historical data rather than just the most recent one as a way of identifying quality issues. They found that 87% of the OSM objects under study were contributed or updated by only 11% of contributors which is in line with the fact that most contributors get disengaged after few initial contributions as reported by Haklay and Weber (2008).

The OSM data provides an additional feature in form of so called tags, key : value pairs, that allows for including contextual, not necessarily geographical, information with each OSM object. Examples include opening hours of shops, purposes of buildings etc. Information in these tags is recognised as of high value especially in terms of analysis as it is not rendered on the map images provided by the OSM portal (Haklay and Weber 2008, Auer *et al.* 2009).

3.4 Conclusions

The Volunteered Geographic Information and challenges when dealing with this type of data were presented in this chapter. Inconsistencies of VGI are repeatedly mentioned across the board in the related research and so it is a problem present in OSM data. Despite the challenges OSM provides a valuable data source and taking into account the additional information encoded in the OSM tags there are open possibilities how to exploit the data for analytical purposes.

The linked data version of OSM data, the LinkedGeoData project, described in chapter 2 is a promising enhancement allowing for further data enrichment with additional information from other linked data sources available on the Semantic Web. However it is important that there are external links established from the LinkedGeoData to the rest of the Semantic Web and number of objects with such links is sufficient for any geographical area to be investigated.

4. EDUCATIONAL PERFORMANCE AND FREE TIME FACILITIES

4.1 Introduction

This chapter provides a review of articles and publications from the social domain as well as research focused on free time and activities of children and adolescents. It ties back to the educational performance where possible but mostly outlines impacts on general development in childhood and later in life.

4.2 Free Time Activities

Reports and studies from social research suggest for example worse educational attainments of young people not engaged in extra-curricular activities (e.g. Eccles and Barber 1999, ‘World Youth Report’ 2003) and significantly higher chance of school dropout due to the same reasons (Wegner *et al.* 2008). It seems common sense that by guiding children and engaging them in particular activities their development can be affected in favour of traits perceived as ‘good’ for them to ensure their well-being throughout their lives.

Osgood *et al.* (2005) express opinion of less support by general public for youth taking care of themselves for prolonged time. They suggest that activities the children and adolescents are engaged after school have very significant impact on at least some of them. They further discuss nature of unstructured activities in context of child development and contrast them with the structured ones. Structured activities are described as being either public, e.g. schools, community centres etc. or private like music lessons but all under supervision of adults. On the other hand facilities like community centres with drop-in sessions are categorised as unstructured which seems reasonable. Among others they conclude that although spending time in unstructured activities for children from particular social classes is not the only problem for such groups, but it is associated with increased rates of problematic behaviour.

Fuligni and Stevenson (1995) address difficulties in assessing the way high school students spend their free time in terms of long time spent studying versus activities involving more socialising and interacting with peers. They compare the free time use with achievements in mathematics test in age group 16 – 17 years across multiple cultural environments. In general their findings show positive correlations (coefficients between 0.22 and 0.36 across the locations) between studying and the test results. Negative correlations (coefficients between -0.17 and -0.34 across the locations) are reported for activities like being with friends, watching television etc.

An opposite argument to the advantage of children’s involvement in structured activities is provided by Barker *et al.* (2014). Their study shows that involving children in ‘less-structured’ activities positively relates to their performance of executive functions. These functions represent processes of controlling thinking and actions toward goal oriented behaviour. Although their experiment is focused on 6 years old children the implications are valid for later age as well. Regarding potential facility types to be included when looking for correlations with educational performance in this project, the study supports merit of facilities with not fully structured activities. Such facilities are a valid counterpart to facilities where children have their time organised by a tutor.

4.3 Free Time and Leisure Facilities

Closely related to the free time activities are facilities providing space where children have adequate opportunities to participate in the activities. Whether man-made or of a natural (outdoor) character the more of such spaces are available locally the better chance for children’s engagement.

But a presence of facilities does not automatically translate into their availability. Byrne *et al.* (2006) study needs of young people in disadvantaged communities in Ireland. They recognise importance of leisure time choices adolescents make on their adulthood lifestyles. The challenges highlighted in their study included restrictions especially in rural areas. Either there were only distant facilities available which required use of limited public transport services to reach them or in new development areas the relevant facilities were privately owned and the young people from the disadvantaged communities were banned from accessing them (p.7).

Van der Merwe (2014) presents provisioning challenges regarding extracurricular programmes and activities in inner-city of Gauteng, South Africa. The ideal of sufficient amount of dedicated facilities expressed by heads of schools from the studied areas was challenged by allocating finances from schools’ scant budgets and high demand for the limited amount of facilities. Nevertheless the participation of learners in the programmes in such constrained conditions was still found beneficial in terms of skills gained and in this particular case with improved school retention.

4.4 Conclusions

There seems to be a lot of support in favour of involving children and adolescents in structured and supervised activities present in the research studies. But there is also disagreement highlighting importance of unstructured activities on development of self-directed and goal focused traits in children.

Not as much attention has been geared in research toward facilities serving as space for the free time activities provision and potential impact of their availability in local areas on educational performance. The advantages of facilities availability as a supplement for positive development of children by their engagement in meaningful activities are however obvious.

Building up from the literature review chapters the next chapter will present design of an experiment to investigate correlation between availability of free time facilities for pupils and their educational achievements while exploiting linked data and volunteered geographic information.

5. EXPERIMENT DESIGN

5.1 Introduction

This chapter presents design of the experiment to investigate correlation between availability of free time facilities for pupils and their educational achievements while exploiting linked data and volunteered geographic information which is the main focus of the research project.

First the scope and geographical area of interest are defined. The potential data sources and technologies identified in the previous chapters are then investigated for their content and suitability to be included in the design of the actual experiment.

Evaluation methods for experiment results or other approaches to assess suitability of experiment components and steps are discussed here. Additional means like survey for community centres and primary schools are presented as well.

5.2 High Level Experiment Outline

Data for the experiment was assumed to be characterised by two main groups. One group to be educational performance data for selected geographical area at the lowest possible level of aggregation. The second group to be geographical data for the area containing as many objects as possible that could be considered as facilities relevant for children’s free time activities.

Educational performance data possibly at individual schools level or any reasonably small administrative area will constitute one of the two variables for correlation calculations. If multiple performance measures are available the correlation will be calculated for each of them separately.

The geographical data will need to be queried to retrieve facilities for each of the entities from the educational data (school, area). The retrieved objects, the facilities, in predefined proximity of each educational entity will be each scored in terms of their relevancy for children’s free time activities and distance. For the purpose of determining the relevancy of individual facility types a survey will be conducted among community centres and relevant educational institutions.

After applying the scoring, i.e. multiplying by score in the interval 0 to 1, the objects will be counted per single facility type or summed across multiple types to constitute overall count per each educational entity and to be used as the second variable for the correlation calculation.

The assumed advantage of using geographical linked data will be ability to follow any relevant external links from each facility object to other data sources on the Semantic Web and investigate if the additional data can improve the resulting correlation calculation. For example if the additional data could help to establish if a community centre actually provides services for children rather than only for adults the score for such facility could be adjusted accordingly to better represent reality and ultimately calculate correlation coefficient that better reflects the reality.

Whether the data would be available via bulk download or required to be retrieved via separate requests / queries against particular interfaces over the internet, once downloaded it will need to be transformed to a format suitable for analysis.

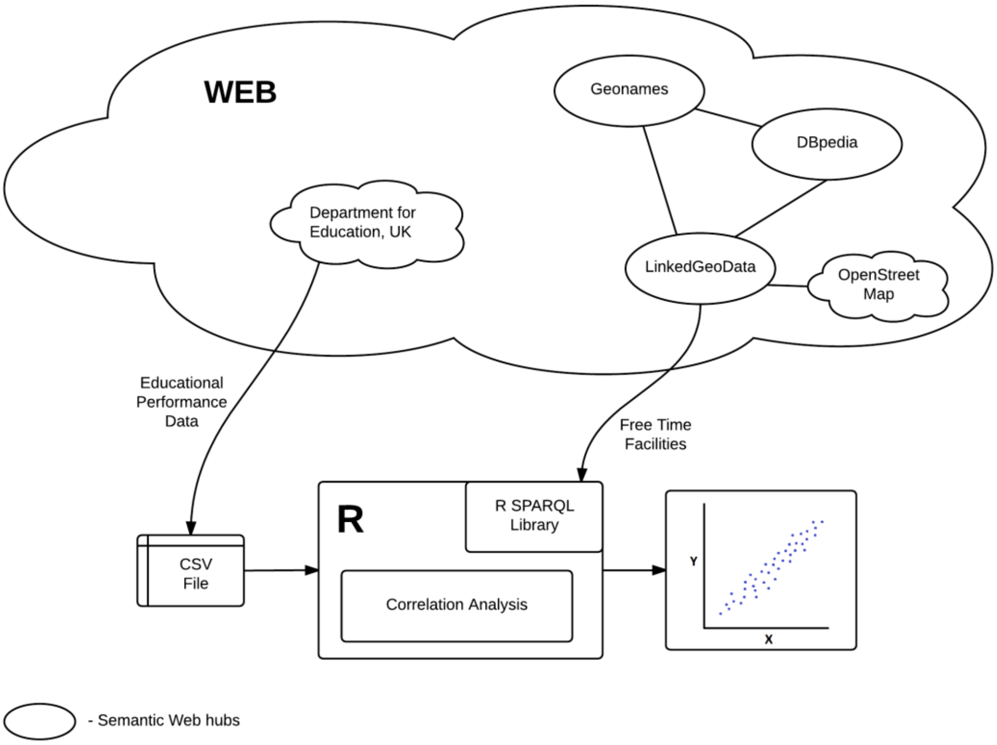


Figure 5.1: Experiment – high level design

The possible correlations will be examined in different combinations. There will be generally three factors from which the combinations will be derived:

- Individual educational performance measures,
- Individual facility types or mixture of various types,
- Geographical groupings at levels like local areas, regions, state.

As the number of possible combinations will be most likely very high only subsets within each of the three factors might be chosen for investigation.

Each correlation calculation will be accompanied by statistical test for its significance under null hypothesis of correlation coefficient being equal to 0.

Multiple regression may also by applied in an attempt to determine particular facility types as the strongest available predictors of certain educational performance measures.

If any correlations of individual combinations are found to be strong enough and statistically significant such combinations could be used to for example recommend local councils investments prioritisation for particular types of facilities. Such recommendation is out of the scope of this project. Simple comparisons between for example smaller geographical areas could be performed though to see if differences in available facility types relate in some way to different educational performance results.

5.3 Data

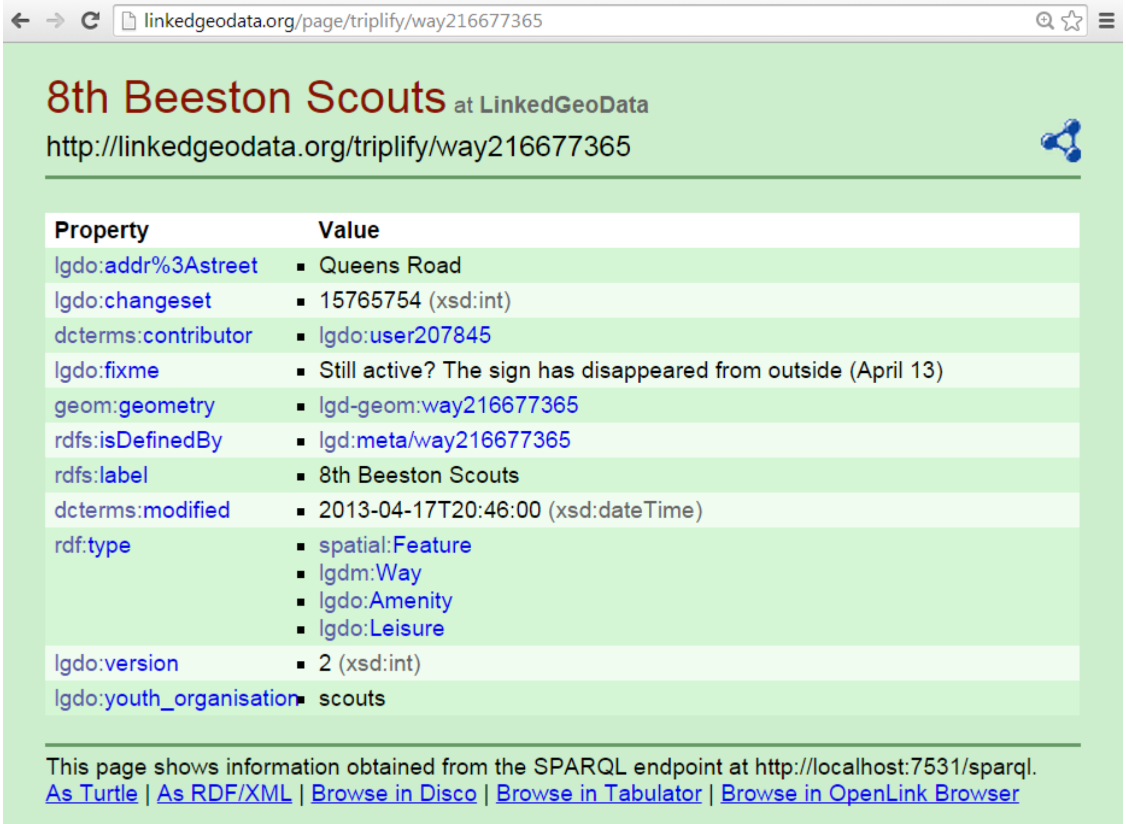
To limit the scope of the experiment the geographical area of England in the UK was chosen. Preliminary investigation discovered suitable data sets with aggregated data about education performance of pupils provided by Department for Education available under Open Government Licence[10].

The open geographical data for England is available from the OpenStreetMap[11] project and from LinkedGeoData[12] both under the Open Database Licence.

[10] https://www.gov.uk/help/terms-conditions
[11] http://www.openstreetmap.org/copyright
[12] http://linkedgeodata.org/Datasets?v=sln

5.3.1 LinkedGeoData

The LinkedGeoData is stored in RDF triples using Virtuoso technology. There are two ways to access the data programmatically, SPARQL end point[13] to issue SPARQL queries against the data set or a REST API. The available data format of the response ranges from usual XML and JSON formats to RDF specific formats like RDF/XML and Turtle. The figures below show examples of the formats.



linkedgeodata.org/page/triplify/way216677365

8th Beeston Scouts at LinkedGeoData

http://linkedgeodata.org/triplify/way216677365

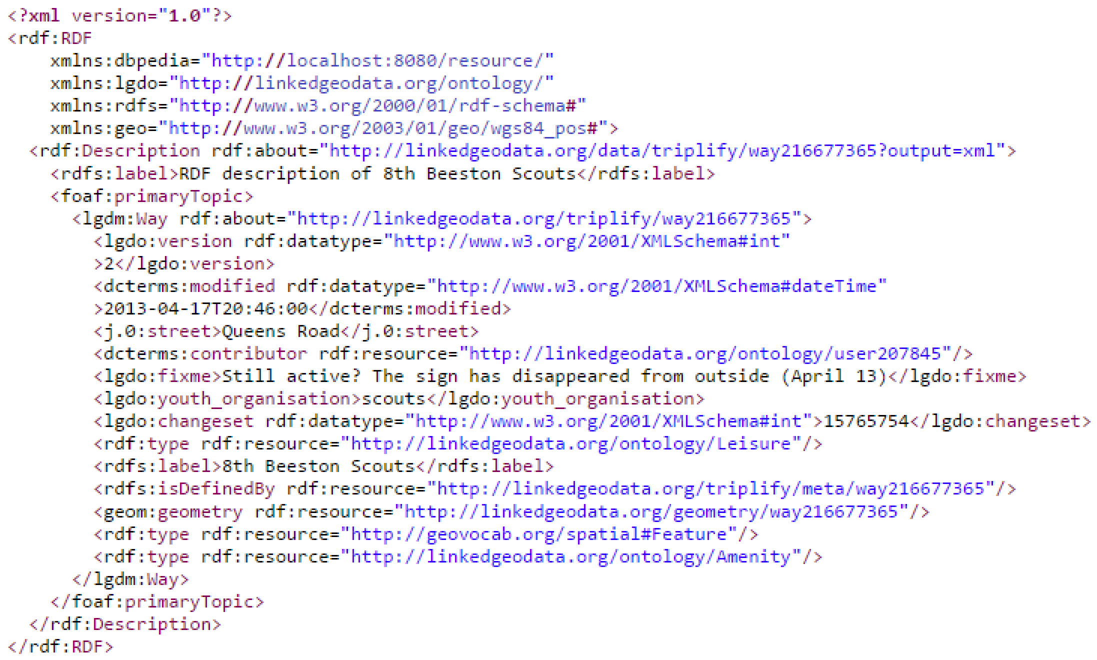
Property	Value
lgdo:addr%3Astreet	▪ Queens Road
lgdo:changeset	▪ 15765754 (xsd:int)
dcterms:contributor	▪ lgdo:user207845
lgdo:fixme	▪ Still active? The sign has disappeared from outside (April 13)
geom:geometry	▪ lgd-geom:way216677365
rdfs:isDefinedBy	▪ lgd:meta/way216677365
rdfs:label	▪ 8th Beeston Scouts
dcterms:modified	▪ 2013-04-17T20:46:00 (xsd:dateTime)
rdf:type	▪ spatial:Feature ▪ lgdm:Way ▪ lgdo:Amenity ▪ lgdo:Leisure
lgdo:version	▪ 2 (xsd:int)
lgdo:youth_organisation	▪ scouts

This page shows information obtained from the SPARQL endpoint at http://localhost:7531/sparql.
As Turtle | As RDF/XML | Browse in Disco | Browse in Tabulator | Browse in OpenLink Browser

Figure 5.2: Nottingham youth facility – linked data in HTML format

Spatial query types supported by both interfaces are limited to only radius or rectangular area. Geographical coordinates are expected to be in WGS84 coordinate system. For the radius queries longitude and latitude of the desired point and size of the radius in km are required. In case of the rectangular area query type south-west and north-east coordinates are required to define the rectangle size.

[13] http://linkedgeodata.org/sparql



```
<?xml version="1.0"?>
<rdf:RDF
    xmlns:dbpedia="http://localhost:8080/resource/"
    xmlns:lgdo="http://linkedgeodata.org/ontology/"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:geo="http://www.w3.org/2003/01/geo/wgs84_pos#">
  <rdf:Description rdf:about="http://linkedgeodata.org/data/triplify/way216677365?output=xml">
    <rdfs:label>RDF description of 8th Beeston Scouts</rdfs:label>
    <foaf:primaryTopic>
      <lgdm:Way rdf:about="http://linkedgeodata.org/triplify/way216677365">
        <lgdo:version rdf:datatype="http://www.w3.org/2001/XMLSchema#int"
        >2</lgdo:version>
        <dcterms:modified rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime"
        >2013-04-17T20:46:00</dcterms:modified>
        <j.0:street>Queens Road</j.0:street>
        <dcterms:contributor rdf:resource="http://linkedgeodata.org/ontology/user207845"/>
        <lgdo:fixme>Still active? The sign has disappeared from outside (April 13)</lgdo:fixme>
        <lgdo:youth_organisation>scouts</lgdo:youth_organisation>
        <lgdo:changeset rdf:datatype="http://www.w3.org/2001/XMLSchema#int">15765754</lgdo:changeset>
        <rdf:type rdf:resource="http://linkedgeodata.org/ontology/Leisure"/>
        <rdfs:label>8th Beeston Scouts</rdfs:label>
        <rdfs:isDefinedBy rdf:resource="http://linkedgeodata.org/triplify/meta/way216677365"/>
        <geom:geometry rdf:resource="http://linkedgeodata.org/geometry/way216677365"/>
        <rdf:type rdf:resource="http://geovocab.org/spatial#Feature"/>
        <rdf:type rdf:resource="http://linkedgeodata.org/ontology/Amenity"/>
      </lgdm:Way>
    </foaf:primaryTopic>
  </rdf:Description>
</rdf:RDF>
```

Figure 5.3: Nottingham youth facility – linked data in RDF/XML format

For either type of interface separate requests with queries will have to be issued for each individual entity from the educational data set. Potentially useful external links to other data sources on the Semantic Web will need to be examined for any additional data about the retrieved facility objects. For this purpose Uniform Resource Identifiers for predicates and objects of individual facilities can be used.

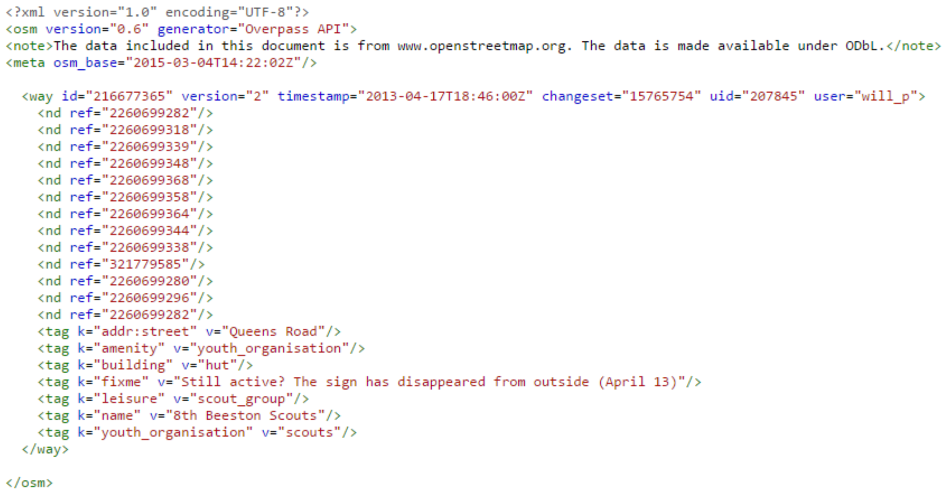
5.3.2 OpenStreetMap data

There are multiple interfaces to access the OpenStreetMap (OSM) data depending on the purpose. Most of them are independent open-source projects. They provide services like bulk downloads (GeoFabrik[14]), reverse geocoding (OSM Nominatim[15]), querying the map data (overpass API[16], overpass turbo[17]) etc.

The overpass turbo provides graphical WEB interface where query results can be overlaid on base map image and as such comprises a useful tool for inspection by humans. The overpass interfaces uses XML query format or specific query language OverpassQL[18]. It supports certain spatial queries and allows among others for use of

[14] http://www.geofabrik.de/
[15] http://wiki.openstreetmap.org/wiki/Nominatim
[16] http://overpass-api.de/
[17] http://overpass-turbo.eu/
[18] http://wiki.openstreetmap.org/wiki/Overpass_API/Language_Guide

regular expressions which could be advantageous for querying multiple facility types at once. Available download formats include XML and JSON.

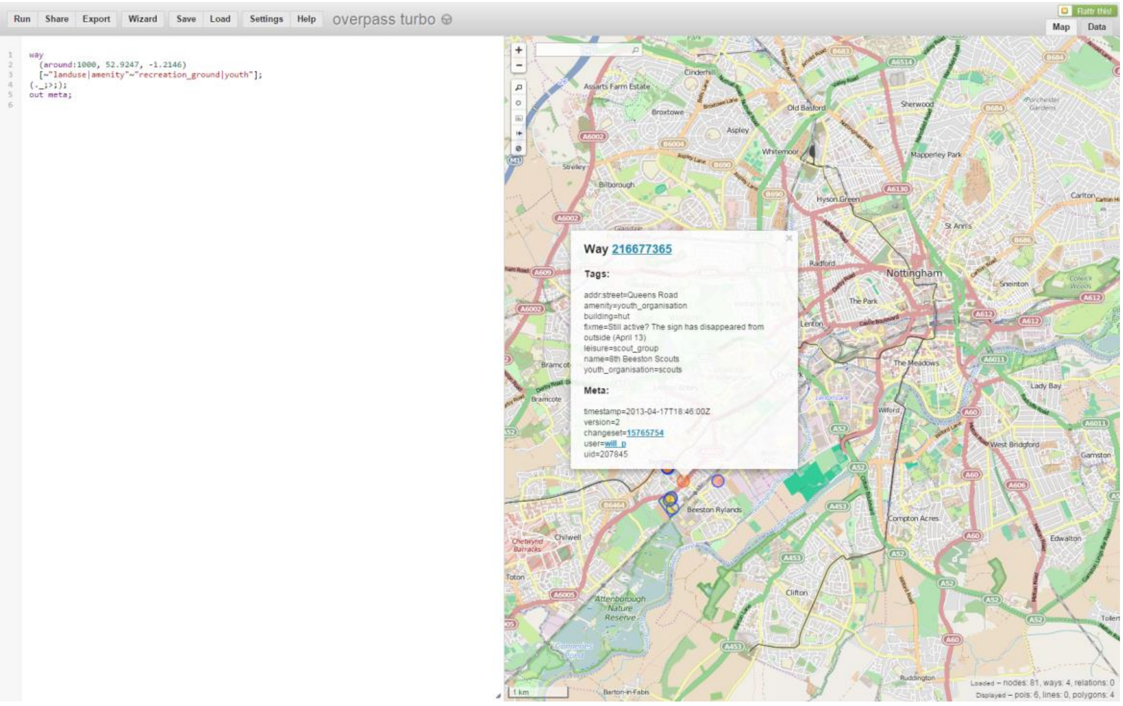


```
<?xml version="1.0" encoding="UTF-8"?>
<osm version="0.6" generator="Overpass API">
<note>The data included in this document is from www.openstreetmap.org. The data is made available under ODbL.</note>
<meta osm_base="2015-03-04T14:22:02Z"/>

  <way id="216677365" version="2" timestamp="2013-04-17T18:46:00Z" changeset="15765754" uid="207845" user="will_p">
    <nd ref="2260699282"/>
    <nd ref="2260699318"/>
    <nd ref="2260699339"/>
    <nd ref="2260699348"/>
    <nd ref="2260699368"/>
    <nd ref="2260699358"/>
    <nd ref="2260699364"/>
    <nd ref="2260699344"/>
    <nd ref="2260699338"/>
    <nd ref="321779585"/>
    <nd ref="2260699280"/>
    <nd ref="2260699296"/>
    <nd ref="2260699282"/>
    <tag k="addr:street" v="Queens Road"/>
    <tag k="amenity" v="youth_organisation"/>
    <tag k="building" v="hut"/>
    <tag k="fixme" v="Still active? The sign has disappeared from outside (April 13)"/>
    <tag k="leisure" v="scout_group"/>
    <tag k="name" v="8th Beeston Scouts"/>
    <tag k="youth_organisation" v="scouts"/>
  </way>

</osm>
```

Figure 5.4: Nottingham youth facility – OSM data in XML format

The OSM data as it can be seen in the figure above contains various parameters about a single map object. In the example a youth organisation facility is represented as a way type which describes the grounds on which the facility is located. Each OSM object has its unique ID though this can change over time and it is recommended to distinguish objects based on their surroundings rather than rely on the ID. The metadata attributes contain among others time stamp, contributor identification and change set identifier which is used to track changes in the data over time. Because a way represents polygon geometrical shape individual points of the polygon are described by the 'nd' type.

Of vital importance for this project is type 'tag' which can in form of key : value pairs provide multitude of additional information about the object collected by the community of contributors with local knowledge. In the example provided the youth facility can be identified as relevant for purposes of this project by any of the amenity : youth_organisation, leisure : scout_group or youth_organisation : scouts pairs. On the other hand the multiple tagging approach allows for inconsistencies that may present a challenge when counting facilities by types as part of the experiment.



```
way
  (around:1000, 52.9247, -1.2146)
  [~"landuse|amenity"~"recreation_ground|youth"];
(._;>;);
out meta;
```

Way 216677365

Tags:

addr:street=Queens Road
amenity=youth_organisation
building=hut
fixme=Still active? The sign has disappeared from outside (April 13)
leisure=scout_group
name=8th Beeston Scouts
youth_organisation=scouts

Meta:

timestamp=2013-04-17T18:46:00Z
version=2
changeset=15765754
user=will_p
uid=207845

Figure 5.5: Overpass Turbo – OSM interface

Accessing OSM through these interfaces may prove necessary if the LinkedGeoData APIs will be found too limiting for purposes of the experiment.

5.3.3 Educational Performance Data for England

The educational performance data was decided to be limited to England and acquired from the performance tables provided by the Department for Education[19] (DfE) in the UK. The data is provided for previous years and for different levels of the primary and secondary education.

To set a scope for the experiment the data was narrowed down to only Key Stage 2 attainment measures, tests taken by pupils at the age of 10 to 11 years, and only to the school year ending in 2013. The age group was selected on the base of a rationale that children around this age are able and allowed to move around their neighbourhoods independently to some degree and thus making use of relevant free time facilities relatively close to their homes. Younger age groups might be limited to travel and transport by their guardians. On the other hand older age groups under reasonable assumption are more likely to travel further distances on their own and relevant

[19] http://www.education.gov.uk/schools/performance/

facilities for them may be beyond the planned facility distances to be investigated as part of the experiment.

The data is available for download in CSV format from the DfE web page[20]. There are also files available containing metadata and general details about each school among others. Three of these files will be used for the purposes of the experiment. The first explains educational measure name abbreviations. The second is for matching local authority area code assigned to each school with actual names and mapping them to higher level administrative units which are in this case represented by 25 regions in England. The third provides per school information in terms of address, contacts, type, age ranges etc. The only information relating to the geographical position of a school is UK post code. The post codes will have to be converted into suitable geographical reference system for querying the map data sets as part of the experiment.

There are about 270 attainment measures describing school performance in various groupings, e.g. by gender, previous stage attainments, fulfilling criteria of being disadvantaged etc. The data is in form of numeric score points, percentages of pupils achieving particular levels or ordinal values containing alphanumeric combinations representing grade levels. The approach to measure selection for purposes of the experiment is described in the next section.

5.4 Methodologies

5.4.1 Facility Type Selection

The OSM provides overview of tag key : value pairs used to describe map objects on their Map Features[21] wiki page. Tag keys represent categories like amenity, leisure, natural etc. each of which can have many different values. The values are accompanied by textual description along with an image example in some cases. The list is though not exhaustive in terms of values being in use.

In the selection process firstly relevant categories will be narrowed down and then relevant values in each category. During the exploration and test downloads phase of the experiment more values may be discovered as useful and included in the final

[20] http://www.education.gov.uk/schools/performance/2013/download_data.html
[21] http://wiki.openstreetmap.org/wiki/Map_Features

spatial query. As relevant will be considered places where reasonable assumption about structured as well as unstructured activities provision for the target age group can be made. These may include youth or community centres or less obvious facility types like sport centres, stadiums etc. Outdoor places like recreation grounds, forests etc. will also be considered as they provide space for play and peer interactions.

5.4.2 Educational Performance Measures Selection

The rationale behind selecting Key Stage 2 (KS2) performance data and focusing on the age group 10 to 11 years old is described in the previous section. The selection of individual performance measures will follow some of the DfE recommendations in terms of informative value the measures provide.

The DfE provides grouping of schools using a ‘Similar Schools’[22] measure enabling performance comparison based on similar intakes. This is an estimate calculated from the Key Stage 1 results aggregated for a school. Based on explanations in the guiding notes it seems appropriate to focus mainly on measures of progression within a stage rather than using measures reflecting single point in time. The progression measures are based on estimated number of pupils achieving certain level of attainment at KS2 where the estimation takes attainment level from previous stage into account. Additionally measures representing combination of skills (reading, writing and maths) rather than each of them separately seem to be used across the guiding examples from DfE.

5.4.3 Correlation Measurement

The correlation between individual educational performance measures and the number of facilities within particular distance from the school will be calculated using Pearson correlation coefficient.

For testing the significance of the calculated coefficient following hypotheses and significance level will be used:

- null hypothesis H_0: correlation coefficient = 0,
- alternative hypothesis H_a: correlation coefficient $\neq 0$,
- significance level of rejecting H_0 $\alpha = 0.05$.

[22] http://www.education.gov.uk/schools/performance/primary_13/p13.html

There will be multiple combinations of educational measures and facility types investigated for correlation relationship. Facility counts will be used as an overall facility count per school, individual facility type count and various groupings of facility types in which case the counts of individual types will be summed. The facility counts in facility type groupings will be investigated with and without the scoring applied.

5.4.4 Facility Scoring

The retrieved objects from OSM, the facilities, in predefined proximity of each school will be each assigned a score of 1 and then multiplied by predetermined scoring value 0 to 1 based on their type. For the purpose of determining relevancy of individual facility types for children's free time activities and setting their scoring values a survey will be conducted among community centres and relevant educational institutions.

Additionally the facilities will be discriminated based on their distance using inverse distance weight. For each facility its Euclidean distance from the school will be calculated. The inverse distance (1 / d) will then provide the distance score.

5.4.5 Linked Data Assessment

For this assessment the linked data version of retrieved OSM objects will need to be explored whether it provides any relevant external links to other data sources on the Semantic Web. Stadler *et al.* (2012) reported over 100,000 links to DBpedia and over 570,000 to GeoNames established from LinkedGeoData. They used 'sameAs', a property of the WEB Ontology Language, predicate for the external links being the objects in the created triples. This fact does not guarantee that such external links are available for many OSM objects located in England thus the assessment in the experiment will be approached as a two-step process.

First, the distinct OSM objects retrieved as part of the experiment will be explored for containing an external links. Due to the likely high number of objects retrieved a sample will be taken from objects belonging to the most prevailing facility types, those appearing among many schools. These objects will be checked for containing the 'sameAs' predicate and if found to be present in more than 50% of the objects the second step will be performed.

In the second step the external links will be followed using SPARQL queries to discover any additional information from other data sources on the Semantic Web that could provide better context for the OSM objects on top of the information encoded in the OSM tags. Such information will be ultimately used for adjusting the facility score of each object described in the previous section.

To measure the contribution of linked data the correlation coefficients will be calculated before and after adjusting the facility score and compared for any change in the strength of the relationships.

If on the other hand the step 1 shows insufficient external links present in the previously retrieved OSM objects the step 2 will be omitted.

5.4.6 Survey

An online survey was created to support design of the scoring mechanism for the facility types in terms of their relevancy for provision free time activities for children. The questions are available in Appendix A.

After establishing what facility types are present in the OSM data these will be categorised and participants will be able to score them on a scale from 1 to 5. For each question the scale numbers will be multiplied by the number of their responses individually, then summed and divided by the total number of responses. The acquired numbers will be converted to 0 to 1 scale.

Some of the questions are open-ended and will try to elicit knowledge about problems not only with provisioning of free time activities but also other problems on a more general level that local communities face regarding children’s development. It is assumed that this might help to recognise additional facility types to be included in the analysis and decide upon their scores.

The requests for survey participation will be sent to community centres and primary schools across England as well as Ireland. The email contacts of the institutions will be looked up on the internet.

5.5 Technologies

The main technology used will be statistical package R with all additional libraries necessary to download, process and analyse data from the previously identified sources. It provides data analytical platform with rich choice of libraries to communicate with remote servers over HTTP, to query remote SPARQL end points, spatial and geographical calculations, and database connectors among others. As a statistical package it contains functions for correlation and linear regression calculations. It also has wide spectrum of graphing and data visualisation options in form of static images as well as interactive outputs using WEB technologies.

A potential limitation may become the available amount of operating memory. As R session keeps all the data in memory it might not be possible to process all the ‘to be’ retrieved OSM objects at once. To overcome this limitation use of RDBMS or NoSQL databases may be considered in course of the experiment.

5.6 Conclusions

This chapter described how the experiment is designed, including steps to acquire data, approaches to its processing and analysis and suggested methods for comparison and evaluation of the results. The scope of the experiment was defined as working with geographical area of England and using school performance data for the Key Stage 2 from year 2013.

The necessary data sources and technologies for the experiment were identified with regard to the previous chapters. The linked data from LinkedGoeData source will be investigated for its suitability as part of the experiment and its impact on final findings will be evaluated in later chapters.

The experiment as it was performed is presented in the next chapter along with its results. The evaluation of results and discussion about problems encountered will be presented in later chapters.

6. EXPERIMENT IMPLEMENTATION AND RESULTS

6.1 Introduction

Individual steps of the project experiment are detailed in this chapter. The description includes approaches taken from the technical point of view, challenges experienced, application of methodologies outlined in the previous chapter as well as results achieved.

As the data acquisition required data retrieval from multiple sources, different cleaning and transformation steps are presented in a chronological order. Also details about technologies included in due course of the experiment are described.

6.2 School Performance Data from England

After downloading the performance data (england_ks2.csv) and corresponding metadata file (ks2_meta.csv) additional information about geographical location of each primary school was needed. This was available in two additional files (england_spine.csv, spine_meta.csv).

6.2.1 Geo Coordinates from Postal Codes

The Spine data about the schools contained only addresses with UK postal code which needed to be converted into geographical coordinates. The postal codes divide the country into quadrants size of which varies as it depends on the number of delivery points. Although not completely precise indicator for establishing exact position of schools misplacement within a quadrant was considered as acceptable in this experiment.

A file was located on the Office for National Statistics (ONS) Web site which mapped full length postal codes to easting and northing values of the UK coordinate system. This required yet additional conversion into WGS84 (longitude, latitude) coordinate system that is used by OSM. As no tool for batch conversion between these two coordinate systems was found an R script was designed to perform such conversion using functions from ‘rgdal’ library for spatial calculations.

6.2.2 Performance Data

The performance data from CSV file was consequently loaded into an R data frame (table like structure) and the coordinates attached based on the matching postal codes.

The values used for the performance measures were discussed in the previous chapter. Where a measure was not given it meant that either the school did not provide the results, or the value was suppressed due to small number of pupils in such school and to ensure that individuals cannot be recognised from the data. These missing/invalid values differed across the individual measures in the performance table.

6.3 Geographical Data

6.3.1 Download Testing in R

To query OSM data there were two interfaces considered, LinkedGeoData SPARQL end point for downloading linked data version of OSM data and Overpass for downloading the original data. Both of the interfaces were checked for their availability and ability to connect to from R using appropriate libraries. Both interfaces were found working as expected with a set of test queries.

6.3.2 OSM Tag Categories Selection

OSM tags are key : value pairs providing additional information about OSM objects. For clarity the tag key will be referred to as facility category and the tag value as facility type.

It was necessary to establish which combinations of facility category and type should be included in query when retrieving all relevant facilities for each school. The selection was done based on list available on Map Feature OSM wiki page. Six main categories were identified as relevant for free time facility types:

amenity, building, club, landuse, leisure, natural.

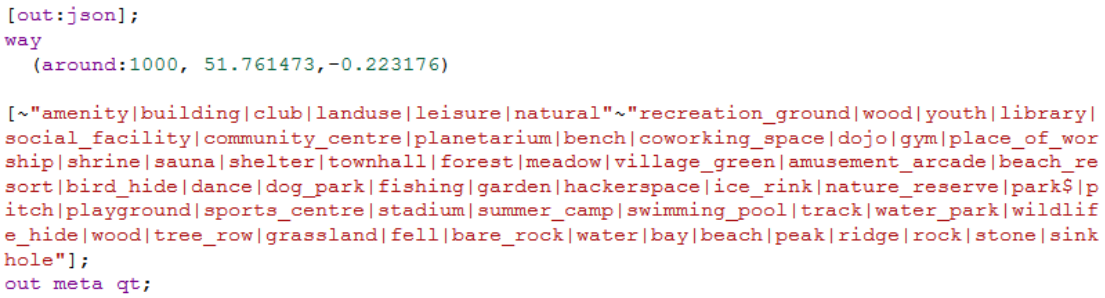
Altogether 53 types across these categories were chosen for inclusion in the final query. Full list can be found in the file ‘POIs.txt’.

The Overpass Turbo Web interface (see Figure 5.5 in the previous chapter) served as testing tool when deciding upon facility categories and types to include in the queries. It allowed for exploring the information stored with OSM objects as well as displaying the objects on a map in form of polygons.

At the time when the list of facility types for the query was finalised the LinkedGeoData SPARQL endpoint started to be randomly inaccessible. A decision was made to use the Overpass interface and download the OSM data rather than the linked form as planned in the experiment design. It did not pose much of an obstacle but for the project objective of assessing linked data contribution to the analysis it required additional queries for linked version of OSM objects once the SPARQL end point became available again.

6.3.3 OSM Data Retrieval

The list of selected facility categories and types was encoded into a query for the Overpass interface. The interface uses its own query language – OverpassQL.



```
[out:json];
way
  (around:1000, 51.761473,-0.223176)

[~"amenity|building|club|landuse|leisure|natural"~"recreation_ground|wood|youth|library|
social_facility|community_centre|planetarium|bench|coworking_space|dojo|gym|place_of_wor
ship|shrine|sauna|shelter|townhall|forest|meadow|village_green|amusement_arcade|beach_re
sort|bird_hide|dance|dog_park|fishing|garden|hackerspace|ice_rink|nature_reserve|park$|p
itch|playground|sports_centre|stadium|summer_camp|swimming_pool|track|water_park|wildlif
e_hide|wood|tree_row|grassland|fell|bare_rock|water|bay|beach|peak|ridge|rock|stone|sink
hole"];
out meta qt;
```

Figure 6.1: OverpassQL query

The query language allows for using regular expressions which enabled search for all desired facility types at once. The query in the figure above contains following information:

- Result in JSON format,
- Spatial query for OSM object of type ‘way’ within 1 km radius from point given by the latitude and longitude decimal numbers,
- REGEX ‘~’ for any combination of the 6 categories and 53 facility types,
- Result to contain also metadata.

The process of retrieval was automated building an R script that iteratively retrieved corresponding facilities for each of the over 18,000 relevant primary schools. Each of the retrieved OSM objects in JSON format was assigned URN unique identifier referring to school for which it was retrieved and eventually added to the final data frame. In this approach the resulting data frame became a sparse table as each of the tag keys contained in an object represented column name. At the end of this process the columns of the resulting data frame reflected all unique tag keys received.

The retrieval took over 4 days and on 6 occasions an error was received in the response from Overpass. Each of these subsets retrieved was saved as a separate CSV file to back up the already retrieved objects.

There were altogether 782,991 objects retrieved which accounted for 240,630 distinct objects based on their OSM IDs when checked later during the experiment.

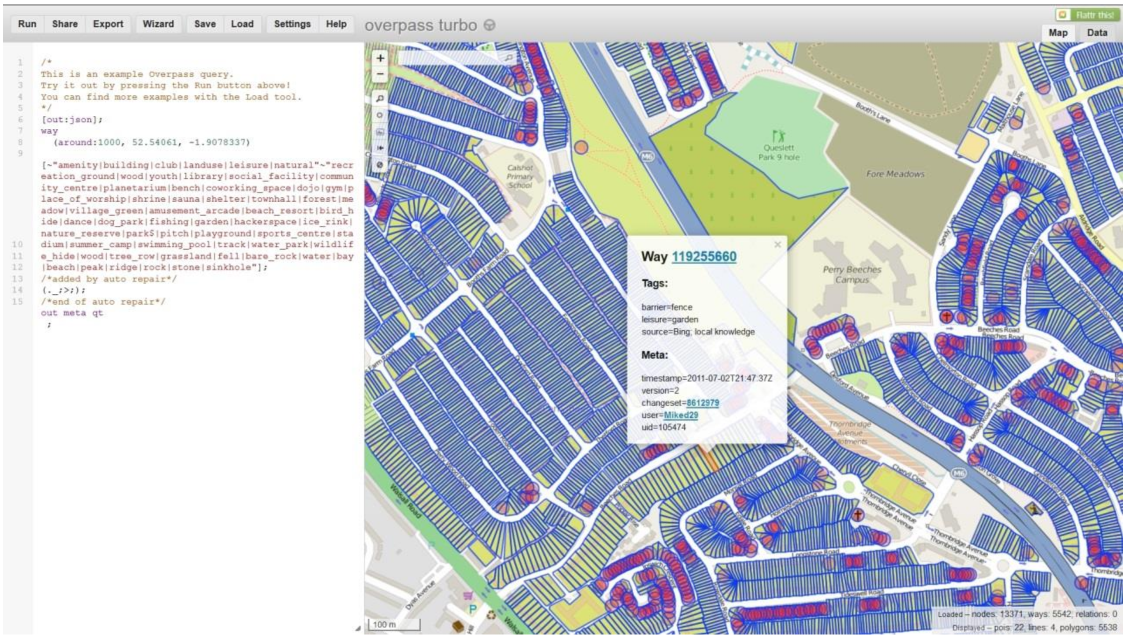
As the number of distinct tag keys thus columns in the data frame was 810 it became impossible to merge all the retrieved objects into a single data frame with 8 GB of available operating memory. This was certainly very inefficient way to store the OSM object, the facilities, and alternative approaches were considered so the data could be manipulated in R.

6.4 Initial Exploration of Facility Data

To enable all the facilities to fit into a single data frame 259 of the tag keys retrieved were selected and a new data frame with all the facilities created. The data frame was used to generate aggregates for overall facility count per school as well as per facility type aggregates.

6.4.1 Issues in Data

When exploring distribution of total facility count values it became obvious there were issues present in the data. 93 schools were found to have more than 500 facilities assigned to them while the highest count of facilities for one school was 5,896. The Overpass Turbo Web page was used to explore and understand the merit of the problem.



```
/*
This is an example Overpass query.
Try it out by pressing the Run button above!
You can find more examples with the Load tool.
*/
[out:json];
way
  (around:1000, 52.54061, -1.9078337)

[~"amenity|building|club|landuse|leisure|natural"~"recr
eation_ground|wood|youth|library|social_facility|commun
ity_centre|planetarium|bench|coworking_space|dojo|gym|p
lace_of_worship|shrine|sauna|shelter|townhall|forest|me
adow|village_green|amusement_arcade|beach_resort|bird_h
ide|dance|dog_park|fishing|garden|hackerspace|ice_rink|
nature_reserve|park$|pitch|playground|sports_centre|sta
dium|summer_camp|swimming_pool|track|water_park|wildlif
e_hide|wood|tree_row|grassland|fell|bare_rock|water|bay
|beach|peak|ridge|rock|stone|sinkhole"];
/*added by auto repair*/
(._;>;);
/*end of auto repair*/
out meta qt
 ;
```

Way 119255660

Tags:

barrier=fence
leisure=garden
source=Bing; local knowledge

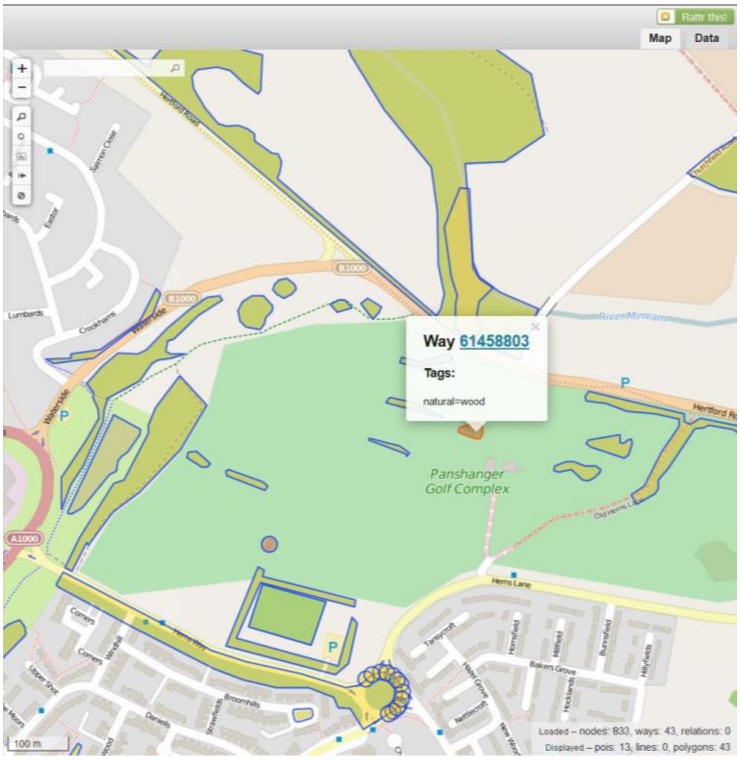
Meta:

timestamp=2011-07-02T21:47:37Z
version=2
changeset=8612979
user=Miked29
uid=105474

Loaded – nodes: 13371, ways: 5542, relations: 0
Displayed – pois: 22, lines: 4, polygons: 5538

Figure 6.2: House back garden tagged as leisure : garden

In the case illustrated in Figure 6.2 it was found that particular contributor to OSM tagged most of back gardens belonging to private houses as leisure : garden. Other issues discovered included private houses tagged as building-roof-shape : pitched which satisfied partial matching when using regular expressions in the query.



Way 61458803

Tags:

natural=wood

Loaded – nodes: 833, ways: 43, relations: 0
Displayed – pois: 13, lines: 0, polygons: 43

Figure 6.3: Woods inside a restricted area

A specific issue was identified where small patches of woods tagged as natural : wood were located inside an area with restricted access. Example from Welwyn Garden City in Figure 6.3 shows woods inside a golf complex. In order to eliminate such objects it would be required establishing containment and then check whether the containing object has or imposes restricted access.

The Overpass API provides 'recurse' functionality to find nodes that a way (polygon) comprises of or the opposite to find 'relation' that a node or way is contained in. This would work if the golf complex was of type 'relation' but because it was of type 'way' it could not be established whether or not it contained another way.

The SPARQL end point on LinkedGeoData which is run by Virtuso technology provides functionality for spatial predicates like 'st_contains'. Following SPARQL query was constructed to see if the desired containment could be established:

```
Prefix ogc: <http://www.opengis.net/ont/geosparql#>
Prefix geom:<http://linkedgeodata.org/geometry/>
Prefix rdfs:<http://www.w3.org/2000/01/rdf-schema#>

Select * {
    ?s

    geom:geometry [
      ogc:asWKT ?geom
    ] .


    Filter(bif:st_contains (
      ?geom,
      (Select * WHERE {
        geom:way61458803 ogc:asWKT ?geom2 .
      })
    )) .
}
```

The query searching the underlying graph of RDF triples should find 'subjects' with geometry in well-known text (asWKT) format such that it contains geometry of object 'way61458803' (the wood as can be seen in Figure 6.3). Unfortunately the query could not be fully tested as the patches of wood did not have their corresponding objects converted to the linked data form and present on LinkedGeoData.

Although the containment could not be established there was a problem with size of the wood objects anyway. Patches of wood like this or tree rows provide limited space for pupils' activities and should have been ignored in the first place based on their size. This would require not only calculating area of the polygon but also verify ratios between its edges.

These and other issues found while cleaning the data are detailed in the section 6.6.

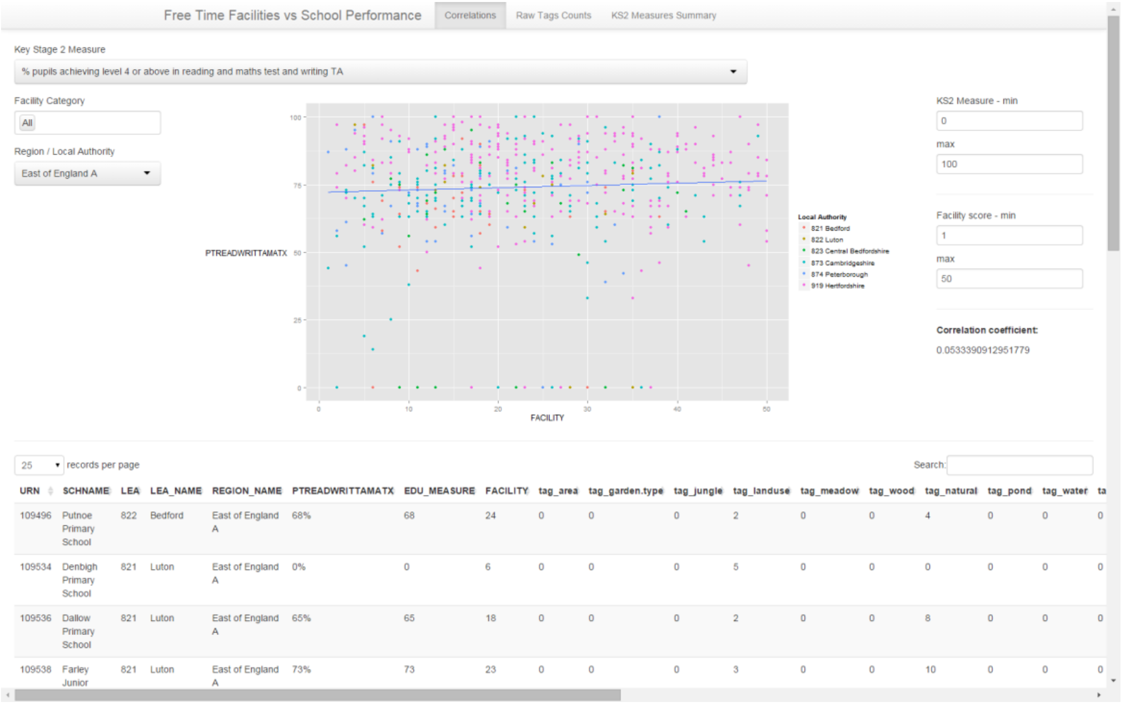
6.4.2 Web Interface for Exploration

Certain amount of time was invested in building an interactive Web interface to support the data exploration for purposes of the experiment using R library Shiny.

The interface allowed for applying various selection options in terms of the educational measure, facility type and geographical area. As problems with unreasonably high facility counts in the data were recognised it also allowed for setting the min and max counts to be included in correlation calculations.

Relevant facility types were split into 6 groups for purposes of exploration at this stage - land, nature, community, heritage, sport, animal. The groups were derived from all tag key : value pairs in the retrieved OSM objects. The Web interface allowed for selecting individual groups, multiple combinations of them as well as to select all of them.

Interactive table with its searching and per column ordering capabilities helped to explore distributions of facility type counts as well as the educational measures.



Key Stage 2 Measure: % pupils achieving level 4 or above in reading and maths test and writing TA

Facility Category: All

Region / Local Authority: East of England A

KS2 Measure - min: 0

max: 100

Facility score - min: 1

max: 50

Correlation coefficient:

0.0533390912951779

25 records per page

Search:

URN	SCHNAME	LEA	LEA_NAME	REGION_NAME	PTREADWRITTAMATX	EDU_MEASURE	FACILITY	tag_area	tag_garden.type	tag_jungle	tag_landuse	tag_meadow	tag_wood	tag_natural	tag_pond	tag_water	ta
109496	Putnoe Primary School	822	Bedford	East of England A	68%	68	24	0	0	0	2	0	0	4	0	0	0
109534	Denbigh Primary School	821	Luton	East of England A	0%	0	6	0	0	0	5	0	0	0	0	0	0
109536	Dallow Primary School	821	Luton	East of England A	65%	65	18	0	0	0	2	0	0	8	0	0	0
109538	Farley Junior	821	Luton	East of England A	73%	73	23	0	0	0	3	0	0	10	0	0	0

Figure 6.4: R-Shiny Web Interface - Correlations

6.4.3 Outcomes

Correlations coefficients for random combinations of educational measure, facility types and geographical groupings showed weak positive correlation relationships and in some cases even negative relationships.

Colour coding of individual data points in the scatter plot that was applied when regions or all England option were selected did not show any significant clusters.

On the other hand there was a shortcoming identified. Except where overall facility count was selected using facility groups resulted in counting some facilities multiple times per school.

A decision was made to attempt data cleaning focusing on the issues around schools with facility count over 500 and then to aggregate facility counts based on facility categories that were part of the query when retrieving data from OSM.

6.5 Deployment of MongoDB

To be able to clean the data it was necessary to merge the full set of OSM objects retrieved including all their properties and tags. Due to the structure of the OSM data where objects have different sets of properties and tags 'key-value store' and 'document-based database' technologies where considered rather than relational databases. MongoDB[23] was selected as there is R library to connect to it supported by the vendor and the extent of online documentation available provided good resource of information to deal with learning curve of using this technology.

It was relatively easy to load all 7 CSV files storing the retrieved OSM objects into the database. In MongoDB terminology a collection is equivalent of a database table and a document represents a record in a table.

The technology proved to be performant with reasonably low operating memory requirements and allowed for easy data manipulation from R. The R data structure that mostly fits data structure of MongoDB is list where a list item is a key : value pair. R provides functionality to convert between lists and data frames which made the data processing reasonably simple.

[23] http://www.mongodb.org

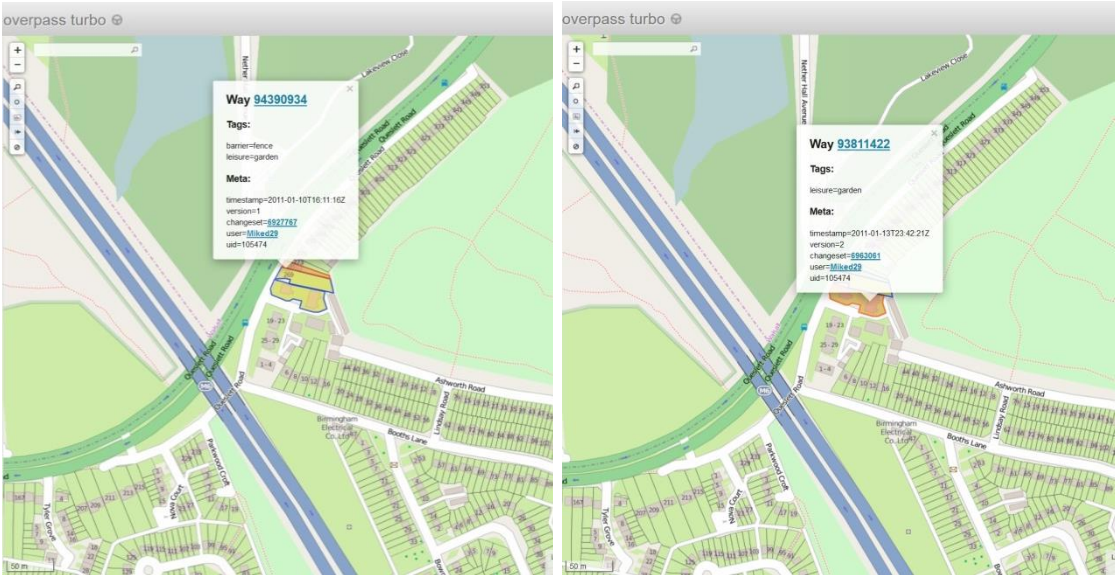
This type of out of memory processing with R was advantageous when cleaning data and calculating various aggregated values in consequent steps of the experiment.

6.6 Data Cleaning

After loading the retrieved OSM objects into MongoDB it was possible to start cleaning the issues identified in section 6.4.1 above. The focus was put on the 93 schools with total count of facilities over 500. Certain other issues were also discovered and addressed in the process.

6.6.1 Private Gardens Issue

The case of back gardens belonging to private houses and being tagged as leisure : garden posed a challenge in distinguishing public gardens or green areas tagged as leisure from the private ones. One approach could be eliminating the private gardens based on the OSM user name or ID but depending on the users' tagging activities publicly accessible gardens could get removed as well.



Way 94390934

Tags:

barrier=fence
leisure=garden

Meta:

timestamp=2011-01-10T16:11:16Z
version=1
changeset=6927767
user=Miked29
uid=105474

Way 93811422

Tags:

leisure=garden

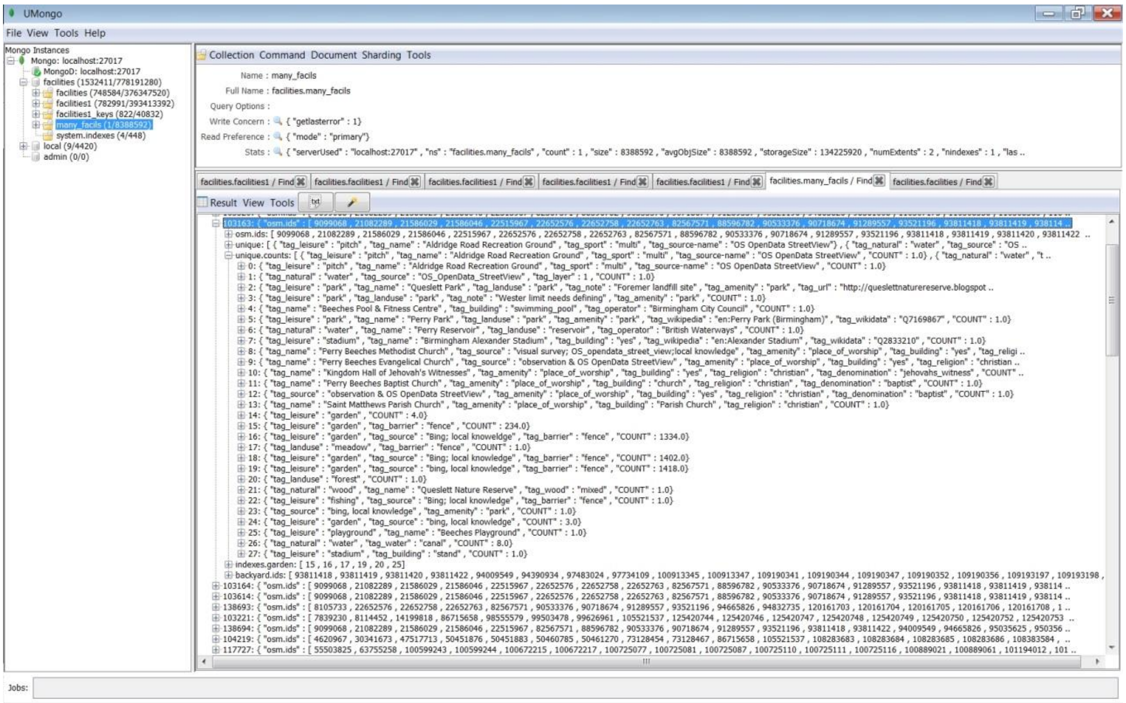
Meta:

timestamp=2011-01-13T23:42:21Z
version=2
changeset=6963061
user=Miked29
uid=105474

Figure 6.5: Private and public gardens tagged by the same contributor

The figure above illustrates this problem where among many private back gardens in an area there is a publically accessible green area around block of apartments that cannot be distinguished and preserved (right side of the figure). The tag combinations are not distinct enough and the OSM user name / ID are the same. This particular case was investigated using freely available aerial imagery to establish accessibility of the areas.

In an attempt to correctly identify private back gardens in order to eliminate them with reasonable level of precision unique tag combinations of facilities were counted on per school basis. A collection for this purpose was created in MongoDB.



UMongo
File View Tools Help
Collection Command Document Sharding Tools
facilities.facilities1 / Find
facilities.facilities1 / Find
facilities.facilities1 / Find
facilities.facilities1 / Find
facilities.facilities1 / Find
facilities.many_facils / Find
facilities.facilities / Find
Result View Tools
txt
Jobs:

```
Mongo Instances
Mongo: localhost:27017
  MongoD: localhost:27017
  facilities (1532411/778191280)
    facilities (748584/376347520)
    facilities1 (782991/393413392)
    facilities1_keys (822/40832)
    many_facils (1/8388592)
    system.indexes (4/448)
  local (9/4420)
  admin (0/0)
```

```
             Name : many_facils
        Full Name : facilities.many_facils
    Query Options :
    Write Concern : { "getlasterror" : 1}
  Read Preference : { "mode" : "primary"}
            Stats : { "serverUsed" : "localhost:27017" , "ns" : "facilities.many_facils" , "count" : 1 , "size" : 8388592 , "avgObjSize" : 8388592 , "storageSize" : 134225920 , "numExtents" : 2 , "nindexes" : 1 , "las ..
```

```
103163: { "osm.ids" : [ 9099068 , 21082289 , 21586029 , 21586046 , 22515967 , 22652576 , 22652758 , 22652763 , 82567571 , 88596782 , 90533376 , 90718674 , 91289557 , 93521196 , 93811418 , 93811419 , 938114 ..
  osm.ids: [ 9099068 , 21082289 , 21586029 , 21586046 , 22515967 , 22652576 , 22652758 , 22652763 , 82567571 , 88596782 , 90533376 , 90718674 , 91289557 , 93521196 , 93811418 , 93811419 , 93811420 , 93811422 ..
  unique: [ { "tag_leisure" : "pitch" , "tag_name" : "Aldridge Road Recreation Ground" , "tag_sport" : "multi" , "tag_source-name" : "OS OpenData StreetView"} , { "tag_natural" : "water" , "tag_source" : "OS ..
  unique.counts: [ { "tag_leisure" : "pitch" , "tag_name" : "Aldridge Road Recreation Ground" , "tag_sport" : "multi" , "tag_source-name" : "OS OpenData StreetView" , "COUNT" : 1.0} , { "tag_natural" : "water" , "t ..
    0: { "tag_leisure" : "pitch" , "tag_name" : "Aldridge Road Recreation Ground" , "tag_sport" : "multi" , "tag_source-name" : "OS OpenData StreetView" , "COUNT" : 1.0}
    1: { "tag_natural" : "water" , "tag_source" : "OS_OpenData_StreetView" , "tag_layer" : 1 , "COUNT" : 1.0}
    2: { "tag_leisure" : "park" , "tag_name" : "Queslett Park" , "tag_landuse" : "park" , "tag_note" : "Foremer landfill site" , "tag_amenity" : "park" , "tag_url" : "http://queslettnaturereserve.blogspot ..
    3: { "tag_leisure" : "park" , "tag_landuse" : "park" , "tag_note" : "Wester limit needs defining" , "tag_amenity" : "park" , "COUNT" : 1.0}
    4: { "tag_name" : "Beeches Pool & Fitness Centre" , "tag_building" : "swimming_pool" , "tag_operator" : "Birmingham City Council" , "COUNT" : 1.0}
    5: { "tag_leisure" : "park" , "tag_name" : "Perry Park" , "tag_landuse" : "park" , "tag_amenity" : "park" , "tag_wikipedia" : "en:Perry Park (Birmingham)" , "tag_wikidata" : "Q7169867" , "COUNT" : 1.0}
    6: { "tag_natural" : "water" , "tag_name" : "Perry Reservoir" , "tag_landuse" : "reservoir" , "tag_operator" : "British Waterways" , "COUNT" : 1.0}
    7: { "tag_leisure" : "stadium" , "tag_name" : "Birmingham Alexander Stadium" , "tag_building" : "yes" , "tag_wikipedia" : "en:Alexander Stadium" , "tag_wikidata" : "Q2833210" , "COUNT" : 1.0}
    8: { "tag_name" : "Perry Beeches Methodist Church" , "tag_source" : "visual survey; OS_opendata_street_view;local knowledge" , "tag_amenity" : "place_of_worship" , "tag_building" : "yes" , "tag_religi ..
    9: { "tag_name" : "Perry Beeches Evangelical Church" , "tag_source" : "observation & OS OpenData StreetView" , "tag_amenity" : "place_of_worship" , "tag_building" : "yes" , "tag_religion" : "christian ..
    10: { "tag_name" : "Kingdom Hall of Jehovah's Witnesses" , "tag_amenity" : "place_of_worship" , "tag_building" : "yes" , "tag_religion" : "christian" , "tag_denomination" : "jehovahs_witness" , "COUNT" ..
    11: { "tag_name" : "Perry Beeches Baptist Church" , "tag_amenity" : "place_of_worship" , "tag_building" : "church" , "tag_religion" : "christian" , "tag_denomination" : "baptist" , "COUNT" : 1.0}
    12: { "tag_source" : "observation & OS OpenData StreetView" , "tag_amenity" : "place_of_worship" , "tag_building" : "yes" , "tag_religion" : "christian" , "tag_denomination" : "baptist" , "COUNT" : 1.0}
    13: { "tag_name" : "Saint Matthews Parish Church" , "tag_amenity" : "place_of_worship" , "tag_building" : "Parish Church" , "tag_religion" : "christian" , "COUNT" : 1.0}
    14: { "tag_leisure" : "garden" , "COUNT" : 4.0}
    15: { "tag_leisure" : "garden" , "tag_barrier" : "fence" , "COUNT" : 234.0}
    16: { "tag_leisure" : "garden" , "tag_source" : "Bing; local knoweldge" , "tag_barrier" : "fence" , "COUNT" : 1334.0}
    17: { "tag_landuse" : "meadow" , "tag_barrier" : "fence" , "COUNT" : 1.0}
    18: { "tag_leisure" : "garden" , "tag_source" : "Bing; local knowledge" , "tag_barrier" : "fence" , "COUNT" : 1402.0}
    19: { "tag_leisure" : "garden" , "tag_source" : "bing, local knowledge" , "tag_barrier" : "fence" , "COUNT" : 1418.0}
    20: { "tag_landuse" : "forest" , "COUNT" : 1.0}
    21: { "tag_natural" : "wood" , "tag_name" : "Queslett Nature Reserve" , "tag_wood" : "mixed" , "COUNT" : 1.0}
    22: { "tag_leisure" : "fishing" , "tag_source" : "Bing; local knowledge" , "tag_barrier" : "fence" , "COUNT" : 1.0}
    23: { "tag_source" : "bing, local knowledge" , "tag_amenity" : "park" , "COUNT" : 1.0}
    24: { "tag_leisure" : "garden" , "tag_source" : "bing, local knowledge" , "COUNT" : 3.0}
    25: { "tag_leisure" : "playground" , "tag_name" : "Beeches Playground" , "COUNT" : 1.0}
    26: { "tag_natural" : "water" , "tag_water" : "canal" , "COUNT" : 8.0}
    27: { "tag_leisure" : "stadium" , "tag_building" : "stand" , "COUNT" : 1.0}
  indexes.garden: [ 15 , 16 , 17 , 19 , 20 , 25]
  backyard.ids: [ 93811418 , 93811419 , 93811420 , 93811422 , 94009549 , 94390934 , 97483024 , 97734109 , 100913345 , 100913347 , 109190341 , 109190344 , 109190347 , 109190352 , 109190356 , 109193197 , 109193198 ,
103164: { "osm.ids" : [ 9099068 , 21082289 , 21586029 , 21586046 , 22515967 , 22652576 , 22652758 , 22652763 , 82567571 , 88596782 , 90533376 , 90718674 , 91289557 , 93521196 , 93811418 , 93811419 , 938114 ..
103614: { "osm.ids" : [ 9099068 , 21082289 , 21586029 , 21586046 , 22515967 , 22652576 , 22652758 , 22652763 , 82567571 , 88596782 , 90533376 , 90718674 , 91289557 , 93521196 , 93811418 , 93811419 , 938114 ..
138693: { "osm.ids" : [ 8105733 , 22652576 , 22652758 , 22652763 , 82567571 , 90533376 , 90718674 , 91289557 , 93521196 , 94665826 , 94832735 , 120161703 , 120161704 , 120161705 , 120161706 , 120161708 , 1 ..
103221: { "osm.ids" : [ 7839230 , 8114452 , 14199818 , 86715658 , 98555579 , 99503478 , 99626961 , 105521537 , 125420744 , 125420746 , 125420747 , 125420748 , 125420749 , 125420750 , 125420752 , 125420753  ..
138694: { "osm.ids" : [ 9099068 , 21082289 , 21586029 , 21586046 , 22515967 , 82567571 , 88596782 , 90533376 , 90718674 , 91289557 , 93521196 , 93811418 , 93811422 , 94009549 , 94665826 , 95035625 , 950356 ..
104219: { "osm.ids" : [ 4620967 , 30341673 , 47517713 , 50451876 , 50451883 , 50460785 , 50461270 , 73128454 , 73128467 , 86715658 , 105521537 , 108283683 , 108283684 , 108283685 , 108283686 , 108383584 ,  ..
117727: { "osm.ids" : [ 55503825 , 63755258 , 100599243 , 100599244 , 100672215 , 100672217 , 100725077 , 100725081 , 100725087 , 100725110 , 100725111 , 100725116 , 100889021 , 100889061 , 101194012 , 101 ..
```

Figure 6.6: Unique tag combinations counts

The figure above shows screen shot from UMongo, a graphical user interface for MongoDB, with an example where school with URN identifier 103163 has 28 unique tag combinations each of which has aggregated count assigned to it. The most prevailing garden related tag combinations are in elements with indexes 15, 16, 18 and 19 having counts 234, 1334, 1402 and 1418 respectively.

The private gardens problem for the 93 schools with facility counts over 500 was approached by removing OSM objects having garden related tag combinations with extremely high counts. Objects representing gardens with smaller counts were investigated individually in order to decide whether to retain or to remove them. Overall only 4 gardens among the 93 schools were found as to be retained which proved marginal gain in using this approach.

Correcting this issue resulted in 183,339 records being removed. Other issues discovered in the process are described in the next section.

6.6.2 Other issues

Some of the private gardens happened to also contain tag ‘access’ with values ‘private’ or ‘no’. Based on the OSM documentation the access tag can have multiple values among which private, no, restricted and prohibited were selected to eliminate any OSM objects with limited access from the data. It helped to remove 12,061 records.

There were 13,901 records discovered as being private houses tagged as building-roof-shape : pitched. These were retrieved coincidentally as both the key and the value satisfied partial matching when using regular expressions in the query. The quite specific tag key used in this case allowed targeting the records quite precisely and all were removed.

Objects having varying tagging relating them to some kind of car parking or car repair were found as a less significant issue. They were present in categories amenity, building, landuse as well as shop. Altogether 52 records were removed regarding the issue.

The last set of objects removed in this step represented irrelevant facility types especially in category amenity, e.g. bank, bar, beer_garden, bus_shelter, bus_station, coach park, restaurant etc. Targeting these facility types resulted in 174 records being removed.

6.6.3 Outcome

There were over 209,000 invalid records eliminated at this stage but the task proved to be very time consuming and different techniques would need to be investigated to deal with issues in the whole dataset retrieved from OSM in an appropriate and efficient manner.

The focus was moved to categories used while initially retrieving data from OSM in an attempt to reduce number of facility types for inclusion in the correlation calculations.

6.7 Aggregating Facility Counts based on OSM Tag Categories Used during Initial Retrieval

The cleaning in the previous step showed that the initial approach to look for correlations among all the retrieved tag categories from OSM was inefficient due to the extent of inconsistencies in the data. Additionally the sparseness of values for majority of these tag categories on one hand and the huge concentration of irrelevant values of certain tag categories in particular areas on the other, e.g. leisure : garden pair referring to a back garden of a private residence, yielded significantly skewed distributions when aggregating the facility counts.

The decision was made to focus on the tag categories included in the query against the Overpass interface when retrieving OSM entities for each school. Such approach was expected to narrow down the scope of facility types included in the correlation calculations. For each of the 6 categories, e.g. amenity, leisure etc., the distributions of their values were examined to find facility types with significant presence among the schools. The number of distributions to examine was reduced from 810 to 399 and further to 392 when category ‘club’ with its 7 distinct values was found to be present only for 50 schools. As this accounted only for 13 distinct OSM entities each of them also belonging to either ‘leisure’ or ‘amenity’ category the ‘club’ category was skipped in the consequent analysis.

Following two tables show the first 15 values in each of the remaining 5 facility categories along with the number of schools they were retrieved for and the distinct count regarding the OSM entities based on their unique ID respectively. Full tables can be found in ‘tag categories.xls’ file.

These groupings allowed for highlighting the most appearing facility types to focus the correlation search on. They also helped to discover types that would be potentially assigned to individual schools in unreasonable high counts like it was in case of leisure : garden type and should be subject to additional cleaning. Among these was type natural : wood which after investigating its distribution across the schools proved to be excessive in many cases. There were 1,235 schools with over 20 woods out of which 9 schools with over 150 woods in 1 km radius.

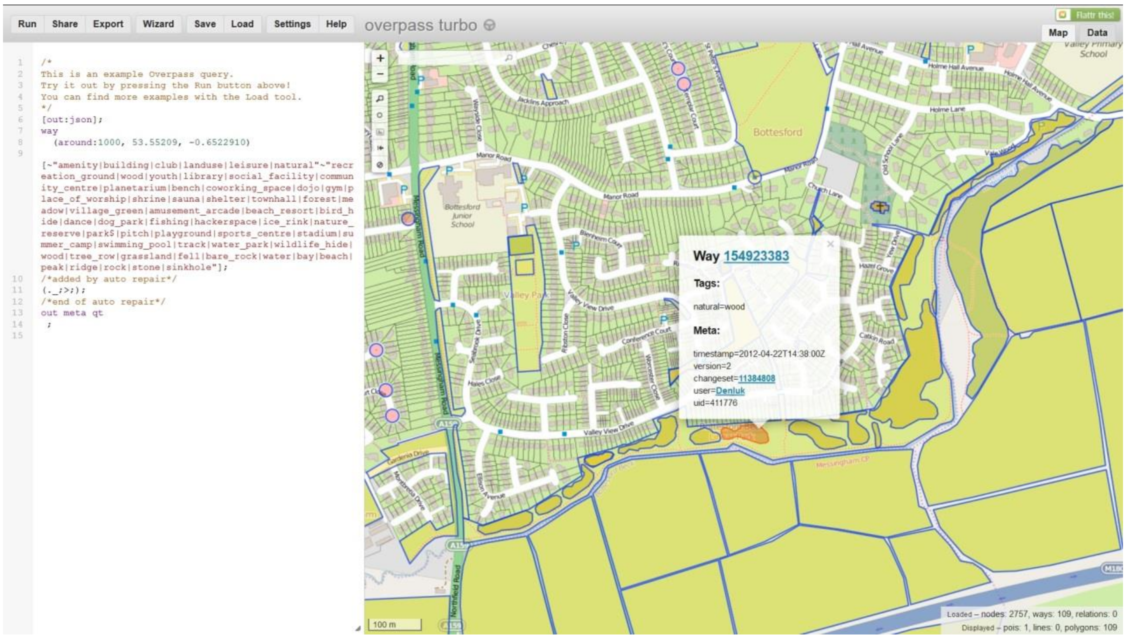
tag_amenity	school count	tag_building	school count	tag_landuse	school count	tag_leisure	school count	tag_natural	school count
place_of_worship	43362	yes	40888	meadow	28981	pitch	98558	wood	137694
library	5045	church	9045	forest	22489	park	66002	water	43356
community_centre	3964	beach_hut	601	recreation_ground	21219	playground	30788	grassland	8632
social_facility	1460	public	421	village_green	3167	garden	24533	tree_row	1841
shelter	1399	house	298	grass	3139	recreation_ground	9453	beach	1269
townhall	1267	library	269	park	870	sports_centre	8360	grass	456
swimming_pool	1223	shelter	247	garden	298	nature_reserve	2692	bare_rock	162
park	563	civic	220	basin	190	track	2005	scrub	150
bench	373	place_of_worship	220	reservoir	146	stadium	1708	heath	116
gym	254	chapel	208	conservation	123	swimming_pool	1273	wetland	48
youth_centre	165	mosque	164	recreational	115	common	207	meadow	44
fountain	145	residential	161	pond	65	fishing	192	no	39
grave_yard	79	office	158	pitch	54	marina	156	forest	31
youth_club	76	bridge	147	leisure	48	dog_park	154	woodland	29
school	71	cathedral	123	residential	43	water_park	95	rock	28

Table 6.1: Facility types in 5 categories with school counts

tag_amenity	distinct OSM ID	tag_building	distinct OSM ID	tag_landuse	distinct OSM ID	tag_leisure	distinct OSM ID	tag_natural	distinct OSM ID
place_of_worship	9861	yes	8545	meadow	12656	pitch	26540	wood	62186
library	1063	church	2521	forest	9738	park	13146	water	20144
community_centre	980	beach_hut	421	recreation_ground	5961	playground	7741	grassland	2540
swimming_pool	442	public	90	village_green	1189	garden	5052	tree_row	846
townhall	345	shelter	81	grass	714	recreation_ground	2941	beach	479
shelter	327	house	80	park	132	sports_centre	1979	grass	115
social_facility	318	civic	68	basin	66	nature_reserve	723	bare_rock	86
bench	111	office	67	garden	51	track	501	scrub	44
park	86	library	59	reservoir	48	swimming_pool	410	heath	26
gym	54	chapel	54	conservation	33	stadium	376	forest	17
youth_centre	36	residential	53	pond	20	fishing	77	rocks	13
fountain	29	place_of_worship	49	residential	20	common	59	wetland	12
school	24	bridge	29	pitch	11	marina	52	no	11
youth_club	19	community_centre	29	recreational	10	water_park	26	rock	11
grave_yard	14	retail	27	wood	10	dog_park	24	meadow	10

Table 6.2: Facility types in 5 categories with distinct OSM ID counts

To illustrate the case of the natural : wood pair following are map pictures from Scunthorpe city. The map is overlaid with multiple wood entities which happen to be inside a bigger entity, namely area tagged as leisure : recreation_ground. This goes back to the problem of not being able to verify such containment using queries against the available OSM data interfaces (see section 6.4.1).



```
/*
This is an example Overpass query.
Try it out by pressing the Run button above!
You can find more examples with the Load tool.
*/
[out:json];
way
  (around:1000, 53.55209, -0.6522910)

[~"amenity|building|club|landuse|leisure|natural"~"recreation_ground|wood|youth|library|social_facility|community_centre|planetarium|bench|coworking_space|dojo|gym|place_of_worship|shrine|sauna|shelter|townhall|forest|meadow|village_green|amusement_arcade|beach_resort|bird_hide|dance|dog_park|fishing|hackerspace|ice_rink|nature_reserve|park$|pitch|playground|sports_centre|stadium|summer_camp|swimming_pool|track|water_park|wildlife_hide|wood|tree_row|grassland|fell|bare_rock|water|bay|beach|peak|ridge|rock|stone|sinkhole"];
/*added by auto repair*/
(._;>;);
/*end of auto repair*/
out meta qt
 ;
```

Way 154923383

Tags:

natural=wood

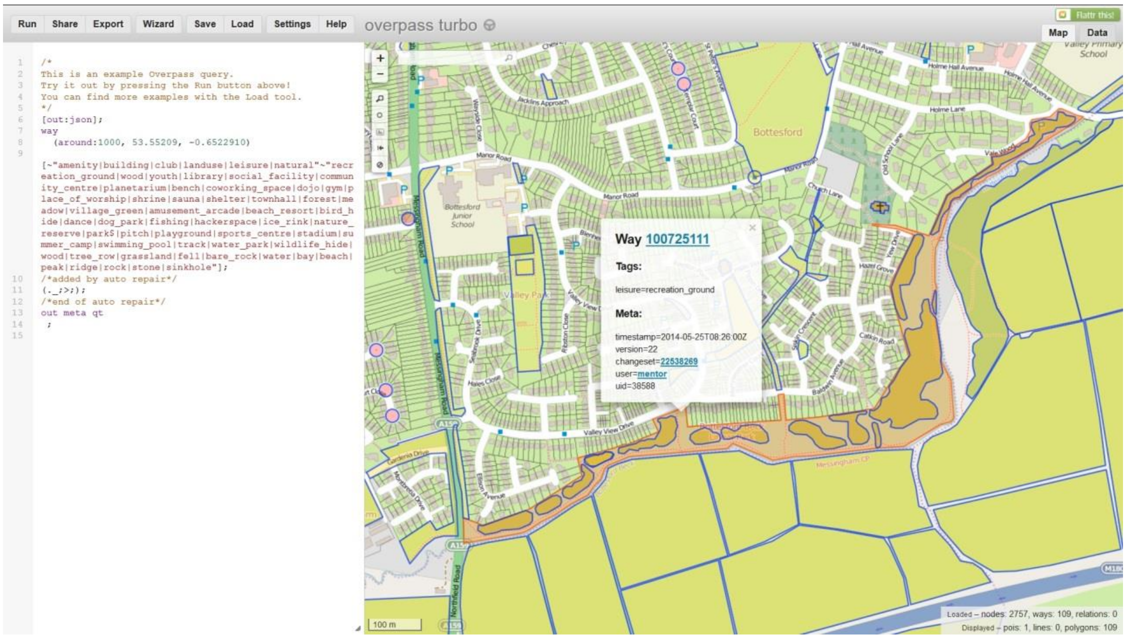
Meta:

timestamp=2012-04-22T14:38:00Z
version=2
changeset=11384808
user=Denluk
uid=411776

Loaded – nodes: 2757, ways: 109, relations: 0
Displayed – pois: 1, lines: 0, polygons: 109

Figure 6.7: Multiple wood entities

In this case the recreation ground is a long strip of grass area with many patches of trees each of which is tagged as wood. The correct approach here would be to only count the recreation ground and ignore all of the woods.



```
/*
This is an example Overpass query.
Try it out by pressing the Run button above!
You can find more examples with the Load tool.
*/
[out:json];
way
  (around:1000, 53.55209, -0.6522910)

[~"amenity|building|club|landuse|leisure|natural"~"recreation_ground|wood|youth|library|social_facility|community_centre|planetarium|bench|coworking_space|dojo|gym|place_of_worship|shrine|sauna|shelter|townhall|forest|meadow|village_green|amusement_arcade|beach_resort|bird_hide|dance|dog_park|fishing|hackerspace|ice_rink|nature_reserve|park$|pitch|playground|sports_centre|stadium|summer_camp|swimming_pool|track|water_park|wildlife_hide|wood|tree_row|grassland|fell|bare_rock|water|bay|beach|peak|ridge|rock|stone|sinkhole"];
/*added by auto repair*/
(._;>;);
/*end of auto repair*/
out meta qt
 ;
```

Way 100725111

Tags:

leisure=recreation_ground

Meta:

timestamp=2014-05-25T08:26:00Z
version=22
changeset=22538269
user=mentor
uid=38588

Loaded – nodes: 2757, ways: 109, relations: 0
Displayed – pois: 1, lines: 0, polygons: 109

Figure 6.8: Multiple wood entities on a recreation ground

The groupings showed some inconsistencies in naming as well as irrelevant facility types with significant numbers, e.g. building : yes which being simply tagged that an entity is a building but had no value for purposes of this project.

Using grouping created in this manner imposed a limitation on how the facility types could be combined. The OSM entities retrieved have one value per category assigned but can be tagged as belonging to multiple categories. Due to this fact it would be reasonable to combine aggregated per school counts only for facility types within the same category when looking for correlations. Including combinations across multiple categories would result in higher counts of facilities than they were actually present. Over 60,000 thousand records were found having tags from the 5 categories combined.

6.8 Verifying Linked Data for Presence of External Links

Having established counts of facility types in 5 categories in the previous step certain OSM objects were selected to be verified for having external links to other data sources on the Semantic Web. Six facility types were selected based on their overall count as well as relevancy to provide free time activities for children:

- amenity : community_centre,
- amenity : youth_centre,
- amenity : social_facility
- leisure : pitch,
- leisure : recreation_ground,
- leisure : water_park.

Firstly OSM IDs of all objects belonging to these facility types were retrieved from MongoDB which yielded total number of 31,416 distinct OSM objects. Using the IDs all ‘predicate’-‘object’ pairs belonging to each object were retrieved from LinkedGeoData SPARQL end point using following query:

```
Prefix lgd:<http://linkedgeodata.org/triplify/>
SELECT *
  lgd:way<OSM_ID> ?p ?o .
}
```

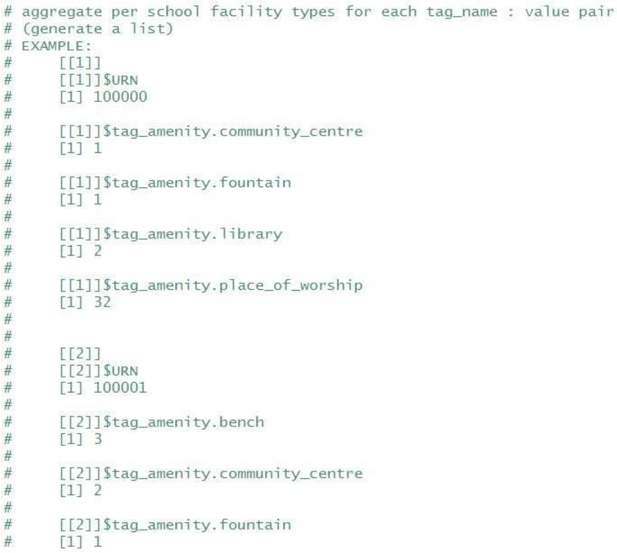
Once downloaded they were stored in R list with each item having the OSM ID as a key. Out of 31,416 OSM objects only 29,380 were found having their linked data equivalent on LinkedGeoData.

All ‘predicates’ retrieved were then checked for containing string ‘sameAs’ which would indicate presence of an external link to other source on the Semantic Web. None of the predicates was found containing the string so it was not possible to continue in assessing the linked data as planned in Methodologies section in the previous chapter.

6.9 Aggregating Facility Type Counts Per School

Some of the facility types identified in the section 6.7 were addressed after building a list structure in R with aggregated facility type counts for each school based on the groupings.

The list structure contained items each uniquely identified by URN : value pair. Facility counts were encoded in form of tag_category.value : count pairs. See an example below. The 'flat' item structure with URN key being at the same level as the facility counts rather than the school number being a key for the whole item (`{URN : [{tag_category.value : count }]}`) was chosen for optimal insertion into the MongoDB database one school per document in the collection.



```
# aggregate per school facility types for each tag_name : value pair
# (generate a list)
# EXAMPLE:
#     [[1]]
#     [[1]]$URN
#     [1] 100000
#
#     [[1]]$tag_amenity.community_centre
#     [1] 1
#
#     [[1]]$tag_amenity.fountain
#     [1] 1
#
#     [[1]]$tag_amenity.library
#     [1] 2
#
#     [[1]]$tag_amenity.place_of_worship
#     [1] 32
#
#
#     [[2]]
#     [[2]]$URN
#     [1] 100001
#
#     [[2]]$tag_amenity.bench
#     [1] 3
#
#     [[2]]$tag_amenity.community_centre
#     [1] 2
#
#     [[2]]$tag_amenity.fountain
#     [1] 1
```

Figure 6.9: Facility Type Counts Per School – R list structure

The list was then converted into the R data frame structure which resulted in a sparse table with 392 columns of facility type counts. Any non-present facility types for each school were assigned count 0. Although not very efficient the operations on data frame are more suitable for analytical tasks than those of list structures.

Having such summary highlighted facility types with unreasonably high counts in each category. Running overview queries against the facility counts data frame uncovered the problematic natural : wood type described in the previous section as well as the not fully cleaned leisure : garden type. There were still 379 schools with over 10 gardens out of which 8 schools had over 300 gardens in 1 km radius.

6.10 Correlations

Due to the sheer amount of data and thus possible combinations of school performance measures against the facility types with opportunity for different geographical as well as facility type groupings, a ‘brute-force’ approach was taken to calculate correlation coefficients to cover large enough amount of these combinations. The overall count of facilities was also included in the calculation although this count was derived separately to avoid the limitation described in section 6.7 above.

R function ‘cor.test’ was used for correlation calculations as it performs also hypothesis test and provides statistical information like p-value for the null hypothesis that correlation is 0, adjustable confidence interval for the calculated coefficient in addition to the correlation coefficient itself. The result is an R object containing all the information described.

To systematically investigate the performance measure versus facility type combinations the problem was approached at multiple levels. The naming convention for distinguishing calculation for individual combinations followed the scheme:

<ks2_measure_code>@<facility_category.facility_type>@<area_name>

6.10.1 School Performance Measures

Certain school performance measures from the Key Stage 2 data set were selected. The focus was put on those measures covering multiple skills, e.g. reading, writing maths, and those that the Department for Education in the UK uses when comparing similar schools and measures taking progression within the stage into account. Such progression measures are regarded as having better indicative value than measures concerned with only a single point in time.

There were following 14 measures selected:

- Overall value added lower 95% confidence limit,
- Overall value added upper 95% confidence limit,
- Average point score,
- % pupils achieving level 4 or above in reading and maths test and writing TA,

- % pupils achieved level 4B or above in reading and maths test and level 4 in writing TA,
- % pupils achieving level 5 or above in reading and maths test and writing TA,
- % pupils achieving 3 or below in reading and maths test and writing TA,
- % pupils with low key stage 1 attainment achieving level 4 or above in reading and maths test and writing TA,
- % pupils with middle key stage 1 attainment achieving level 4 or above in reading and maths test and writing TA,
- % pupils with high key stage 1 attainment achieving level 4 or above in reading and maths test and writing TA,
- % pupils achieving level 4 or above in reading and maths test and writing TA: 2012,
- % pupils achieving level 5 or above in reading and maths test and writing TA: 2012,
- % pupils achieving level 4 or above in reading and maths test and writing TA: 2013,
- % pupils achieving level 5 or above in reading and maths test and writing TA: 2013.

For each of the measures only schools with valid values were included. Invalid values included no value, LOWCOV and SUPP. These represent a fact that a school did not provide the results or the results were supressed due to the low number of pupils and thus potential risk of identifying the individual pupils.

6.10.2Facility Types and their Groupings

Out of the 392 facility types only those being present in more than 50 schools were included in correlation calculations (see facility types in 5 categories described in section 6.7 above). Irrelevant types like building : yes etc. were excluded from the calculations.

Where facility type counts for individual schools where found over 10 they were adjusted to this number except following which were decided upon based on the findings from the cleaning step and the built up knowledge of the retrieved OSM data set:

Facility type	**Adjusted count**
natural : wood	20
leisure : garden	5
leisure : swimming_pool	2

Table 6.3: Adjusted counts for selected facilities

Additionally groups within some of the categories were defined with the aim to investigate correlation where the facility types could be regarded as related, e.g. community centre, youth centre, youth club in category amenity etc. In some cases a group contained other facility types from the category that would not be present in 50 or more schools as the counts within the group were summed.

6.10.3Facility Type Scoring

The scoring mechanism was applied when calculating correlation coefficient with facility groups. Using the score when dealing with only single facility type would not have any effect on the correlation.

First the main facility data frame was subset to contain only facility types assigned to the particular group. Each facility type count per school was multiplied by predetermined score in range from 0 to 1 and only then the total sum for the school was calculated. The scoring levels were decided upon taking the results of the survey conducted into account. The responses received in the survey for the presented facility type categories can be seen below. The average score of the scale 1 to 5 (1 – not important, 5 – very important) was projected into scale 0, 25, 50, 75, 100%.

Facility group	Average score	Projected percentage
Community (community centre, youth club, leisure centre etc.)	4.91	100%
Sport (football, cricket, climbing etc.)	4.36	75%
Heritage (memorial, history museum, ruins etc.)	3.36	50%
Land (woods, grassland, public garden etc.)	3.8	75%
Nature (lake, beach etc.)	3.45	75%
Animal (animal shelter, bird hide, dog fowling etc.)	3.5	75%

Table 6.4: Facility scoring based on the survey

The facility type grouping in the experiment did not fully match the groups presented in the survey therefore individual facility types were scored differently within each facility group used for calculating the correlation coefficient.

The planned use of inverse distance weight to discriminate facility based on its distance from the school was not introduced as it would require querying all the distinct OSM entities retrieved previously (over 240 thousand) for their geographical coordinates before the actual distance could be calculated.

6.10.4Geographical Areas

As each school is assigned to one of the 152 Local Authorities which in turn belong to one of the 25 regions of England the schools were combined accordingly and correlation coefficient calculated for each of them including the whole England. This increased the multitude of correlation combinations by factor of 178.

6.11 Multiple Comparison Problem

While the calculations from the previous section were still running it was recognised that the approach of testing so many samples will indeed find strong correlations with p-value < 0.05 but it will lead to inflating the possibility of a type I error, a false positive. Please see discussion about the problem in the next chapter.

At this stage the time allocated for the experiment was over and the work on it had to be put on hold.

6.12 Conclusions

The chapter described implementation of the experiment by chronologically detailing individual steps undertaken. Problems and challenges encountered during the course of the experiment were included along with presenting the alternative approaches taken to deal with them. The results of the experiment were presented and their evaluation is discussed in the next chapter.

7. EVALUATION

7.1 Introduction

This chapter presents evaluation of results and findings from the experiment detailed in the previous chapter while reflecting knowledge elicited from the literature review. Conducted survey is also presented and evaluated in this chapter.

The steps and attempts undertaken within the experiment are evaluated in terms of their contribution towards the project objectives. Potential problems with soundness of results achieved are discussed where appropriate.

7.2 Linked Data Evaluation

As set in the Methodologies section of the Experiment Design chapter to be able to evaluate linked data contribution to the analysis of correlation relationships it was necessary to establish if the linked data version of OSM data that was subject of the experiment had enough external links to other data sets on the Semantic Web present.

Due to the limitation described in the previous chapter LinkedGeoData was not used for retrieving the OSM objects. The data was downloaded directly from OSM database via the Overpass interface. Only when the LinkedGeoData SPARQL end point became available again distinct OSM objects from selected facility types were queried and checked for presence of the 'sameAs' predicate.

It was found that out of the total number of 31,416 distinct OSM objects belonging to the non-overlapping six selected facility types only 29,380 had their RDF version present on the LinkedGeoData. When checking for the presence of the 'sameAs' predicate it was not found in any of the 29,380 objects.

Despite the number of links created to DBpedia and GeoNames hubs on the Semantic Web as reported by Stadler *et al.* (2012) OSM objects located in England and used in this experiment did not seem to be among those with external links. Although only approx. 5% of all the previously retrieved OSM objects were checked for presence of the 'sameAs' predicate the fact that absolutely none contained this predicate was deemed satisfactory to assume that there is insufficient number of external links for purposes of the experiment.

Having established insufficiency of external links presence the step 2 in the process to assess contribution of linked data to investigation of correlation relationships as defined in the Methodologies section in chapter 5 was omitted. The assessment of advantages having linked data included in the analysis is thus very limited.

The potential of publishing data following the linked data principles including its interlinking with the rest of the Semantic Web, as it is presented in current research of this area of interest, is certainly indisputable. Within the scope and settings of this project the potential could not have been exploited for purposes of analysis.

7.3 Evaluation of Correlation Results

The investigation of correlation relationships between samples of individual educational performance measures and individual as well as grouped facility types was undertaken with the generated dataset at various stages where attempts were made to clean the data and rearrange facility aggregates to find any promising combinations.

7.3.1 Initial Correlation Calculations

First random check for correlations took place after all OSM objects were downloaded. The data yielded 810 tag keys, i.e. categories like amenity and leisure but also many others that were mostly irrelevant and came as additional attributes of the retrieved objects. Only 259 of these categories were selected in an attempt to merge all the retrieved OSM objects from the multiple download batches into a single table. It was not possible to include more categories due to restriction of available operating memory. The table was then aggregated and facilities counted in each category per school. Also overall facility count per school was calculated while making sure each facility was counted only once.

At that stage simple Web interface was created using available R technology to aid the exploration by means of scattered plot with fitted line, correlation value calculated and interactive table to explore the underlying data.

Random combinations of the three factors (educational measure, facility type, area) showed weak positive correlations of up to 0.2 or even negative correlations in some cases. Three main problems were recognised in this setup:

- The aggregated data contained facility counts of over 100 or even 5,000 for some schools,
- When combining multiple facility types some of the facilities were counted more than once and
- By setting the min/max counts schools with counts outside of the selected range got ignored.

Due to these problems the data was processed as described in the previous chapter and an extensive search for correlations was attempted.

7.3.2 Extensive Search for Correlations

After the OSM data was partially cleaned and a new table created with aggregated facility counts for each facility type in the five selected categories an R script was built to iteratively apply correlation tests for over 900 thousand combinations.

The aim was to explore the range of correlation relationships present in the data and if any were found strong enough those cases could be further investigated. It was assumed that knowing which facility types strongly correlate with which educational performance measure in which area would provide a basis for establishing the distinguishing school properties that could be generalised to some degree.

There were strong positive correlations found and in 23 cases the correlation coefficient was greater than 0.8. Except five cases the rest of these had p-value of the t-test lower than the predetermined 0.05 which would indicate their significance.

Only at this stage of the experiment it was recognised that performing the correlation coefficient calculations for so many samples from the data set inevitably lead to inflating the possibility of a type 1 error, a false positive. Any discovered correlations, which warrant the rejection of the null hypothesis of zero correlation, could therefore not be deemed reliable due to the possibility of type 1 error. Further analysis would need to control for the inflation of error and/or to impose a different statistical approach, such as a t-test between top and bottom percentiles, so as to draw reliable and valid statistical relations from the dataset.

7.4 Evaluation of the Experiment

The experiment was designed to investigate correlations between educational performance and available free time facilities for children for relatively large geographical area. Although the scope was narrowed down for England due to dealing with geographical data this led to creation of data set with over 780 thousand records out of which over 240 thousand represented distinct OSM objects. As the OSM data used in the experiment is subject to inconsistencies it imposed significant challenges.

The automated retrieval of OSM data by issuing spatial queries against the Overpass interface took over 4 days. The transfer of data was interrupted multiple times resulting in 7 separate batches of data. When trying to merge all the data retrieved into single table the R limitation of storing all the data from a session in operating memory became an obstacle. This could have been avoided to some degree if a normalisation of some sort was applied, i.e. storing only distinct OSM objects and the using their unique IDs to reference them from each school for which they were downloaded.

The inclination to generate sparse table with one row per school each containing overall facility count and then counts for individual facility types came from the need to create an analytical base table for the analytical step.

After an attempt to process subset of the retrieved OSM data in R it was decided to deploy NoSQL document-based database technology, namely MongoDB. This proved to be advantageous for the purposes of the experiment and helped in the consequent data processing. The OSM data was downloaded in JSON format and NoSQL database technology should have been recognised during the design phase of the experiment as a native persistent storage for this data format and included from the beginning.

There were numerous objects found in the retrieved OSM data that skewed distributions of facility counts for schools due to a specific tagging applied to them and their high concentrations in some places, e.g. private house back garden tagged as leisure : garden etc. When trying to clean OSM data from such objects the data was aggregated per school based on all unique combinations of tag key : value pairs. This helped to find most prevalent issues in the data relatively quickly. The cleaning process though proved to be very tedious and time consuming. After eliminating over 200 thousand OSM object related records from the data set there were still many schools with over 100 facilities.

A decision was made to focus only on 5 out of 6 main facility categories that were used in the spatial queries when initially retrieving data from OSM. Across these categories there were 392 different values present. An aggregation was performed to find most occurring values in each category and so to narrow down subset of values to be included when looking for correlation relationship with selected educational performance measures. The limitation of this approach was that facility types could be combined only within single category to avoid counting the same facility more than once. This was step in the right direction and would have been taken earlier the scope of the experiment would become more manageable especially regarding the time restrictions of the whole project.

The last attempt to find any strong positive correlation relationship in the data that was performed in a brute-force fashion is evaluated in the previous section.

The experiment proved that inconsistencies pose a significant challenge when dealing with Volunteered Geographic Information as it can be found in the related research (Coleman *et al.* 2009; Mooney *et al.* 2012; Karam and Melchiori 2013). The inconsistencies are hardly avoidable due to nature of VGI and how it is generated. The more appropriate approach would have been to firstly identify areas of England where the OSM data is relatively consistent in terms of the tag relevancy for free time facility categorisation and then perform the experiment only in that scope.

Similar holds for cases where no facilities where retrieved. Although it is possible that in rural areas no relevant facilities can be found within 1 km radius from a school it can also be the case of missing OSM data for particular area due to low interest to map such area among the volunteer contributors (Ciepłuch *et al.* 2010; Haklay 2010).

7.5 Survey Results

The survey was conducted to support adjustment of scores for the facility types in terms of their relevancy for provision of free time activities for children. It also contained open-ended questions aimed to get a better understanding of problems faced by local communities regarding children in general as well as in the free time opportunities context.

There were altogether 386 email requests for participation in the survey sent to community centres and primary schools across England and Ireland. The email addresses had to be sourced manually from the internet. The biggest challenge was to find community centres with email contacts as many of them seemed to use only postal address and phone contact.

Prior to publishing the survey there were 6 categories recognised which covered any relevant OSM tag values in the downloaded data set:

- Community (community centre, youth club, leisure centre etc.),
- Sport (football, cricket, climbing etc.),
- Heritage (memorial, history museum, ruins etc.),
- Land (woods, grassland, public garden etc.),
- Nature (lake, beach etc.),
- Animal (animal shelter, bird hide, dog fowling etc.).

The participants rated importance of each of them on scale 1 to 5. The conversion of the received ratings into the scores used as part of the experiment is described in section 6.10.3 above.

When directly asked about their opinion on impact of availability of relevant facilities on pupils' school performance all participants except one selected the 'positive impact' option.

Two questions were concerned with the possibility to build a model for recommending facilities to be provided by the local council. Most participants agreed on usefulness of such model though only 3 of them thought that an adoption of such model by local authorities was realistic.

The open-ended questions could be split into two groups, services provided for pupils and issues faced with regard to pupils in the target age group.

Among the services provided currently the sport and after-school dominated the responses. As per services that should be provided but are not at present the ‘more sport’ appeared in multiple responses. Participants in two cases also expressed problem with financial barriers for children to access already available services.

The issues regarding pupils in the target age group described in the responses include lack of funding for space and activities, poverty, inequality and problems with transport. The issues seem to be addressed by mentoring programmes and support groups as well as by securing funds via charities and lobbying. In general the participants saw the best solutions in direct or indirect increased support from the governments and better interactions between agencies.

These responses helped to recognise additional facility types that might be relevant when looking for correlation with educational performance. Music and arts appeared repeatedly among the activities mentioned. Also outdoor spaces for play that are lit were mentioned. Regarding transport issues the OSM data contains quite detailed information about public transport for some areas so there is a potential to quantify such information and include in relevant facility types.

The number of responses received was 11. The number would be probably slightly higher if the questions were not limited to the age group of 8 to 11 years old. Some community centres replied via email stating that they do not provide services for children in this age group.

The questions along with the responses can be found in Appendix A.

7.6 Conclusions

This chapter provided evaluation of the experiment and its results and findings. Evaluation was presented reflecting the literature review conducted as part of this project. Potential problems regarding the results were recognised and discussed where appropriate. Separate section was devoted to evaluation of the online survey results.

Next chapter draws conclusions from the findings and elaborates on contribution to the body of knowledge, reflection on the project as well as future work and research this project may help to initiate.

8. CONCLUSIONS AND FUTURE WORK

8.1 Introduction

This chapter presents conclusions of the project and summarises its results and findings. It provides account for contribution to the body of knowledge, research overview and limitations, and suggests ideas for future work and research.

8.2 Problem Definition and Research Overview

Research in the area of the Semantic Web and the linked data indicates big potential in publishing data on the Web in a structured and machine readable form to support information discovery and inference. Structuring data based on its meaning and linking it to other related data in the same or even different data sets provides new opportunities for combining data from multiple sources and across diverse domains of interest. The idea of combining data from diverse sources is always appealing as it promises new ways of using data and gaining meaningful insights. On the other hand not many research projects present practical applications of linked data especially for analytical purposes. It seemed reasonable to explore possibility of combining linked data sources with those in other open formats and evaluate how they contribute to the analytical process and its outcomes.

From these reasons topic of pupils' educational achievement being correlated with their engagement in meaningful extra-curricular activities was selected for investigation. As a measure of opportunities for such an engagement availability of free time facilities in local areas was identified. To support discovery of facilities as entities placed in space a geographical data source was needed with sufficient coverage and being freely available. One of the Volunteered Geographic Information projects, namely OpenStreetMap, appeared as the right fit due to being available also as linked data.

The research focused on three areas: Linked Data, Volunteered Geographic Information and Social research. An experiment was designed to combine data from these diverse domains and to investigate potential correlation between availability of free time facilities for pupils and their educational achievements while at the same time trying to assess how link data impacts the analysis.

8.3 Contributions to Body of Knowledge

The project's contribution to the body of knowledge achieved as part of the dissertation can be considered as follows:

- Regarding Semantic Web and linked data especially from the point of its spatial dimension the project identified relatively large geographical area were in many thousands of objects no external links to other datasets were found.
- The project presents an unusual use case for Volunteered Geographic Information in terms of research.
- Combining educational performance data with Volunteered Geographic Information shows potential value of including data originating from crowed-sourced projects in social research.

8.4 Experimentation, Evaluation and Limitations

The experiment aimed to combine data from multiple sources to allow for investigation of correlation between availability of free time facilities for pupils and their educational achievements. It tried to make use of linked data and take advantage of the potentials the technology promises.

The data was successfully acquired though challenges regarding inconsistencies in the geographical data led to lengthy data cleaning step with not fully satisfying results. The sheer amount of retrieved data required deployment of additional technologies during the course of the project. It helped to eliminate certain amount of misclassified records and speed up aggregations in consequent steps.

The investigation of correlations was attempted in different ways. The partial results were identified as falsely rejecting null hypothesis of correlation coefficient being equal to 0 which was important to avoid drawing incorrect conclusions from the analysis. Possible way to remedy the issue was suggested. The experiment did not establish any significant correlation relationships.

The objective of assessing impact of linked data on the analysis was achieved by proving absence of external links from the linked data version of OSM data. It can be concluded that linked data was not ready yet in given settings to support the analysis with additional data.

Results of online survey conducted among community centres and primary schools as part of the project helped to set scores for individual facility types retrieved as well as recognise other facility types worth including in the analysis.

Another outcome of the experiment is a voluminous geographical data set of potential value for research in social domain being generated for area of England, UK that has references to the official educational performance data for the same area and thus can be easily merged as needed.

The scope of the experiment proved to be difficult to manage within the given project’s time constraints. Limiting the scope to smaller area in the design phase would have helped to avoid at least some of the experienced problems with geographical data retrieved.

Limitation of OSM data in terms of areas not being mapped sufficiently due to low interest of volunteer contributors could be supplemented by combining OSM data with data from other map providers like Google Maps and Bing Maps. Though these commercial providers impose certain limitations as how many requests a user can send for free per certain time units.

8.5 Future Work and Research

As this project was not able to verify usefulness of including linked data in the data analysis the future work targeted purely on such assessment could firstly establish for which geographical areas the LinkedGeoData as well as any other linked data source provides sufficient interlinking to other data sources on the Semantic Web to derive the scope of the consequent experiment.

To avoid problem with inconsistencies among OSM data in terms of how much is mapped and how well the objects are tagged between different regions it would advantageous in future works to firstly establish which geographical area (administrative region) has been mapped sufficiently enough for the purposes of a project and then limit the focus on that area. Regarding similar future projects the educational performance data is available for the whole England so OSM data fitness should dictate which regions of England would be included for investigation.

Although no correlations were successfully identified the problems encountered may serve as a guidance of what to avoid when including Volunteered Geographic Information data sources in the social research.

8.6 Conclusions

This chapter concluded the project investigating correlation between availability of free time facilities for pupils and their educational achievement by providing a short summary of the research and experiment conducted along with the results achieved. It discussed the limitations and possible future work and research areas.

The project generated an interesting geographical data set that might be of potential value for social research domain. The data is arranged to follow structure of the official educational performance data for England in terms of administrative areas which might serve as a basis for similar future work in this area. Moreover the project represents an example of including Volunteered Geographic Information in social domain related projects and shows one of the many ways this kind of data can be organised to fit purposes of social research.

In the settings of the project it was not possible to assess usefulness of including linked data in the analytical process. Selected linked data source did not provide sufficient interlinking to other data sources on the Semantic Web. Nevertheless with the continuing research in the area of linked data and the increasing number of data sets being published following the linked data principles it can be expected that more opportunities for exploiting the linked data potential will emerge in the future.

BIBLIOGRAPHY

Auer, S, Lehmann, J and Hellmann, S (2009) LinkedGeoData - Adding a Spatial Dimension to the Web of Data, in *8th International Semantic Web Conference, ISWC 2009, Chantilly, VA, USA, October 25-29, 2009. Proceedings*, Lecture Notes in Computer Science, Volume 5823, Presented at the 8th International Semantic Web Conference, ISWC 2009, Springer Berlin Heidelberg: Chantilly, VA, USA, 731–46. Available: http://link.springer.com/chapter/10.1007%2F978-3-642-04930-9_46 [Accessed 22 October 2013].

Barker, J E, Semenov, A D, Michaelson, L, Provan, L S, Snyder, H R and Munakata, Y (2014) Less-structured time in children's daily lives predicts self-directed executive functioning, *Frontiers in Psychology*, 5(593). Available from http://www.frontiersin.org/developmental_psychology/10.3389/fpsyg.2014.00593/abstract [Accessed 5 January 2015].

Berners-Lee, T (2006) Linked Data - Design Issues [online], Available: http://www.w3.org/DesignIssues/LinkedData.html [Accessed 18 December 2013].

Berners-Lee, T, Hendler, J and Lassila, O (2001) The Semantic Web, *Scientific American*, 284(5), 34–43. Available from http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21 [Accessed 20 November 2013].

Bizer, C, Heath, T and Berners-Lee, T (2009) Linked Data - The Story So Far, *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(3), 1–22. Available from http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/jswis.2009081901 [Accessed 7 January 2014].

Byrne, T, Nixon, E, Mayock, P and Whyte, J (2006) *Free-Time and Leisure Needs of Young People Living in Disadvantaged Communities*, Combat Poverty Agency, Dublin, Ireland. Available: https://www.tcd.ie/childrensresearchcentre/assets/pdf/Publications/freetime.pdf [Accessed 26 October 2014].

Capadisli, S, Auer, S and Riedl, R (2013) Linked Statistical Data Analysis, Presented at the Semantic Web Challenge. Available: http://challenge.semanticweb.org/2013/submissions/swc2013_submission_7.pdf [Accessed 15 March 2014].

Chen, H, Wu, Z and Cudre-Mauroux, P (2012) Semantic Web Meets Computational Intelligence: State of the Art and Perspectives [Review Article], *Computational Intelligence Magazine, IEEE*, 7(2), 67–74. Available from http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6183732 [Accessed 5 December 2013].

Ciepłuch, B, Jacob, R, Mooney, P and Winstanley, A (2010) Comparison of the accuracy of OpenStreetMap for Ireland with Google Maps and Bing Maps, in *Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resuorces and Enviromental Sciences*, Presented at the Accuracy 2010 Conference, Leicester, UK, 337. Available: http://www.spatial-accuracy.org/CipeluchAccuracy2010 [Accessed 20 November 2014].

Coleman, D J, Georgiadou, Y and Labonte, J (2009) Volunteered Geographic Information: The Nature and Motivation of Produsers, *International Journal of Spatial Data Infrastructures Research*, 4, 332–58. Available from http://ijsdir.jrc.ec.europa.eu/index.php/ijsdir/article/view/140/223 [Accessed 31 January 2015].

Eccles, J S and Barber, B L (1999) Student Council, Volunteering, Basketball, or Marching Band: What Kind of Extracurricular Involvement Matters?, *Journal of Adolescent Research*, 14(1), 10–43. Available from http://jar.sagepub.com/content/14/1/10 [Accessed 10 May 2014].

Egenhofer, M J (2002) Toward the Semantic Geospatial Web, in *Proceedings of the 10th ACM International Symposium on Advances in Geographic Information Systems*, GIS ’02, ACM: McLean, Virginia, USA, 1–4. Available: http://doi.acm.org/10.1145/585147.585148 [Accessed 29 November 2013].

Fuligni, A J and Stevenson, H W (1995) Time Use and Mathematics Achievement among American, Chinese, and Japanese High School Students, *Child Development*, 66(3), 830–42. Available from http://dx.doi.org/10.1111/j.1467-8624.1995.tb00908.x [Accessed 20 January 2015].

Goodchild, M (2007) Citizens as sensors: the world of volunteered geography, *GeoJournal*, 69(4), 211–21. Available from http://link.springer.com/article/10.1007%2Fs10708-007-9111-y [Accessed 5 April 2014].

Haklay, M (2010) How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets, *Environment and Planning B: Planning and Design*, 37(4), 682–703. Available from http://ideas.repec.org/a/pio/envirb/v37y2010i4p682-703.html [Accessed 31 January 2015].

Haklay, M and Weber, P (2008) OpenStreetMap: User-Generated Street Maps, *Pervasive Computing, IEEE*, 7(4), 12–8. Available from http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=4653466 [Accessed 5 November 2014].

Hausenblas, M (2009) Exploiting Linked Data to Build Web Applications, *Internet Computing, IEEE*, 13(4), 68–73. Available from http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=5167270 [Accessed 4 December 2014].

Janowicz, K, Scheider, S, Pehle, T and Hart, G (2012) Geospatial Semantics and Linked Spatiotemporal Data - Past, Present, and Future, *Semantic Web*, 3(4), 321–32. Available from http://iospress.metapress.com/content/m065w1130043w3p4 [Accessed 6 October 2013].

Kalampokis, E, Tambouris, E and Tarabanis, K (2013) Linked Open Government Data Analytics, in Wimmer, M., Janssen, M. and Scholl, H., eds., *Electronic Government*, Lecture Notes in Computer Science, Springer Berlin Heidelberg, 99–110. Available: http://dx.doi.org/10.1007/978-3-642-40358-3_9 [Accessed 9 April 2014].

Kämpgen, B (2011) DC Proposal: Online Analytical Processing of Statistical Linked Data, in *Proceedings of the 10th International Conference on The Semantic Web - Volume Part II*, ISWC'11, Springer-Verlag: Berlin, Heidelberg, 301–8. Available: http://dl.acm.org/citation.cfm?id=2063076.2063100 [Accessed 26 March 2014].

Karam, R and Melchiori, M (2013) Improving Geo-spatial Linked Data with the Wisdom of the Crowds, in *Proceedings of the Joint EDBT/ICDT 2013 Workshops*, EDBT '13, ACM: New York, NY, USA, 68–74. Available: http://doi.acm.org/10.1145/2457317.2457329 [Accessed 30 October 2014].

McGuinness, D L and Van Harmelen, F (2004) OWL Web Ontology Language Overview [online], Available: http://www.w3.org/TR/owl-features/ [Accessed 15 November 2013].

Van der Merwe, H (2014) 'Do what you can with what you have where you are': Extracurricular provisioning in an inner-city environment, *South African Journal for Research in Sport, Physical Education and Recreation*, 36(2), 195–210. Available from http://repository.up.ac.za/handle/2263/43622 [Accessed 20 January 2015].

Mooney, P and Corcoran, P (2012) Characteristics of Heavily Edited Objects in OpenStreetMap, *Future Internet*, 4(1), 285–305. Available from http://www.mdpi.com/1999-5903/4/1/285 [Accessed 11 December 2014].

Mooney, P, Corcoran, P, Sun, H and Yan, L (2012) Citizen Generated Spatial Data and Information: Risks and Opportunities, in *Industrial Control and Electronics Engineering (ICICEE), 2012 International Conference on*, 1990–3. Available: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6322820 [Accessed 30 October 2014].

Mülligann, C, Janowicz, K, Ye, M and Lee, W-C (2011) Analyzing the Spatial-Semantic Interaction of Points of Interest in Volunteered Geographic Information, in Egenhofer, M., Giudice, N., Moratz, R. and Worboys, M., eds., *Spatial Information Theory*, Lecture Notes in Computer Science, Springer Berlin Heidelberg, 350–70. Available: http://dx.doi.org/10.1007/978-3-642-23196-4_19 [Accessed 30 April 2014].

Osgood, D W ., Anderson, A L . and Shaffer, J N . (2005) *Unstructured Leisure in the after-School Hours* [online], Organized activities as contexts of development:

Extracurricular activities, after-school and community programs., Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-58149230899&partnerID=40&md5=67d7f5e11fc452c33365bc4f14ac27bc [Accessed 20 January 2015].

Shadbolt, N, O'Hara, K, Berners-Lee, T, Gibbins, N, Glaser, H, Hall, W and Schraefel, M C (2012) Linked Open Government Data: Lessons from Data.gov.uk, *Intelligent Systems, IEEE*, 27(3), 16–24. Available from http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6171150 [Accessed 15 December 2013].

Stadler, C, Lehmann, J, Höffner, K and Auer, S (2012) LinkedGeoData: A Core for a Web of Spatial Open Data, *Semantic Web*, 3(4), 333–54. Available from http://dx.doi.org/10.3233/SW-2011-0052 [Accessed 5 November 2013].

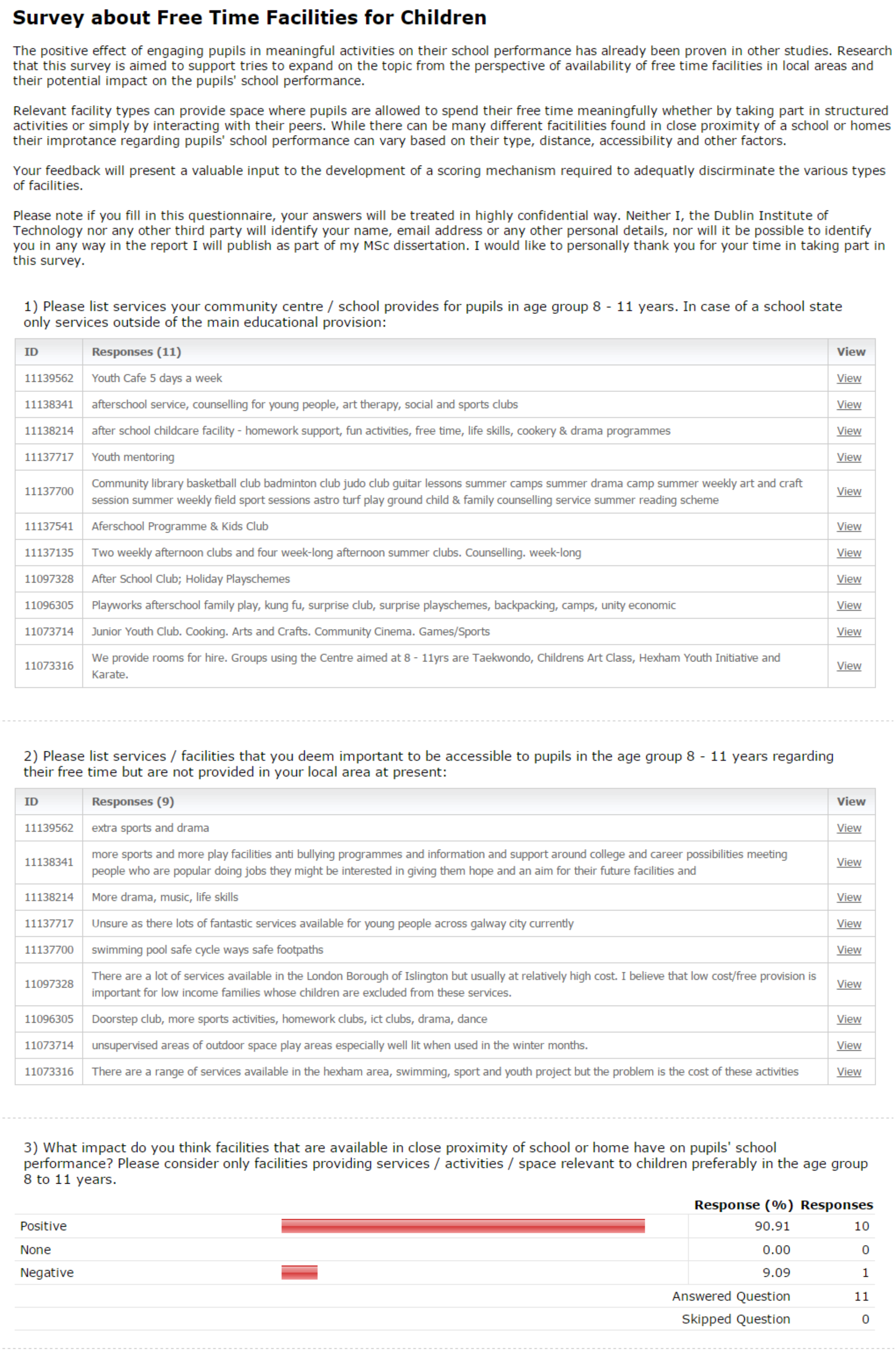
Wegner, L, Flisher, A J, Chikobvu, P, Lombard, C and King, G (2008) Leisure boredom and high school dropout in Cape Town, South Africa, *Journal of Adolescence*, 31(3), 421–31. Available from http://www.sciencedirect.com/science/article/pii/S0140197107000930 [Accessed 7 May 2014].

World Youth Report (2003) in The Global Situation of Young People, Department of Economic and Social Affairs, United Nations: NY, 213–47. Available: http://www.un.org/esa/socdev/unyin/documents/worldyouthreport.pdf [Accessed 6 May 2014].

Zapilko, B, Harth, A and Mathiak, B (2011) Enriching and Analysing Statistics with Linked Open Data, Presented at the Eurostat (Hrsg.), NTTS - Conference on New Techniques and Technologies for Statistics, Brüssel. Available: http://www.researchgate.net/publication/231336781_Enriching_and_Analysing_Statistics_with_Linked_Open_Data [Accessed 2 April 2014].

APPENDIX A - SURVEY

Survey questions along with responses received:



Survey about Free Time Facilities for Children

The positive effect of engaging pupils in meaningful activities on their school performance has already been proven in other studies. Research that this survey is aimed to support tries to expand on the topic from the perspective of availability of free time facilities in local areas and their potential impact on the pupils' school performance.

Relevant facility types can provide space where pupils are allowed to spend their free time meaningfully whether by taking part in structured activities or simply by interacting with their peers. While there can be many different facitilities found in close proximity of a school or homes their improtance regarding pupils' school performance can vary based on their type, distance, accessibility and other factors.

Your feedback will present a valuable input to the development of a scoring mechanism required to adequatly discriminate the various types of facilities.

Please note if you fill in this questionnaire, your answers will be treated in highly confidential way. Neither I, the Dublin Institute of Technology nor any other third party will identify your name, email address or any other personal details, nor will it be possible to identify you in any way in the report I will publish as part of my MSc dissertation. I would like to personally thank you for your time in taking part in this survey.

1) Please list services your community centre / school provides for pupils in age group 8 - 11 years. In case of a school state only services outside of the main educational provision:

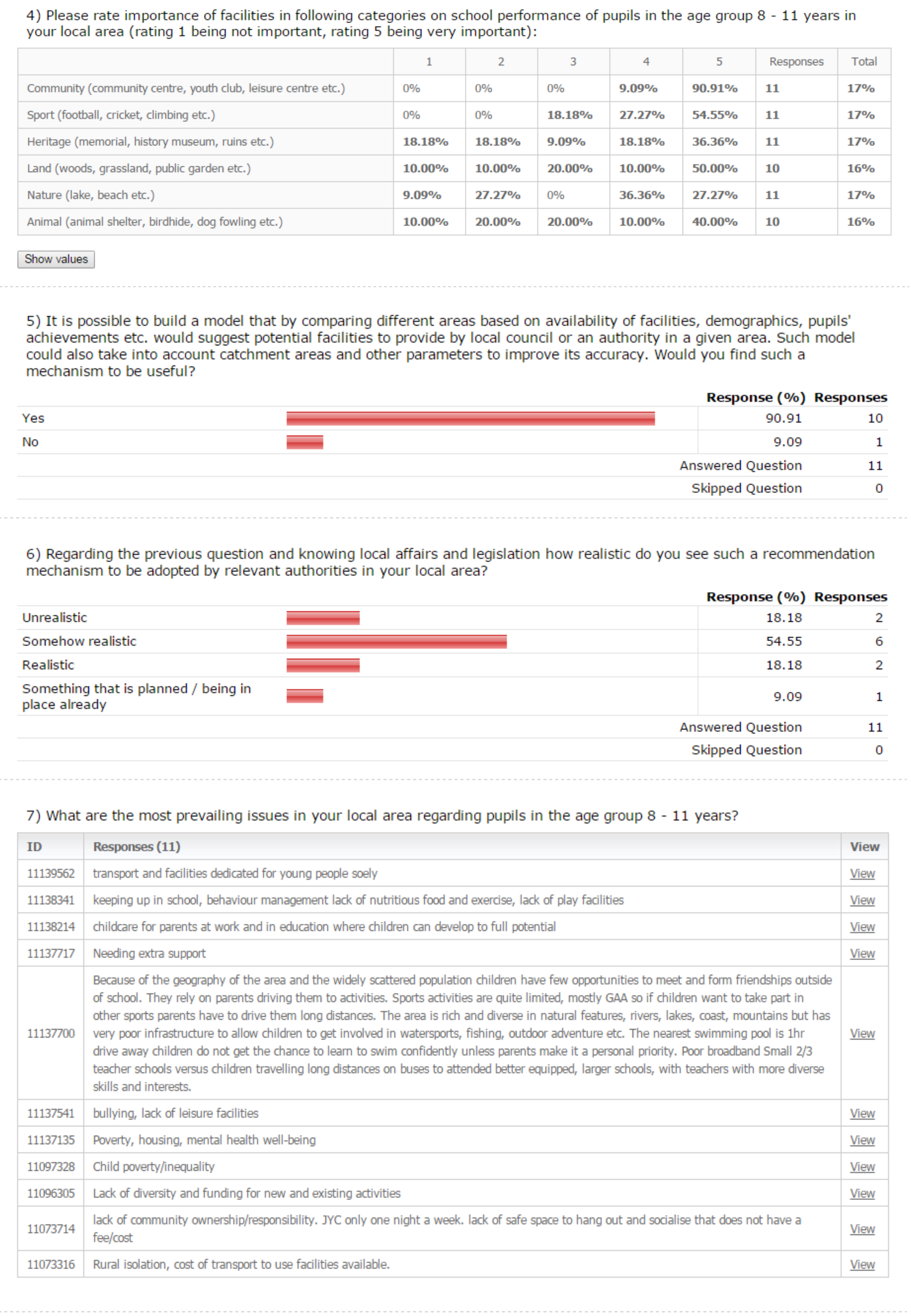
ID	Responses (11)	View
11139562	Youth Cafe 5 days a week	View
11138341	afterschool service, counselling for young people, art therapy, social and sports clubs	View
11138214	after school childcare facility - homework support, fun activities, free time, life skills, cookery & drama programmes	View
11137717	Youth mentoring	View
11137700	Community library basketball club badminton club judo club guitar lessons summer camps summer drama camp summer weekly art and craft session summer weekly field sport sessions astro turf play ground child & family counselling service summer reading scheme	View
11137541	Aferschool Programme & Kids Club	View
11137135	Two weekly afternoon clubs and four week-long afternoon summer clubs. Counselling. week-long	View
11097328	After School Club; Holiday Playschemes	View
11096305	Playworks afterschool family play, kung fu, surprise club, surprise playschemes, backpacking, camps, unity economic	View
11073714	Junior Youth Club. Cooking. Arts and Crafts. Community Cinema. Games/Sports	View
11073316	We provide rooms for hire. Groups using the Centre aimed at 8 - 11yrs are Taekwondo, Childrens Art Class, Hexham Youth Initiative and Karate.	View

2) Please list services / facilities that you deem important to be accessible to pupils in the age group 8 - 11 years regarding their free time but are not provided in your local area at present:

ID	Responses (9)	View
11139562	extra sports and drama	View
11138341	more sports and more play facilities anti bullying programmes and information and support around college and career possibilities meeting people who are popular doing jobs they might be interested in giving them hope and an aim for their future facilities and	View
11138214	More drama, music, life skills	View
11137717	Unsure as there lots of fantastic services available for young people across galway city currently	View
11137700	swimming pool safe cycle ways safe footpaths	View
11097328	There are a lot of services available in the London Borough of Islington but usually at relatively high cost. I believe that low cost/free provision is important for low income families whose children are excluded from these services.	View
11096305	Doorstep club, more sports activities, homework clubs, ict clubs, drama, dance	View
11073714	unsupervised areas of outdoor space play areas especially well lit when used in the winter months.	View
11073316	There are a range of services available in the hexham area, swimming, sport and youth project but the problem is the cost of these activities	View

3) What impact do you think facilities that are available in close proximity of school or home have on pupils' school performance? Please consider only facilities providing services / activities / space relevant to children preferably in the age group 8 to 11 years.

		Response (%)	Responses
Positive		90.91	10
None		0.00	0
Negative		9.09	1
		Answered Question	11
		Skipped Question	0



4) Please rate importance of facilities in following categories on school performance of pupils in the age group 8 - 11 years in your local area (rating 1 being not important, rating 5 being very important):

	1	2	3	4	5	Responses	Total
Community (community centre, youth club, leisure centre etc.)	0%	0%	0%	**9.09%**	**90.91%**	**11**	**17%**
Sport (football, cricket, climbing etc.)	0%	0%	**18.18%**	**27.27%**	**54.55%**	**11**	**17%**
Heritage (memorial, history museum, ruins etc.)	**18.18%**	**18.18%**	**9.09%**	**18.18%**	**36.36%**	**11**	**17%**
Land (woods, grassland, public garden etc.)	**10.00%**	**10.00%**	**20.00%**	**10.00%**	**50.00%**	**10**	**16%**
Nature (lake, beach etc.)	**9.09%**	**27.27%**	0%	**36.36%**	**27.27%**	**11**	**17%**
Animal (animal shelter, birdhide, dog fowling etc.)	**10.00%**	**20.00%**	**20.00%**	**10.00%**	**40.00%**	**10**	**16%**

Show values

5) It is possible to build a model that by comparing different areas based on availability of facilities, demographics, pupils' achievements etc. would suggest potential facilities to provide by local council or an authority in a given area. Such model could also take into account catchment areas and other parameters to improve its accuracy. Would you find such a mechanism to be useful?

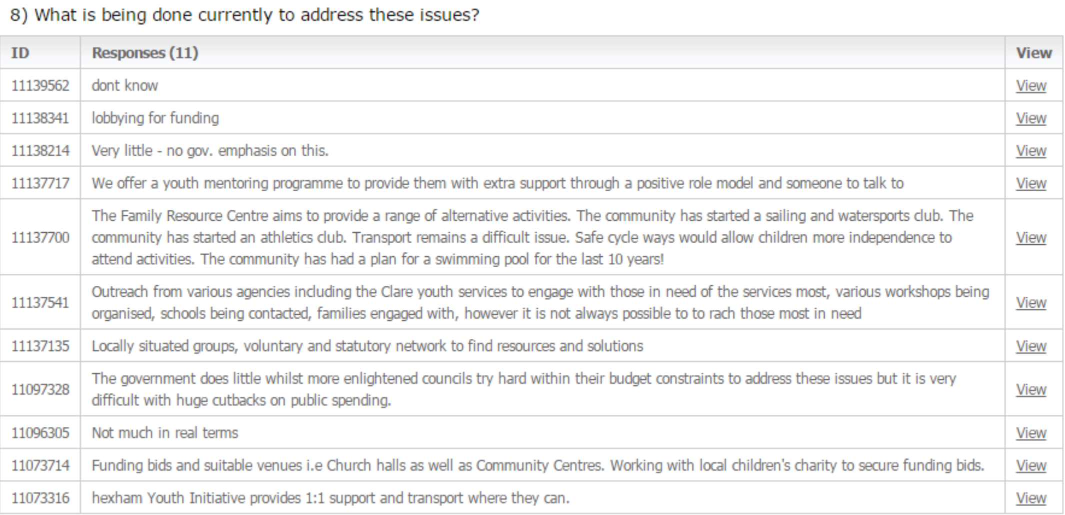
	Response (%)	Responses
Yes	90.91	10
No	9.09	1
	Answered Question	11
	Skipped Question	0

6) Regarding the previous question and knowing local affairs and legislation how realistic do you see such a recommendation mechanism to be adopted by relevant authorities in your local area?

	Response (%)	Responses
Unrealistic	18.18	2
Somehow realistic	54.55	6
Realistic	18.18	2
Something that is planned / being in place already	9.09	1
	Answered Question	11
	Skipped Question	0

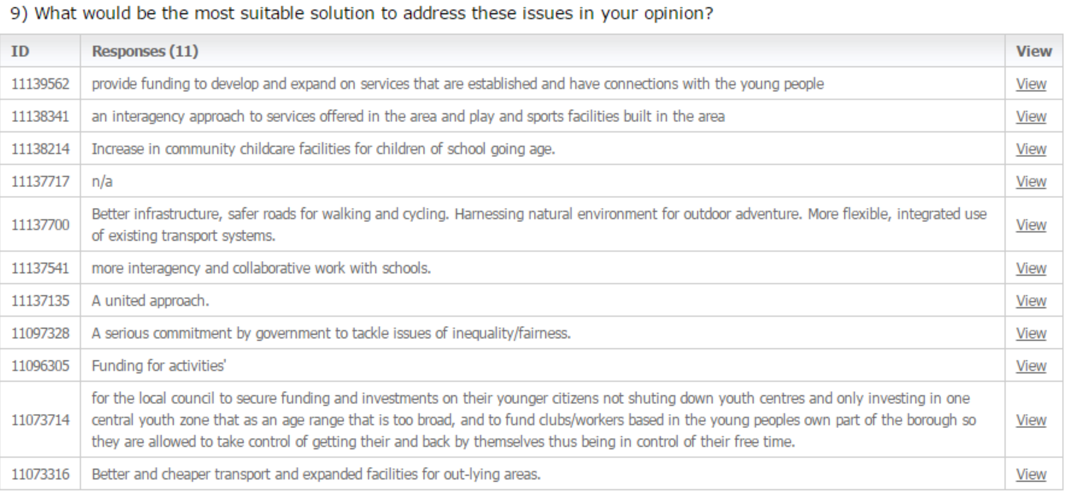
7) What are the most prevailing issues in your local area regarding pupils in the age group 8 - 11 years?

ID	Responses (11)	View
11139562	transport and facilities dedicated for young people soely	View
11138341	keeping up in school, behaviour management lack of nutritious food and exercise, lack of play facilities	View
11138214	childcare for parents at work and in education where children can develop to full potential	View
11137717	Needing extra support	View
11137700	Because of the geography of the area and the widely scattered population children have few opportunities to meet and form friendships outside of school. They rely on parents driving them to activities. Sports activities are quite limited, mostly GAA so if children want to take part in other sports parents have to drive them long distances. The area is rich and diverse in natural features, rivers, lakes, coast, mountains but has very poor infrastructure to allow children to get involved in watersports, fishing, outdoor adventure etc. The nearest swimming pool is 1hr drive away children do not get the chance to learn to swim confidently unless parents make it a personal priority. Poor broadband Small 2/3 teacher schools versus children travelling long distances on buses to attended better equipped, larger schools, with teachers with more diverse skills and interests.	View
11137541	bullying, lack of leisure facilities	View
11137135	Poverty, housing, mental health well-being	View
11097328	Child poverty/inequality	View
11096305	Lack of diversity and funding for new and existing activities	View
11073714	lack of community ownership/responsibility. JYC only one night a week. lack of safe space to hang out and socialise that does not have a fee/cost	View
11073316	Rural isolation, cost of transport to use facilities available.	View



8) What is being done currently to address these issues?

ID	Responses (11)	View
11139562	dont know	View
11138341	lobbying for funding	View
11138214	Very little - no gov. emphasis on this.	View
11137717	We offer a youth mentoring programme to provide them with extra support through a positive role model and someone to talk to	View
11137700	The Family Resource Centre aims to provide a range of alternative activities. The community has started a sailing and watersports club. The community has started an athletics club. Transport remains a difficult issue. Safe cycle ways would allow children more independence to attend activities. The community has had a plan for a swimming pool for the last 10 years!	View
11137541	Outreach from various agencies including the Clare youth services to engage with those in need of the services most, various workshops being organised, schools being contacted, families engaged with, however it is not always possible to to rach those most in need	View
11137135	Locally situated groups, voluntary and statutory network to find resources and solutions	View
11097328	The government does little whilst more enlightened councils try hard within their budget constraints to address these issues but it is very difficult with huge cutbacks on public spending.	View
11096305	Not much in real terms	View
11073714	Funding bids and suitable venues i.e Church halls as well as Community Centres. Working with local children's charity to secure funding bids.	View
11073316	hexham Youth Initiative provides 1:1 support and transport where they can.	View



9) What would be the most suitable solution to address these issues in your opinion?

ID	Responses (11)	View
11139562	provide funding to develop and expand on services that are established and have connections with the young people	View
11138341	an interagency approach to services offered in the area and play and sports facilities built in the area	View
11138214	Increase in community childcare facilities for children of school going age.	View
11137717	n/a	View
11137700	Better infrastructure, safer roads for walking and cycling. Harnessing natural environment for outdoor adventure. More flexible, integrated use of existing transport systems.	View
11137541	more interagency and collaborative work with schools.	View
11137135	A united approach.	View
11097328	A serious commitment by government to tackle issues of inequality/fairness.	View
11096305	Funding for activities'	View
11073714	for the local council to secure funding and investments on their younger citizens not shuting down youth centres and only investing in one central youth zone that as an age range that is too broad, and to fund clubs/workers based in the young peoples own part of the borough so they are allowed to take control of getting their and back by themselves thus being in control of their free time.	View
11073316	Better and cheaper transport and expanded facilities for out-lying areas.	View