A Framework for the Analysis and User-Driven Evaluation of Trust on the Semantic Web

Peter Clarke

A dissertation submitted in partial fulfilment of the requirements of
Dublin Institute of Technology for the degree of
M.Sc. in Computing (Information and Knowledge Management)

March 2014

DECLARATION

I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Information and Knowledge Management), 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: ***24th March 2014***

ABSTRACT

This project will examine the area of trust on the Semantic Web and develop a framework for publishing and verifying trusted Linked Data.

Linked Data describes a method of publishing structured data, automatically readable by computers, which can linked to other heterogeneous data with the purpose of becoming more useful.

Trust plays a significant role in the adoption of new technologies and even more so in a sphere with such vast amounts of publicly-created data. Trust is paramount to the effective sharing and communication of tacit knowledge (Hislop, 2013). Up to now, the area of trust in Linked Data has not been adequately addressed, despite the Semantic Web stack having included a trust layer from the very beginning (Artz and Gil, 2007).

Some of the most accurate data on the Semantic Web lies practically unused, while some of the most used linked data has high numbers of errors (Zaveri et al., 2013). Many of the datasets and links that exist on the Semantic Web are out of date and/or invalid and this undermines the credibility and validity, and ultimately, the trustworthiness of both the dataset and the data provider (Rajabi et al., 2012).

This research will examine a number of datasets to determine the quality metrics that a dataset is required to meet to be considered ‘trusted’. The key findings will be assessed and utilized in the creation of a learning tool and a framework for creating trusted Linked Data.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my dissertation supervisor, Damian Gordon, whose enthusiasm, optimism and encouragement consistently inspired me throughout the dissertation process. His wise words, guidance and feedback contributed to making this a mostly enjoyable journey.

Thank you to my wife, Antoinette, and my children, Pearl and James. Their immense patience and support allowed for me to reach this stage of my studies.

Thanks also go to my colleagues in UCD Library for the help and support provided throughout the dissertation process. Thank you also to the group who participated in the survey and subsequent interviews.

Finally, I would like to thank my family, especially my mother and father and all the friends who have supported me throughout this process.

For Brigid.

TABLE OF CONTENTS

1. INTRODUCTION

1.1 Background

There can be no denying that the ways in which we share knowledge and information have been transformed by the emergence of the Web. The barriers that once existed in publishing and consuming information have been lowered, replaced with user-oriented search engines offering customized searches and inferred results based on machine-learned knowledge.

Commonly, when data has been published on the Web it has been made available as raw formats such as XML, CSV or marked up with HTML. The negative effect of this is that almost all of the structure and meaning, or *semantics*, of this data is stripped out and lost. The *Semantic Web* aims to create the Web of Data, as an extension of the existing Web of Documents. It can be seen as a set of best practices for sharing data over the Web for use by applications (DuCharme, 2011). Linked Data emerged from this grand idea, the fruit of a desire for a more practical attitude with a reduced emphasis on semantics (Heath and Bizer, 2011). Bizer et al. see the Semantic Web as the end goal with Linked Data seen as providing the means to reach that goal (Bizer et al., 2009).

The field of Information and Knowledge Management is concerned with the representation, organization, acquisition, creation and use of information and knowledge (Jurisica et al., 2004). The linked data lifecycle (fig. 1) mirrors this definition (Villazón-Terrazas et al., 2011). Therefore, the techniques chosen for both acquisition and representation together with the quality of their application can determine to what degree a particular endeavor will succeed. These ontological representations operate as a surrogate for real-world entities (Davis et al., 1993), by explicitly expressing the concepts and relationships of Linked Data (Stroka, 2010). Ontologies are therefore paramount in describing the structure and semantics of data (Fensel, 2003).

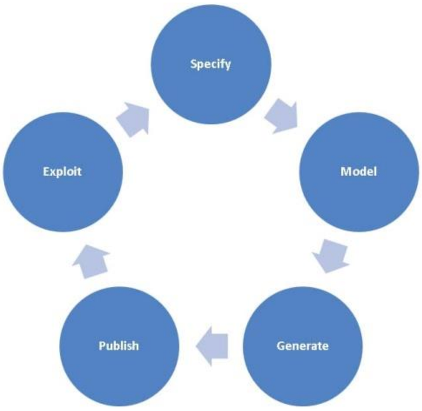


Fig. 1.1: Linked Data lifecycle (Villazón-Terrazas et al., 2011)

The growth of Linked Data is undeniable. Between 2007 and September 2010, 203 datasets were published containing almost 27 billion RDF triples, of which 395 million were RDF links (Bizer et al., 2010). By the following year, this had risen to 295 datasets, 31 billion triples and 503 million RDF links (Bizer et al., 2011). This rise in the number of datasets being published would indicate that Linked Data is widely seen to be a step in the right direction. In recent times, many library institutions such as the Library of Congress (Library of Congress, 2012) and WorldCat (Dishongj, 2012) have published large datasets of Linked (Open) Data.

While there is visible growth in the Linked Data cloud (fig. 2), a number of concerns are raised regarding its usage. Semantic Web technologies have existed for a number of years, however the availability of these tools has had only modest impact on the development of real world applications to date (Hausenblas, 2009). In a study by Moller et al, examining a number of large LOD datasets, it was seen that there has been no increase in the requests for semantic data (Möller et al., 2010). Hausenblas and Karnstedt contend that an understanding of the requirements and the challenges concerning the use of Linked Data is absent (Hausenblas and Karnstedt, 2010). With such tremendous growth in freely accessible interconnected data across a broad range of disciplines, the potential of this vast universe of data has, to date, been left unexploited (Pedrinaci and Domingue, 2011).

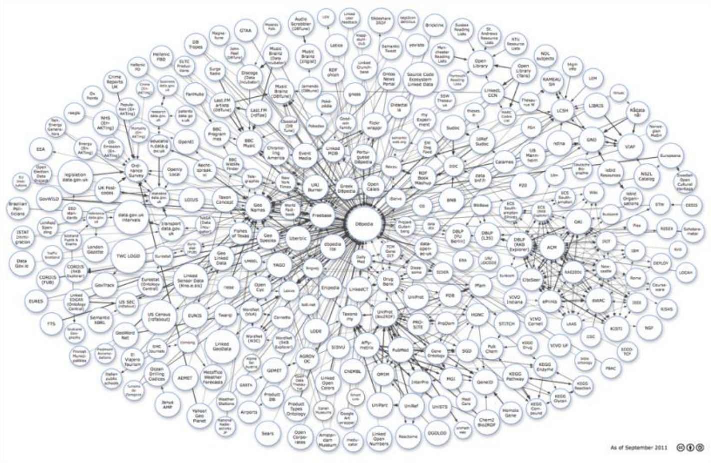


Fig. 1.2: Linked Data cloud, 2011

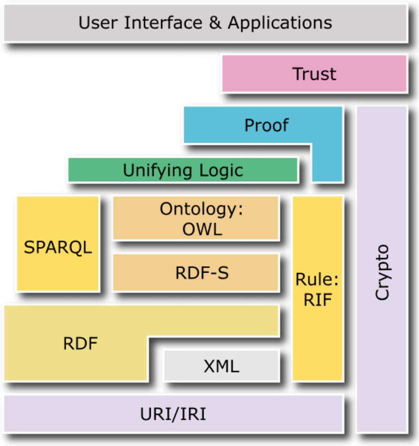


Fig. 1.3: Semantic Web technology stack

Trust plays a hugely significant role in the adoption of new technologies and even more so in a sphere with such vast amounts of publicly-created data. Trust is paramount to the effective sharing and communication of tacit knowledge (Hislop, 2013). It is defined as the belief an entity has in the behavior of others and the assumption that they will honor their obligations. Up to now, the area of trust in Linked Data has not been adequately addressed, despite the Semantic Web stack (fig. 3) having included a trust layer from the very beginning (Artz and Gil, 2007).

Many of the datasets and links that exist on the Semantic Web are out of date and/or invalid and this undermines the credibility and validity, und ultimately, the trustworthiness of both the dataset and the data provider (Rajabi et al., 2012). Datasets should provide users with a means to assess the trustworthiness of the data within. This raises many questions on the provenance, reliability and believability of the data. Therefore, to answer these questions we need to assess trustworthiness of data.

This research hopes to examine a number of datasets to determine the quality metrics that a dataset is required to meet to be considered 'trusted'. The key findings will be assessed and utilized in the creation of an application which evaluates the trust rating of a dataset and will be published to the web alongside a framework for creating trusted Linked Data.

1.2 Description

The principles of Linked Data are widely documented (Berners-Lee, 2009). In 2009, Tim Berners-Lee published a list of five attributes (“five stars”) that all linked data should possess for it to be truly considered ‘linked’ (Berners-Lee, 2009). This was subsequently amended in 2010, with a note suggesting the requirement for a sixth property, related to providing metadata for this linked data. Clearly, the quality, characteristics and challenges of linked data are still evolving.

To examine this further, Pipino et al. suggest an approach based on an objective assessment of the data using predefined criteria, or a subjective assessment of how the data has been put to use (Pipino et al., 2002). Opinion is divided within the linked data community on precise Linked Data quality metrics (semanticweb.com, 2011). Despite this, many agree on data being assessed subjectively, citing Chapman’s espousal of quality being a measure of fitness for use in a specific application (Chapman, 2005). This mirrors the point made previously by Wang and Strong (Wang and Strong, 1996).

This is strongly aligned with the Linked Data spirit of focusing on the "what" and "why" of semantic relationships rather than the "how". Linked Data is concerned with using the Web to connect related data that wasn't previously linked, or using the Web to lower the barriers to linking data currently linked using other methods. More explicitly, Wikipedia defines Linked Data as "a term used to describe a recommended best practice for exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web using URIs and RDF" (Wikipedia, 2013).

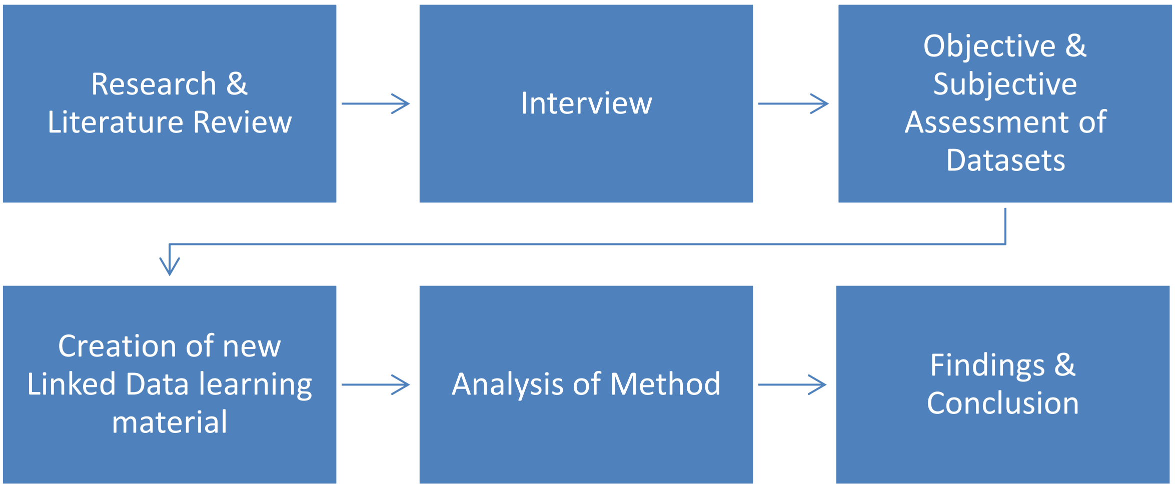
The library field has significant familiarity with being a producer of high-quality structured data, naturally complementing the area of Linked Data (Heath and Bizer, 2011). Highly-curated and highly-trusted library linked datasets, such as OCLC WorldCat or Europeana data, represent a model for trusted linked data. It is hoped that by comparing these resources with those more freely contributed to (or *crowd-sourced*) by the general population, but less-trusted, such as DBpedia, it will be possible to ascertain the characteristics of trusted datasets and develop an understanding of the principles of trust that are at work.

Zaveri *et al.* have identified a number of trust dimensions which should be examined when determining the trustworthiness of Linked Data (Zaveri et al., 2012). Examples of these metrics include *provenance*, *verifiability*, *reputation*, *believability* and *licensing*. These combine both objective and subjective quality metrics and represent a thorough analysis of the trustworthiness of a dataset. Some of the metrics which could be examined include:

> Provenance: This relates to contextual metadata that details how data is represented and managed and, importantly, the origin of the source. In examining provenance, we are assessing the trustworthiness, credibility and reliability of the data which will lead to trusted data being adopted and used further. This can be evaluated by both objective and subjective means.
>
> Verifiability: This is the *"degree and ease with which the information can be checked for correctness"* (Bizer and Cyganiak, 2011). Trusted data is data which has been verified to be correct. In many instances verifiability can be measured objectively but subjective assessment is also valuable. Verifiability can be examined by an unbiased third party or by employing digital signatures.
>
> Reputation: This is a subjective judgement made by a user or group of users, determining the integrity of the data source. Often, a survey of a community is used to define the reputation of a data provider. Based on this reputation score, the user makes a judgement on the trustworthiness of the data presented.
>
> Licensing: This is the granting of permissions to reuse the dataset under specific conditions. This is closely linked to provenance and encourages trust and reuse by informing data consumers of their legal rights in using this data.

Thorough research into the field of Linked Data and a comprehensive literature review will be conducted as a preliminary stage. Following this, interviews of a number of Library Linked Data experts will be conducted with an emphasis on determining the characteristics of trusted data. It is hoped that these interviews, combined with the outcomes of the initial research will shape the design of learning material which can be used to assist the creation of trusted Linked Data.

The application will evaluate the aforementioned datasets by taking random samples of RDF data from each dataset and rating them against these Linked Data trust metrics through user interaction. It should be relatively straightforward to measure much of the data objectively and subjective assessment of the data can be examined in the form of weighted questions.


Research & Literature Review
Interview
Objective & Subjective Assessment of Datasets
Creation of new Linked Data learning material
Analysis of Method
Findings & Conclusion

By examining and assessing these datasets using these metrics it is hoped that a framework or trust maturity model, akin to Tim Berners-Lee's '5 Star' model, can be developed and published on the web. This could lead to the development of the notion of a '*Trusted Data Seal of Approval*' which could be used by data providers to enhance their data and reputation but also act as verification of data quality by parties considering using a particular dataset. This would serve the purpose of increasing both data usage and data trust, while creating a feedback loop which enhances the Semantic Web generally.

1.3 Aims and Objectives

The aim of the project is to assess and evaluate the features of trusted, quality Linked Data. Through the effective execution of a suitable experiment this research will detail the characteristics of trusted Linked Data datasets and summarise these into a framework that can be reused in the creation of trusted linked data.

1. Review the Semantic Web landscape
2. Investigate the standards and tools required to produce, manipulate and exploit this data
3. Investigate the current research in the field of Linked Data
4. Survey and interview expert within the field of Linked Data
5. Develop experiment to ascertain appropriate trust metrics for quality Linked Data
6. Develop learning material in conjunction with data trust metrics
7. Document and evaluate the findings of this experiment
8. Make recommendations for further research in the field

1.4 Thesis Roadmap

Chapters 2 and 3 provide the main literature review for this research. Chapter 2 explains and introduces the ideas of the Semantic Web and Linked Data, and their relationship to Knowledge Management. Chapter 3 explores the notion of trust in semantically marked-up data.

Chapter 4 explores the nature of believability in Linked Data and identifies the five datasets that will be used as part of this experiment.

Chapter 5 discussed the technological deployment of the Virtuoso SPARQL triplestore and explains how to load data into the system.

Chapter 6 outlines the survey that was undertaken to assess peoples' general understanding of trustworthiness in Linked Data, and helps support findings in existing literature.

Chapter 7 presents the technology-oriented assessment of the datasets using the Virtuoso system to explore objectively-measureable characteristics of the datasets.

Chapter 8 focuses on the development of a framework embodied as instructional materials to capture some of the key "knowledge gaps" that exist in the development of Linked Datasets.

Finally, Chapter 9 presents the conclusions of this research and some future directions that this research may be taken in.

2. KNOWLEDGE MANAGEMENT AND THE SEMANTIC WEB

2.1 Introduction

Until recently, much of the data published on the Web has been made available in raw document formats such as XML, CSV or text, marked up with HTML. The negative effect of this is that almost all of the structure and meaning, or *semantics*, of this data is stripped out and lost. The Semantic Web aims to create the *Web of Data*, as an extension of the existing *Web of Documents*. The Semantic Web can be seen as a set of best practices for sharing data over the Web for use by applications (DuCharme, 2011). That is, to make the web more accessible to computers.

In order to make this Web of Data a reality, it is first necessary to publish large amounts of data on the Web, making this available in a standardized format, accessible and manageable by Semantic Web tools. To avoid simply creating a large collection of datasets, it is necessary to make the relationships between the data available also. This collection of interlinked datasets available on the Web is known as *Linked Data*. Linked *Open* Data *(LOD)* is Linked Data that is published under an open license.

Linked Data can be seen as a reference implementation of the Semantic Web, providing *"a publishing paradigm in which not only documents, but also data, can be a first class citizen of the Web"* (Heath and Bizer, 2011). Bizer et al. see the Semantic Web as the end goal with Linked Data seen as enabling the means to reach that goal (Bizer *et al.*, 2009). In 2009, Tim Berners-Lee introduced a *5 Star rating system* for publishing data on the Semantic Web and suggested data publishers follow these design principles (Berners-Lee, 2009).

The Semantic Web enables a new frontier of decentralized knowledge management by enhancing information flow with machine-processable metadata (Cayzer, 2004). Since the vision for the Semantic Web was explicitly laid out in 2000 (Berners-Lee, 2000), Semantic Web technologies have undergone rapid advancement and the Semantic Web community has witnessed tremendous growth in scale and diversity.

2.2 The State of the LOD Cloud

The growth of Linked Data is undeniable. Between 2007 and September 2010, 203 datasets were published containing almost 27 billion RDF triples, of which 395 million were RDF links (Bizer et al., 2010). By the following year, this had risen to 295 datasets, 31 billion triples and 503 million RDF links (Bizer et al., 2011). This rise in the number of datasets being published indicates that Linked Data is widely seen to be a step in the right direction. In recent times, many library institutions such as the Library of Congress (Library of Congress, 2012) and WorldCat (Dishongj, 2012) have published large datasets of Linked (Open) Data.

While there is visible growth in the Linked Data cloud, a number of concerns are raised regarding its usage. Semantic Web technologies have existed for a number of years, however the availability of these tools has had only modest impact on the development of real world applications to date (Hausenblas, 2009). In a study by Moller et al, examining a number of large LOD datasets, it was seen that there has been no increase in the requests for semantic data (Möller *et al.*, 2010). Hausenblas and Karnstedt contend that an understanding of the requirements and the challenges concerning the use of Linked Data is absent (Hausenblas and Karnstedt, 2010). With such tremendous growth in freely accessible interconnected data across a broad range of disciplines, the potential of this vast universe of data has, to date, been left unexploited (Pedrinaci and Domingue, 2011).

Trust plays a hugely important role in the adoption of new technologies and even more so in a sphere with such vast amounts of publicly created data. Trust is paramount to the effective sharing and communication of tacit knowledge (Hislop, 2013). It is defined as the belief an entity has in the behaviour of others and the assumption that they will honour their obligations. Up to now, the area of trust in Linked Data has not been adequately addressed, despite the Semantic Web stack (see Figure 3) having included a trust layer from the very beginning (Artz and Gil, 2007).

There are many examples of Linked Data applications that users interact with on a daily basis without being aware of it. Google's Rich Snippets provides users with several lines of text that appear under every search result and is designed to give their users a sense for what is on the page and why it is relevant to their query. Many cultural heritage institutions, such as libraries and museums, draw additional data from external sources using Linked Data. Examples of this include geographical information or bibliographic information which embellishes the search experience for the user. In recent times, many public organisations have begun publishing Linked Data which has prompted a proliferation of mobile apps which harness this public information for the benefit of the public.

However, many of the datasets and links that exist on the Semantic Web are out of date and/or invalid which undermines the credibility and validity, and ultimately, the trustworthiness of both the dataset and the data provider (Rajabi *et al.*, 2012). Datasets should provide users with a means to assess the trustworthiness of the data within (Dai et al., 2008). This raises many questions on the provenance, reliability and believability of the data. Therefore, to answer these questions we need to assess trustworthiness of data.

2.3 The Semantic Web and Knowledge Management

Knowledge Management (KM) has been defined as "*the process of capturing, distributing, and effectively using knowledge*" (Davenport and Prusak, 2000). This definition is in agreement with that of Bhatt (2001) who defines KM as the process of knowledge creation, knowledge validation, knowledge formatting, distribution and knowledge application. These stages are depicted in Figure 2.1.

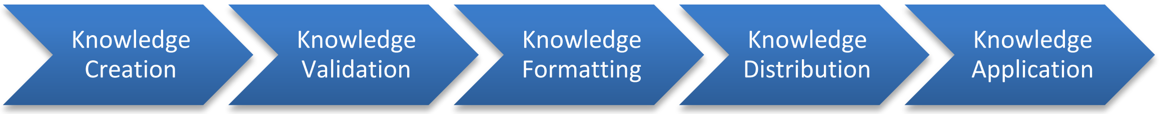


Figure 2.1 Knowledge Management process activities Bhatt (2001)

The process of turning data into knowledge is a complicated task. Data is considered to be basic statements or raw facts, information is when this information has been structured and knowledge is considered to be the understanding of this information. Nonaka and Tekenuchi (1997) discuss state that "*information is a flow of messages, while knowledge is created by that very flow of information anchored in the beliefs and commitment of its holder. This […] emphasizes that knowledge is essentially related to human action."* This concept is elaborated on in Nonaka's 'Spiral of Knowledge' (Figure 2.2).

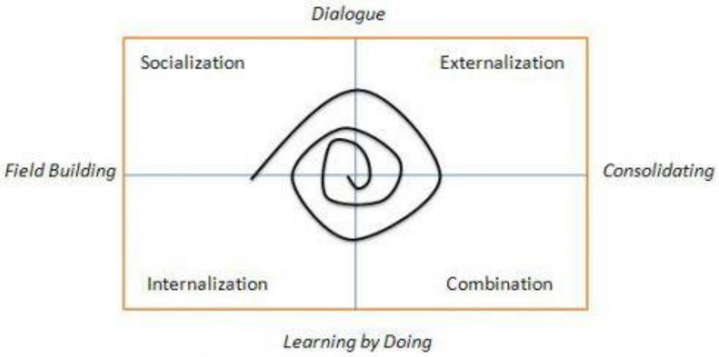


Figure 2.2 Spiral of Knowledge (SECI model) Nonaka and Tekenuchi (1997)

In Nonaka's spiral, tacit knowledge can be exchanged between individuals during interpersonal communications (socialization), and subsequently converted to explicit knowledge through the use of metaphors, analogies, diagrams etc. (externalisation). Explicit knowledge can be evaluated, analysed, enhanced and combined with other knowledge (combination) to simulate new insights and ideas, creating knowledge. Finally, explicit knowledge can be converted back into tacit knowledge (internalisation) through learning and experience. The process repeats and with each iteration, a deeper knowledge is created.

Therefore, data, prior to becoming information, is in a raw state and is not connected in a meaningful way to a context or situation. Knowledge is the result of understanding patterns in information and the ability to synthesize new information based on these patterns. As demonstrated in figure 2.3, when knowledge is accumulated over time, one can learn to understand patterns and principles in human action so that "*knowledge can be put in context, combined and applied appropriately*" (Bellinger *et al.*, 2006).

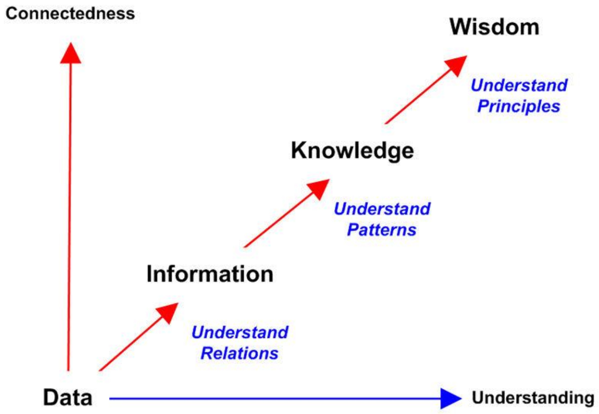


Figure 2.3 DIKW flow (Bellinger et al., 2006)

As a knowledge organisation becomes efficient in the task of processing data it can create more information. There can exist, however, an issue related to the perception or interpretation of this data. The perception of this information is a subjective process, reliant on the interpretation of the person, or machine, being presented with the data. The process of converting data into knowledge should be as swift as possible Bhatt (2001).

Technical documents and instructional material can enable the process of turning data into information, which in turn can become knowledge. The techniques chosen for both acquisition and representation of knowledge, together with the quality of their application can determine the degree to which a particular endeavour will succeed. Similarly, the techniques chosen for the representation of data on the Semantic Web will decide its ultimate success.

As previously identified, trust signifies a thorny issue on the Semantic Web landscape. It has been stated that "*trust is the single most important precondition for knowledge exchange*" (Rolland and Chauvel, 2000). A lack of trust was also recognised by Davenport and Prusak (2000) as a barrier to knowledge management.

With the personal interpretations of data and information contributing so much to the success or failure of a knowledge management endeavour it is imperative that technology does not remain the focus of our considerations. Bhatt advocates a *People-Process-Technology* model of knowledge management (figure 2.4). It is stated that placing too high an emphasis on the technological aspects is insufficient and that only by applying the focus to the interactions between people and process will knowledge management succeed (Bhatt, 2001).

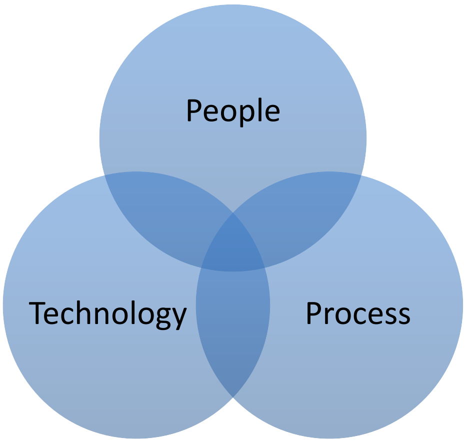


Figure 2.4: People, Process and Technology (Bhatt, 2001)

2.4 The Semantic Web and Linked Data Technologies

This section serves to outline a number of the significant technologies that underpin Linked Data and the Semantic Web. These technologies will be introduced with reference to the Semantic Web technology stack and then briefly described for the benefit of those unfamiliar to the concepts.

Part of Berners-Lee's original vision of the Web (2000) was that it should be used to publish, share and link *data*. The Semantic Web is not simply concerned with connecting datasets, but about linking information at the level of a single statement or fact.

In 2006, Berners-Lee published four principles for the linking of data:

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names
3. When someone looks up a URI, provide useful (RDF) information
4. Include RDF statements that link to other URIs so that they can discover related things

From this it can be seen that the technology that provides the foundation for much of the Semantic Web technology stack (Figure 2.5) is the *Uniform Resource Identifier* (URI). A URI is a string of characters used to uniquely identify a resource on the Web. They can be used to identify resources such as people, places and organisations, and then use web technologies to provide some meaningful and useful information when these URIs are looked up. This ‘useful information’ can then be returned in a various different encodings or formats. The most common standard for encoding this information on the Semantic Web is to use RDF (Resource Description Framework). RDF is a World Wide Web Consortium (W3C) standard that offers a very simple way of encoding data based upon making a series of statements about resources. These statements create a relationship between two objects by way of a property, or *predicate*. Formally, these statements take the form *subject-predicate-object* and are known as ‘*triples*’. Just as HTML provides a standard for linking documents on the web, RDF provides a standard way of linking data on the Semantic Web.

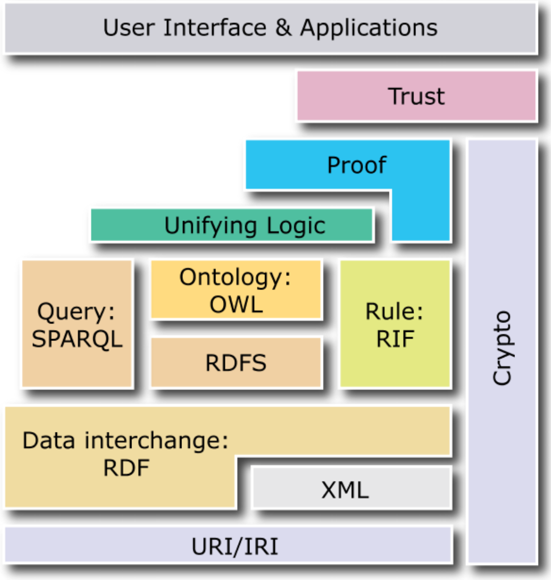


Figure 2.5: Semantic Web technology stack

The fundamental concepts of RDF are *resources, properties, statements* and *graphs*. The resource is the object at the centre of the description, i.e. what is being described. Every resource must be described with a URI. This URI does not need to be *dereferencable*, or accessible on the Web, but it is generally considered to be good practice (Antoniou and Van Harmelen, 2004). *Properties* describe relations between other resources, e.g. created by, is a, located in. A *statement* is the entity-attribute-value triple consisting of the resource, property and value. The value can either be another resource or a *literal* value. The example in Figure 2.6 uses a literal value but this could be replaced by another resource URI, e.g. that of Tim Berners-Lee's FOAF page. A *graph* is a set of RDF statements that have been grouped together, whereas a *named graph* is a set of RDF statements that have been provided an identifier.

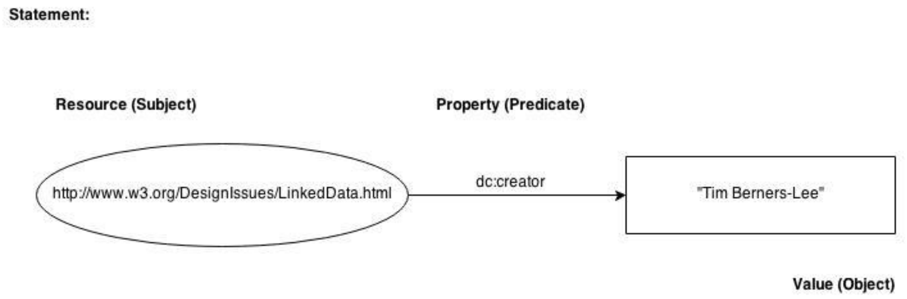


Figure 2.6 A RDF statement represented graphically (source: author)

The example from Figure 2.6 can be represented in RDF in the following manner:

```
<rdf:Description rdf:about="http://www.w3.org/DesignIssues/LinkedData.html">
  <dc:creator>Tim Berners-Lee</dc:creator>
</rdf:Description>
```

Other syntaxes, or *serializations*, of RDF, such as RDF/XML, Turtle, N3, N-Triples and JSON, are often preferred as they provide a more human-readable form of RDF (Decker *et al.*, 2000).

In order to allow for querying of this RDF data, where there will often be hundreds of thousands of RDF statements and files, it is necessary to store this data in a *triplestore*. A triplestore is a specialised database for the storage and retrieval of triples and queried via the SPARQL query language. The following represents a SPARQL query to DBpedia to find all landlocked countries with a population greater than ten million, return a list of countries, in the English language, and their respective population.

```
PREFIX type: <http://dbpedia.org/class/yago/>
PREFIX prop: <http://dbpedia.org/property/>

SELECT ?country name ?population
WHERE {
    ?country a type:LandlockedCountries ;
```

```
            rdfs:label ?country name ;
            prop:populationEstimate ?population .
    FILTER (?population > 10000000 && langMatches(lang(?country name), "en")) .
} ORDER BY DESC(?population)
```

In an RDF context, ontologies are the vocabularies and structures that embody the predicate (property) relations that enable data to be transformed into Linked Data graphs. An ontology is defined as "*a specification of a conceptualization*" (Gruber, 1993). Ontologies aim to make knowledge explicit by expressing concepts and their relationships. They define the common terms and concepts used to describe and represent an area of knowledge or collection of information about data and how the data is related (Wang *et al.*, 2004). Thus, ontologies provide a method for establishing a semantic structure and provide context to the data in question (Fensel, 2003).

Alongside the use of existing ontologies, the data provider should examine how entities in the dataset can be linked to entities in other datasets. This follows the fourth Linked Data principle presented by Berners-Lee, by linking to other URIs so that users can discover more. RDF links between entities in different datasets can be specified on two levels: the instance level and the schema level.

On the instance level links can be made between individual entities (e.g. people, places, objects) using the properties *owl:sameAs* and *rdfs:seeAlso*. The property *owl:sameAs* is used to denote that two URI references actually refer to the exact same entity. The *rdfs:seeAlso* property specifies that more relevant information can be obtained by following the link. The following contains an extract from the FOAF file of Tim Berners-Lee (Berners-Lee, 2011).

```
<rdf:Description rdf:about="http://www.w3.org/People/Berners-Lee/card#i">
  <owl:sameAs rdf:resource="http://identi.ca/user/45563"/>
  <foaf:knows rdf:resource="#dj"/>
</rdf:Description>
<foaf:Person rdf:about="#dj">
  <rdfs:seeAlso rdf:resource="http://www.grorg.org/dean/foaf.rdf"/>
  <foaf:mbox sha1sum>6de4ff27ef927b9ba21ccc88257e41a2d7e7d293</
    foaf:mbox sha1sum>
  <foaf:name>Dean Jackson</foaf:name>
</foaf:Person>
```

On the schema level, which contains the vocabulary used to classify the instance-level items, relationships can be conveyed using RDFS, OWL and the SKOS vocabulary.

The RDFS properties *rdfs:subPropertyOf* and *rdfs:subClassOf* can be used to declare relationships between two properties or two classes from different ontologies as shown below.

```
@prefix dbp: <http://dbpedia.org/ontology/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix mo: <http://purl.org/ontology/mo/>

<actedIn>  rdfs:subPropertyOf    dbp:starring- .
<hasChild> rdfs:subPropertyOf    dbp:parent- .
<isCitizenOf>         rdfs:subPropertyOf    dbp:nationality .
dbpedia-owl:RecordLabel rdfs:subClassOf mo:Label .
<http://rdf.freebase.com/ns/music.record label> rdfs:subClassOf mo:Label .
```

In the Semantic Web, ontologies are semi-structured and depict an open world, which means that an ontological model can grow with the data and does not need to contain every existing real-world entity from the outset. An ontology model can be merged with another ontology model thus they can be viewed as modular.

For a many years, the existence of metadata has been widely considered as a verification of accuracy and trustworthiness, as bad or incorrect metadata can lead to the resource being undiscoverable (Park, 2009). Commonly used metadata ontologies include DCMI and MODS. The focus of these standards has long been the classification by libraries of information resources to aid discoverability and therefore usage. However, these vocabularies have seen widespread usage across a broad range of fields.

The Dublin Core Metadata Initiative (DCMI) offers a core metadata vocabulary, commonly known as Dublin Core. The 15 elements of Dublin Core are broadly defined and contain no strict specifications regarding the range of values that an element can be assigned. In 2010, the Dublin Core vocabulary was further extended to 55 elements. This extension of the vocabulary is known as *terms* and bears the prefix *dcterms* or *dct*. The following is an example of a metadata record that demonstrates these vocabularies:

```
          ex:doc2 dct:title “What is Knowledge Management?” .
          ex:doc2 dct:creator ex:peter .
          ex:doc2 dct:created “2012-02-13” .
          ex:doc2 dct:publisher ex:dit .
          ex:doc2 dct:subject ex:knowledge .
          ex;doc2 dct:issued “2012-02-16” .
          ex:doc2 dct:replaces ex:doc1 .
          ex:doc2 dct:format “PDF” .
```

The example above demonstrates how DCMI includes two forms of metadata, description metadata and provenance metadata. The description metadata in the above example would include the *dct:title*, *dct:subject* and *dct:format*, whereas *dct:creator*, *dct:issued* and *dct:replaces* would be considered provenance metadata.

In April 2011, the W3C Provenance Working Group began developing a specification for the interoperable exchange of provenance information in heterogeneous environments such as the Web. In April 2013, the W3C Provenance Working Group published a family of specifications known as PROV. PROV consists of a number of specifications such as the PROV data model (PROV-DM) and the PROV ontology (W3C, 2013).

These metadata vocabularies are in fact, knowledge representation language. They allow the inference of additional information from the explicitly stated information. Such inferences give publishers of data the potential to create a basic degree of believability regarding the published data.

2.5 Conclusions

This chapter provided an overview of the Semantic Web and its relationship to Linked Data. Following this some of the key papers that relate to the LOD cloud were presented. Next the relationship between Knowledge Management and the Semantic Web were explored. Finally, some of the technology associated with the Semantic Web and Linked Datasets were discussed.

3. TRUST ON THE SEMANTIC WEB

3.1 Introduction

This chapter examines trust on the Semantic Web by exploring the existing research conducted in the area. The goal of this chapter is to explore some of the dimensions that can potentially be used for the experiment element of this project whose key focus is looking at how people determine which semantic web sources they have confidence in. Section 3.2 introduces the concept of trust as it specifically relates to online or web-based content. Section 3.3 discusses the topic of data quality and fitness for use, with an emphasis on the trust dimensions. Section 3.4 examines various dimensions to Semantic Web trust at both an objective and subjective level. Section 3.5 discusses the use of a trust assessment model for use in the Semantic Web.

3.2 What is Trust?

Trust has long been a research topic within the field of computer science. The definition applied is often specifically catered towards the research being conducted but in order to provide a broad understanding, a number of definitions of trust will be provided.

> "[Trust is] *the mutual confidence that one's vulnerability will not be exploited.*" (Barney and Hansen, 1994, p. 177)

> "[Trust is] *a subjective expectation an agent has about another's future behaviour based on the history of their encounters.*" (Mui *et al.*, 2002)

> "*Trust is the firm belief in the competence of an entity to act dependably, securely, and reliably within a specified context.*" (Grandison and Sloman, 2000)

While each of these definitions provides a sound description of the term, it seems that an aggregation of the three would be most appropriate when discussing trust in data sources. The initial definition should be considered the most basic requirement of a trust relationship. The additional definition elements of "subjective expectation" and "belief " map directly to the trust characteristics of reputation and believability.

Trust is an essential component of the initial Semantic Web vision, described by Berners-Lee (2000). Since the outset, the Semantic Web stack (fig. 3.1) has included a trust layer, responsible for representing the ontology, logic and proof layers below it.

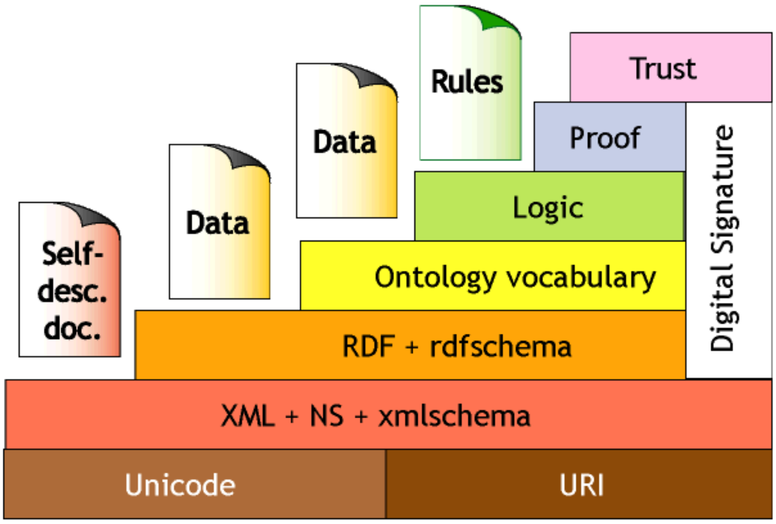


Figure 3.1 The Semantic Web Stack (Berners-Lee, 2000)

Often in technology circles, the notion of trust refers to the technology and tools in place to verify that the source of an information statement is actually who it claims to be. Commonly, encryption mechanisms and digital signatures allow for any individual to verify these sources of information (Stallings et al., 2008). Regardless of the existence of these tools, any information provider should be in a position to provide the consumer of that information with proof that certifies the origins of the data, rather than expect the consumer to generate these proofs themselves in what could be a computationally expensive process. The concept of the *Three A's*, that "*anyone can say anything about anything*" (w3.org, 2002) makes the web a unique source of information, but there is a requirement to understand where one is placing their trust.

As the Semantic Web develops and becomes more centred around agents and reasoning algorithms, trust plays a more prominent role. In the world of Linked Open Data, computer applications will be responsible for making quality and trust judgments on a range of diverse data sources, which contain data of varying degrees of quality. In everyday life, human web users make routine decisions about which data sources to rely on when presented with numerous sources as a response to a query. These sources can vary from blogs to academic institutions, governments to corporations, and objective reports to opinion-based editorial pieces. The decisions made by humans are often then based upon prior experience and knowledge of a source's perceived reputation. In many circumstances, such as in science and commerce, these decisions are formed based upon following a set of policies and procedures in respect to publicly available data and services.

These important trust judgments are currently in the hands of humans on the Semantic Web. This is not the vision of the Semantic Web as initially outlined by Berners-Lee (Berners-Lee, 2000). In the Semantic Web, humans will not be the singular consumer of information and data. Agents will need to be able to automatically make trust judgments to choose a service or information source while performing a task. Automatic reasoners will be expected to judge which of the diverse information sources available, often providing varying results and contradictions, are most acceptable as a response to a query (Hebeler et al., 2011).

3.3 Data Quality and Trust

As discussed previously, the development and formalization of Semantic Web technologies has led to an exceptional growth in the amount of data being published on the Web as Linked Open Data (LOD). Such increased volumes of information can certainly be considered as a step in the right direction. This deluge of information covers a staggeringly broad range of topics and domains, but unfortunately also reveals a large variation in data quality. However, it would not be prudent to discount datasets with quality issues as even data with some quality issues can be of use in certain applications, as long as the quality was within a required range.

This is in line with the typical view that data quality should be considered as its *"fitness for use"* (Wang and Strong, 1996). Any information under quality review should be subject to both an objective and subjective assessment (Pipino *et al.*, 2002). This is an important consideration as a thorough quality review is concerned with not only the objective properties of the data but also those characteristics perceived by the consumers of the data. This is of particular significance when dealing with a subjective property such as trustworthiness. Trust can be seen as one indicator of data quality. This view is held by Hartig who states that *"We understand trustworthiness of Semantic Web data as a criterion of information quality"* (Hartig, 2010).

Existing research on the subject has developed the notion of quality *dimensions* or criteria, which contain metrics and measures that are relevant to the consumer of the data when assessing data quality (Wang and Strong, 1996). These metrics are heuristics that are intended to fit a specific assessment situation (Pipino *et al.*, 2005).

There has been much research on the subject of data quality generally but, to date, little of this provides a singular focus on the topic of trust. Nonetheless, many of the studies up to now feature attributes that together can form a trust dimension even if they have not explicitly been identified as so.

The following sections investigate this topic further in detail by examining objective and subjective assessment metrics for measuring trust in Linked Data. While there are many papers available that discuss Linked Data quality, those papers that did not deal explicitly with the characteristics of trust were not considered for review.

3.4 Assessing the Trustworthiness of Data and Data Sources

As stated previously, trust can be seen as being an indicator of data quality. Thus, datasets perceived to be of high quality can hope to achieve high levels of trust.

Having identified trust as a characteristic of high data quality, it is worthwhile examining the attributes that contribute to the notion of trusted information.

Much of the early work in the domain of data quality remains relevant to the field of Linked Data and much of this early research forms the basis for current best practices. As introduced above, the notion of trust is neither objective nor subjective and that there are aspects of both that contribute to the ultimate decision on whether the data can be considered trustworthy.

Wang and Strong (1996) have classified data quality dimensions under the headings of *Intrinsic*, *Contextual*, *Representational* and *Accessibility*. Hartig and Zhao (2009) have categorized data quality dimensions into three categories; *Content-based*, *Context-based* and *Rating-based* dimensions. Zaveri (2012) elaborates on the categories created by Wang and Strong by adding a *Trust* category. This is divided into five trust dimensions (Figure 3.2); *Provenance*, *Verifiability, Reputation*, *Believability* and *Licensing*.

By taking Zaveri's five trust dimensions and using these as a template for the review of data quality literature, it is hoped that there can be some consensus achieved on the metrics that should be utilized in reviewing the trustworthiness of linked data.

Some of these dimensions cannot solely be assessed objectively or subjectively. In a number of cases, there will be a combination of metrics in place for examining trust qualities of the data under scrutiny.

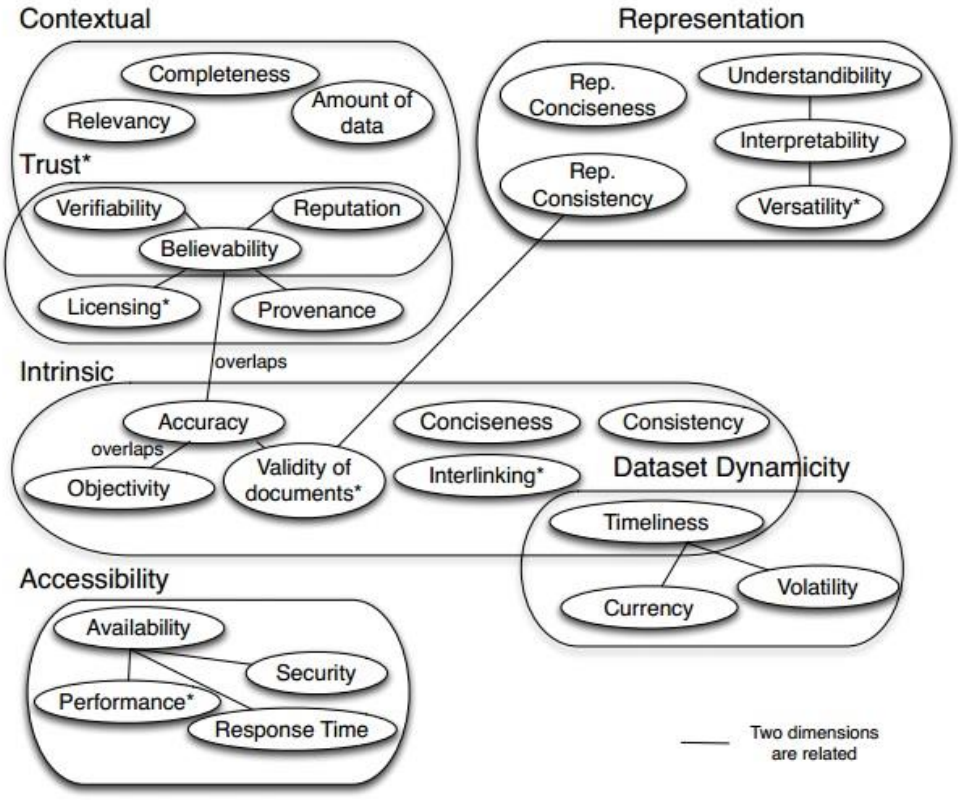


Figure 3.2 Linked Data quality dimensions (Zaveri et al., 2012)

3.4.1 Provenance

Studies show that one of the main factors that influence the trust of users in Web content is *Provenance* (Artz and Gil, 2007) and the literature broadly agrees on this metric. Provenance is a very specialized form of metadata that is defined as "*a record that describes entities and processes involved in producing and delivering or otherwise influencing that resource*" (W3, 2005). Thus, a common approach for data quality assessment is the analysis of provenance information. Tan concurs with this view, stating "*Information about provenance constitutes the proof of correctness [...] and [...] determines the quality and amount of trust [...]"* (Tan, 2007).

Provenance information about a data item is information about the history of the item, starting from its creation, including information about its origins. Tan (2007) distinguishes two granularities of provenance: *Workflow* (or *Coarse-grained*) provenance and *Data* (or *Fine-grained*) provenance. Flemming (2010) identified provenance as one of the primary considerations when assessing the quality of linked data and data sources. Golbeck (2006) also states that provenance tracking is useful when the trustworthiness of linked data is at issue. Although Wang and Strong (1996) list the importance of *Traceability* within their study it was subsequently excluded as one of the final metrics. Given the recent support (Zhao and Hartig, 2012) for this metric within the Semantic Web community it is clear that this should be a consideration.

There are a number of methods that can be utilized to assess the provenance of a data source or dataset. Flemming (2010) suggests the checking for the existence of basic provenance information, such as title, content and URI, within the dataset is one metric.

However, trust assessment becomes challenging when the consumers of this data are applications and machines. In order to automate the allocation of trustworthiness measures, it must be possible for trust values to be associated with different properties of the data such as the actual data content, the source of the data, how recently the data has been updated, the ontologies being used, and the creator, and for these, trust values be merged together to assess trust in the actual data. For example, there may be multiple *Friend of a Friend* (FOAF) files for an individual that describe their social profile in *Resource Description Framework* (RDF), but the one that is most trusted is the one available on their faculty website. This is because the trustworthiness of the source, their university, is higher than that of the other sources. Different trust levels may also be assigned to sources relative to their contents. For example, a laboratory may be trusted with information about a possible contagious infection but may not be trusted with respect to its financial predictions. Jacobi *et al.* (2011) suggests that the trust associated with any Web data is some combination of these different trust values associated with the content of the data as well as metadata about the data such as its source, creator, etc.

3.4.2 Verifiability

Verifiability is described as "*the degree by which a data consumer can assess the correctness of a dataset and as a consequence its trustworthiness*" (Zaveri *et al.*, 2012). This metric is closely linked with provenance and the term used synonymously by Flemming (2010). The usage of a dedicated provenance vocabulary is also considered to be measure of verifiability (Flemming and Hartig, 2010). It is listed by Zaveri (2012) under the heading of verifiability but clearly also related to provenance. While verifiability and provenance are linked, they remain individual dimensions. In many cases, such as a large heterogeneous dataset, the accuracy of the data may not be immediately verifiable without some statistical analysis. In cases such as this, the usage of a trusted methodology and ontology, not exclusive to provenance, can go a significant distance as a guarantee of quality (Bruce and Hillmann, 2004). This metric becomes important when a dataset contains information with a low believability or reputation.

Verifiability is a trust dimension that can be measured subjectively by a trusted, unbiased third party or objectively by the presence of a digital signature (Flemming, 2010). Bizer (2007) suggests a subjective assessment by verifying the correctness of the dataset. Flemming (2010) recommends verifying the publisher information subjectively. Wang and Strong (1996) also make reference to verifiability under the term *Traceability*, which was eliminated from the final criteria selected for use.

3.4.3 Reputation

The trust dimension with broadest agreement across the existing literature is reputation (see Figure 3.4). Reputation is defined as *"a judgment made by a user to determine the integrity of a source. It is mainly associated with a data published, a person, organization, group of people or community of practice rather than being a characteristic of a dataset. The data publisher should be identifiable for a certain (part of) a dataset"* (Zaveri *et al.*, 2012).

"*Reputation and trust on the semantic web have been gaining particular attention for their application to questions of provenance.* […] *provenance alone does not give any information about whether the specified source should be trusted*" (Golbeck and Hendler, 2004).

Wang and Strong (1996) uses reputation as a measure of data quality. Gil & Artz (2007) state that reputation of the publisher is formed by a subjective view held by a third party. This is determined either by the experience or recommendations of others (Artz and Gil, 2007). One method used to assess this metric is to conduct a survey, asking the community to rate the data provider. Generally a value of 0 (low) to 1 (high) is provided as a measure of reputation. Zaveri (2012) also suggests a less manual approach of assessing reputation using external links and page ranks.

Zaveri (2012) points to an interdependency existing between the data provider and the data source itself. Data is likely to be accepted as true if a trustworthy provider delivers it. On the other hand, the data provider is likely to be considered trustworthy if it provides true data. Thus, both the provider and the data can be measured for trustworthiness (Zaveri et al., 2012). This view is shared by Hartig and Zhao (2010).

Naumann (2002) defines reputation as "*the extent to which data are trusted or highly regarded in terms of their source*". It was cited that the reasons that data consumers choose one source over another is not always obvious. It has been noted that users tend to favour sources of information that are well known to them, rather than being the authoritative source of the most appropriate data (Naumann, 2002). Gamble and Goble (2011) share this opinion by stating that individuals are likely to select data from a source known to them or widely regarded as trustworthy, even if objective measures of accuracy reveal this trust to be misguided.

Flemming (2010) agrees with this but suggests using caution when utilizing this metric and ruled out reputation as an indicator of quality linked data. It was argued that reputational trust often stems from the prominence of a source, rather than an objective assessment of the source. An emerging authoritative provider of high quality data may not receive any consumer trust for these reasons, despite it perhaps having met all other criterion for high quality data.

Mendes *et al.*, (2012) agree with these common definitions. In that work, a subjective reputation score is assigned to data sets, e.g. data published in the English language is judged to have a higher reputation and the reputations of subsequent languages are rated accordingly (Mendes et al., 2014).

3.4.4 Believability (Accuracy)

Believability and accuracy also represent important trust dimensions that span the existing data quality literature. These two dimensions are closely related and although not entirely synonymous, they are commonly used interchangeably. Believability is defined as the measure "*to which the information is accepted to be correct, true, real and credible*" (Zaveri *et al.,* 2012). With this being a highly personal interpretation of accuracy, in many ways believability can be considered *perceived accuracy*.

Wang and Strong (1996) have identified believability as one of the main characteristics of high data quality. Bizer (2007) suggests that this can be objectively measured by checking the data provider is contained within a list of trusted providers.

Gamble and Goble (2011) also use believability as an intrinsic measure of trust, albeit a separate metric to accuracy. Naumann (2002) uses a metric of *reliability* to measure the likelihood of the data being correct. This is very different to his *accuracy* metric that objectively measures accuracy.

3.4.5 Licensing

Licensing is defined as a granting of explicit permission for a consumer to re-use a dataset under defined conditions (Open Data Institute, 2014). Applications that consume data from the Web must be able to access the exact conditions under which data can be reused and republished. The availability of suitable frameworks for publishing such requirements is an essential requirement to inspiring data providers to participate in the Web of Data, and in assuring data consumers that they are not infringing the rights of others by using data in a certain way (Bizer *et al.*, 2009).

Fleming and Hartig (2010) are strong advocates of this dimension of trusted data and suggests five licensing conditions. Machine-readable and human-readable indications of a license, permission to use the dataset, attribution, and a CopyLeft or ShareAlike license if appropriate.

As detailed in Figure 3.3, publishing under an open license is the first criteria in Tim Berners-Lee's *5 Star Open Data* model for Linked Open Data (Berners-Lee, 2009). Hogan et al. (2012) demonstrate how licensing should be applied to linked data resources in RDF. Publishing data using an open license is also the fourth *shamrock* of Cyganiak's *5 Shamrock* model for publishing open data (Cyganiak, 2011).

Star/Shamrock	**Berners-Lee (2005)**	**Cyganiak (2011)**
*	**Open license**	Publish data on the web
**	Structured data	Machine-readable
***	Non-proprietary formats	Open standards
****	Use URIs	**Open license**
*****	Link data to other data	List data in data catalogue

Figure 3.3 Comparison of 5 Star and 5 Shamrock publishing models (author)

3.4.6 Summary of Analysis

The following table, Figure 3.4, summarises the findings of the literature review and outlines the key features, researchers and metrics for each of the five characteristics of trust in Linked Data.

	Key Features	**Key Researchers**	**Metrics**
Provenance	A record of origin Describes entities and processes influencing the resource Proof of correctness Often dictates the quality and amount of trust associated with a resource Can be objectively assessed	(Golbeck and Mannes, 2006) (Artz and Gil, 2007) (Tan, 2007) (Flemming, 2010) (Flemming and Hartig, 2010) (Hartig and Zhao, 2010) (Zaveri et al., 2012)	Verify VoID description exists and is correctly located Check for basic provenance metadata (title, creator, content, URI)
Verifiability	Enables assessment of correctness Linked with the notion of provenance	(Wang and Strong, 1996) (Bizer, 2007)	Check for the existence and usage of dedicated provenance vocabularies Check for the existence of digital

	Can be objectively and/or subjectively assessed	(Flemming, 2010) (Flemming and Hartig, 2010) (Zaveri et al., 2012)	signatures and verifying their authenticity Survey a community to rate a dataset’s verifiability
Reputation	A judgment made by a user to determine integrity Associated with data, individuals, organisations, groups and communities of practice Broad agreement on this metric throughout research Can be subjectively assessed	(Wang and Strong, 1996) (Naumann, 2002) (Artz and Gil, 2007) (Flemming, 2010) (Hartig and Zhao, 2010) (Gamble and Goble, 2011)	Survey a community to rate a data provider’s reputation

		(Mendes et al., 2012) (Zaveri et al., 2012)	
Believability	The degree to which information is accepted to be correct and true “Perceived accuracy” Intrinsic measure of trust Can be assessed subjectively	(Wang and Strong, 1996) (Naumann, 2002) (Bizer, 2007) (Gamble and Goble, 2011) (Zaveri et al., 2012)	Survey a community to rate the believability of a dataset and data provider
Licensing	Granting of permission to use a dataset Provides the legal terms of its use Legal requirements for attribution and replication of data	(Berners-Lee, 2009) (Bizer et al., 2009) (Flemming and Hartig, 2010)	Verify the existence of a machine-readable license Verify the existence of a human-readable license

	Can be assessed objectively	(Cyganiak, 2011) (Hogan et al., 2012) (Zaveri et al., 2012)	Verify the existence of permission information Verify the existence of attribution requirements Verify the existence of a CopyLeft or ShareAlike condition statement

Figure 3.4 Summary of literature review (Source: author)

3.5 A Trust Assessment Model for Linked Data

By taking these dimensions as a means for assessing trust within the context of Linked Data an assessment model can be applied to a dataset. The following (Figure 3.5) is a model developed as part of this research that endeavours to characterize the relationships and dependencies that exist between the trust criteria outlined previously. In this model, provenance is regarded as the root of trusted data. Data with provenance metadata can then be assessed on reputation or believability. Should any of the dimensions of provenance, reputation or believability be under question, the data can be assessed under the metrics associated with the dimension of verifiability. Following these assessments, all data is required to meet the metrics specified within the license dimension.

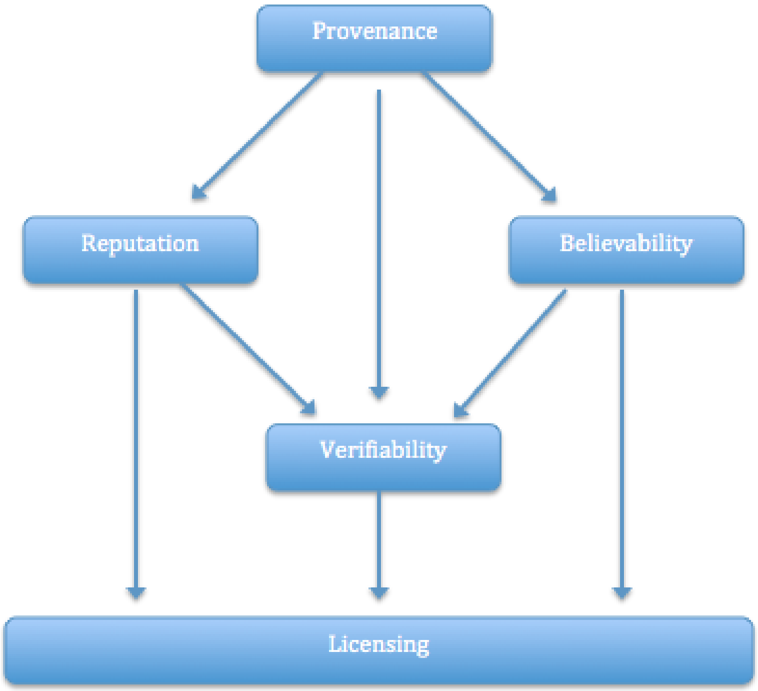


Figure 3.5 Trust Assessment Model for Linked Data (author)

3.6 Conclusions

The typical view is that data quality should be considered as its *"fitness for use"*. Any information under quality review should be subject to both an objective and subjective assessment (Pipino *et al.*, 2002). This is an important consideration as a thorough quality review is concerned with not only the objective properties of the data but also those characteristics perceived by the consumers of the data. This is of particular significance when dealing with a subjective property such as trustworthiness. Trust can be seen as one indicator of data quality.

This chapter examined the existing literature in relation to trust of linked data. First a background to the notion of trust and how it applies to the field of linked data was discussed. Following this, the topic of Data Quality and how trust can be identified as one factor of data quality was examined. Next the assessment of trust in linked data was investigated and how data should be assessed at an objective and subjective level was examined, as well as individual trust dimensions, which can be used towards this assessment. Finally a trust assessment model that takes the dimensions identified and formalizes a method for assessment of linked datasets was outlined.

Using the knowledge ascertained from performing this literature review, the following chapter outlines the design of the experiment to assess the trustworthiness of the selected linked datasets, with the intention of developing a set of guidelines that can be used in the creation and assessment of trustworthy linked data.

4. ASSESSING LINKED DATA

4.1. Introduction

This chapter explores the datasets that this research will use. It begins with Section 4.2, a reminder of the architecture of the experiment; following this, Section 4.3 provides an overview of linked datasets in general. Section 4.4 lists a series of criteria as to what represents a quality dataset, highlighting the importance of characteristics such as *Currency*, *Size* and *Internationality*. Section 4.5 follows this with a list of potential datasets and they are evaluated with respect to the criteria outlined in the previous section, until the best-fit linked datasets are identified. Each of these datasets is described in detail, and finally in Section 4.6 the five quality criteria (*Provenance*, *Licensing*, *Reputation*, *Believability*, and *Verifiability*) are explored as either subjective, objective or both.

4.2. Overview of Experiment

The experiment focuses on the assessment of Linked Open Data (LOD) with the intention of determining the key characteristics of trusted linked data. Once identified, these features can be elaborated into a set of policies and procedures to aid in the creation and assessment of high quality trusted linked data. The literature review from the previous chapters demonstrated a number of trust dimensions within the field of data quality. By examining these dimensions, a series of metrics can be created with which to assess linked datasets for trustworthiness.

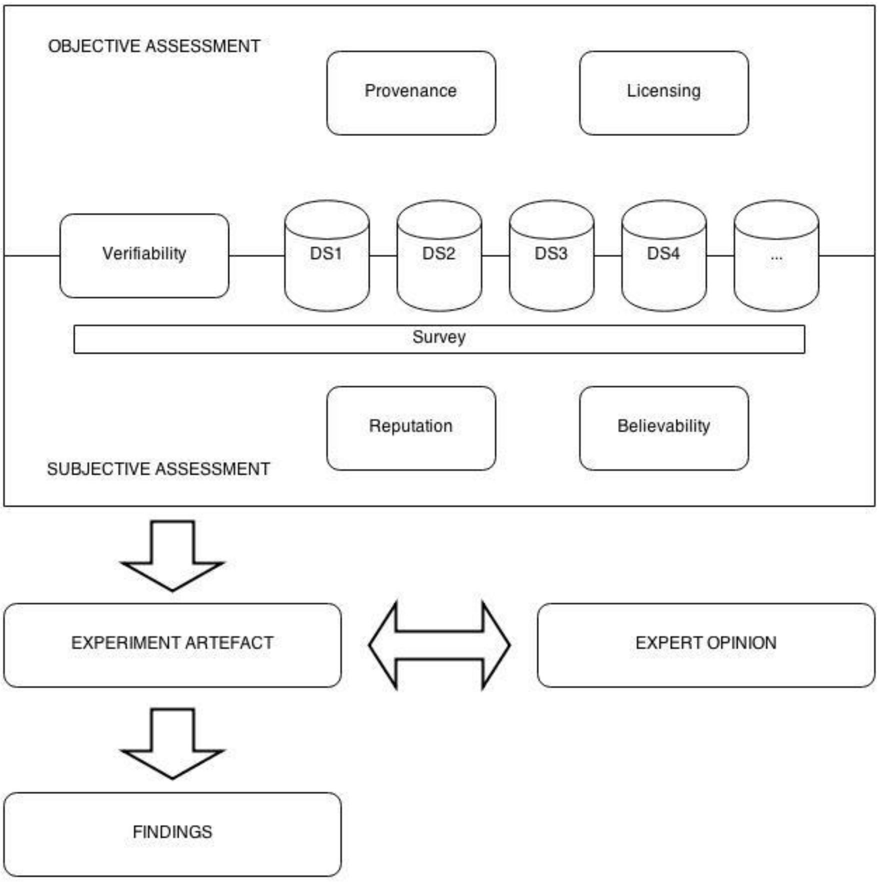


Figure 4.1 Graphical Representation of Experiment (author)

Figure 4.1 provides a graphical representation of the experiment. As outlined in the previous chapter, it is recommended that data is evaluated using both objective and subjective measures (Wang and Strong, 1996). The dimensions of provenance and licensing have been identified as demanding objective analysis, due to the requirement that they be assessed for the existence of specific attributes. The characteristics of reputation and believability will be examined subjectively as their assessment is based entirely on the subjective opinion of those interacting with the data. The final dimension of verifiability will be assessed both objectively and subjectively as recommended in the previous chapter. This is due to a requirement to objectively verify the usage of dedicated provenance ontologies but also to gain the subjective opinion from a community on the verifiability of a dataset.

4.3. Linked Datasets Background

From an examination of the Linked Data Cloud there are approximately 295 datasets available for investigation (Bizer *et al.*, 2011). Figure 4.2 provides a recent view of the Linked Open Data Cloud and Table 4.1 outlines the breakdown of these datasets by domain. It can be clearly seen that government data accounts for the largest proportion of available Linked Data while user-generated, or crowd-sourced, data accounts for the least.

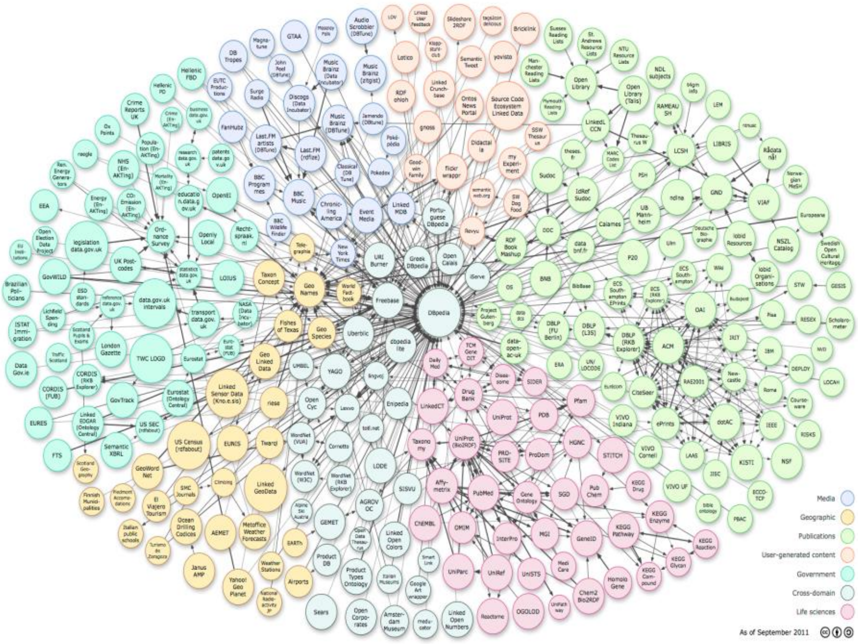


Figure 4.2 LOD Cloud image (Bizer *et al.*, 2011)

Domain	No. of Datasets	Triples	%
Media	25	1,841,852,061	5.82 %
Geographic	31	6,145,532,484	19.43 %
Government	49	13,315,009,400	42.09 %
Publications	87	2,950,720,693	9.33 %
Cross-Domain	41	4,184,635,715	13.23 %
Life Sciences	41	3,036,336,004	9.60 %
User-Generated Content	20	134,127,413	0.42 %
Totals	295	31,634,213,770	99.92 %

Table 4.1 Chart of Dataset Breakdown (Bizer *et al.*, 2011)

4.4. Database Selection Rationale

The aim of this process was to identify a number of datasets that could be used in an objective and subjective assessment of trust in linked data. There were a number of criteria utilized in the selection process that focused on demonstrating and representing the broad range of data available on the linked data cloud. It was also important to mitigate against imbalances and bias when selecting data sources. Together with the criteria described below it was also necessary that these data sources adhere to the *'5 Stars of Linked Data'* as outlined by Berners-Lee (2009).

The following criteria have been identified by the author as a means of selecting, and in some cases de-selecting, datasets for examination.

CONSIDERATION	DESCRIPTION
Currency	In order to gain a clear understanding of the Linked Data landscape as it currently stands, it is important to use data that is up-to-date. In deciding the datasets to use, datasets that had a publication date prior to 2012 were eliminated from consideration. To demonstrate the subjective nature of reputation, it is imperative that a new, largely unknown database is examined also.
Technology Agnostic	The Linked Data Cloud features a plethora of technology standards and applications. It was decided that the dataset selection process should be technology agnostic meaning that the standards and technologies used to present the data would not have a bearing on the process. By doing so it is expected that a more representative view of the Linked Data landscape can be achieved.
Data Provider	The Linked Data Cloud features a broad range of data providers across a broad range of industries. It is hoped that by selecting data sources from a wide range of institutions there can be balance and any inherent bias eliminated. To allow for a

	representative sample, government data, scientific research data, user-generated data and cultural heritage data will be chosen
Size	The size of the dataset does not go any distance to infer its utility to the Linked Data Cloud. For this reason, datasets will not be chosen based on the size and number of triples within. Datasets, both large and small, will be considered for selection.
Internationality	In order to provide a fair representation of the Linked Data landscape as it currently stands, data was not selected based on geographic location of the data provider. Where language was a consideration, only data sources provided in English were considered. The datasets selected will not originate from solely one country and endeavour to represent the international, borderless, nature of the Web.
Subjective Perception	Some datasets are generally perceived to be more trustworthy than others. Datasets developed through crowdsourcing information from the general public can be as significant and accurate as data curated by governments or academic institutions (Casebourne *et al.*, 2012). For this reason, a crowd-sourced dataset must be chosen for assessment.

4.5. Datasets Selected

The following datasets were considered as candidates for examination as part of the experiment.

- **LinkedGeoData**

  LinkedGeoData uses the information collected by the OpenStreetMap project and makes it available as an RDF knowledge base according to the Linked Data principles.

- **OCLC WorldCat**

  OCLC WorldCat is a downloadable dataset of the 1.2 million most widely held works in WorldCat.

- **ChEMBL**

  ChEMBL is freely available data from life science experiments covering the full spectrum of molecular biology.

- **Linked Logainm**

  Linked Logainm is an online database containing Irish geographical names generated by the Placenames Branch of the Department of Arts, Heritage and the Gaeltacht, developed in collaboration with Fiontar, DCU.

- **UCD Data Hub**

  The UCD Data Hub is a repository of digitised cultural heritage data and research data made available in many formats, including Linked Data serializations.

- **education.data.gov.uk**

  education.data.gov.uk contains a snapshot of Edubase taken in 2009 and published as linked data.

- **DBpedia**

  DBpedia is a crowd-sourced community effort to extract structured information from Wikipedia and make this information available on the Web.

- **Geonames**

  The GeoNames geographical database covers all countries and contains over eight million placenames that are available for download free of charge.

- **Musicbrainz**

  MusicBrainz is an open music encyclopaedia that collects music metadata and makes it available to the public.

- **International Monetary Fund (IMF)**

  This dataset contains statistical observations from a number of studies published by the International Monetary Fund (IMF).

A number of datasets were not chosen due to their similarity with other datasets. For example, DBpedia was considered to be worthy of analysis due to it’s crowd-sourced origins, therefore Musicbrainz was deselected as the resource bore too many similarities and covered a more narrow field of data. For the same reasons LinkedGeoData was selected above Geonames. A number of data sources were disqualified from selection due to technical or administrative considerations. The Linked Logainm dataset proved inaccessible and unreliable on a number of occasions and attempts to download the data dumps were also unsuccessful. UCD Data Hub was identified for assessment but was still undergoing rapid development and so was eliminated from the study.

The following datasets were selected for assessment:

- OCLC WorldCat
- DBpedia
- International Monetary Fund (IMF)
- Education.data.gov.uk
- LinkedGeoData

The following section details the datasets chosen for examination. It provides and explanation for this decision and additional details that contribute to a broader understanding of the dataset.

4.5.1. The OCLC WorldCat Dataset

OCLC WorldCat is a downloadable dataset of the 1.2 million most widely held works in the WorldCat catalogue and was published in 2012.

OCLC WorldCat	
Why is this dataset suitable?	The OCLC WorldCat dataset represents the federation of many library collections from around the world. It is a large, heavily curated dataset from a data provider with experience in library metadata and Linked Data. WorldCat has been selected for examination due to its size, internationality and it represents a data provider with world-leading expertise in metadata and data curation.
Type of data (e.g. federated, descriptive, longitudinal)	The data is descriptive metadata related to library collections, authors, published works and publishers. The data is federated from member libraries and wide variety of partners in order to leverage collective data from the world's libraries in ways that benefit scholarship, research, business and civic life. The dataset represents the 1.2 million of the most widely held works in WorldCat.
Location	This dataset is made available for download from the following website: http://www.oclc.org/data/data-sets-services.en.html
Size	69,760,417 triples 8GB download
Format	The data is presented for downloading in a 8GB .nt file dump. There are no publicly accessible SPARQL endpoints or mirrors available for this data at present.

4.5.2. The DBpedia Dataset

DBpedia is a crowd-sourced community effort to extract structured information from Wikipedia and make this information available on the Web. The most recent release of this data was published in December 2013.

DBpedia	
Why is this dataset suitable?	DBpedia has been selected for examination due to its large size, internationality and it representing a well-renowned crowd-sourced dataset. This is a dataset that is widely used throughout the Linked Data field due to the number of links it can provide to a broad range of resources across the web. The crowd-sourced nature of this dataset allows for the perception of untrustworthiness, thus making it an important dataset to examine.
Type of data (e.g. federated, descriptive, longitudinal)	DBpedia.org is a community-driven effort to extract structured information from Wikipedia and to make this information available on the Web. DBpedia favourably compares to traditional encyclopaedias and contains descriptive metadata from all aspects of the known-world (Casebourne *et al.*, 2012).
Location	The most current release of this dataset is made available for download from the following website: http://wiki.dbpedia.org/Downloads39
Size	825,761,509 triples 45GB download
Format	The data is presented for downloading in a range of file formats. The DBpedia datasets are available to download individually and in 119 different languages.

4.5.3. The International Monetary Fund (IMF) Dataset

This dataset contains statistical observations from a number of studies published by the International Monetary Fund (IMF). The most recent release of this data originates from 2013.

International Monetary Fund (IMF)	
Why is this dataset suitable?	The International Monetary Fund (IMF) dataset represents the federation of many statistical observations and analysis from around the world. It is a moderately sized, public data source with an international focus and is of interest globally. It has been selected for examination due to its currency, size, internationality and it represents a data provider with world-leading expertise in statistical data curation.
Type of data (e.g. federated, descriptive, longitudinal)	This is statistical information made available by the IMF through a REST API accessible to the general public. The IMF data available for consumption as Linked Data has been scraped from the IMF REST API and transformed to Linked Data as outlined in Capadisli *et al.* (2013).
Location	The most current release of this dataset is made available for download from the following location: http://imf.270a.info/data/data.tar.gz
Size	40,036,129 triples 58mb download
Format	The data is presented for downloading as one tar.gz file containing 104 RDF files.

4.5.4. The LinkedGeoData Dataset

LinkedGeoData uses the information collected by the OpenStreetMap project and makes it available as an RDF knowledge base according to the Linked Data principles. The most current release of this data is from August 2013.

LinkedGeoData	
Why is this dataset suitable?	LinkedGeoData has been selected for examination due to its currency, large size, internationality and it representing another well-known crowd-sourced dataset. This is a dataset that is widely used throughout the Linked Data field and links to other crowd-sourced datasets, such as DBpedia and Geonames. The crowd-sourced nature of this dataset allows for the perception of untrustworthiness, thus making it an important dataset to examine.
Type of data (e.g. federated, descriptive, longitudinal)	LinkedGeoData is an effort to add a spatial dimension to the Semantic Web. LinkedGeoData uses the information collected by the OpenStreetMap project and makes it available as an RDF knowledge base according to the Linked Data principles.
Location	The most current release of this dataset is made available for download from the following location: http://linkedgeodata.org/Datasets
Size	226,403,937 triples 121GB download
Format	The data is presented for downloading in a range of formats, including a dump of the entire dataset in one (.nt) 121GB file.

4.5.5. The UK Government Education Dataset

education.data.gov.uk contains a snapshot of Edubase taken in 2009 and published as Linked Data in 2012.

UK Government Education data	
Why is this dataset suitable?	Data.gov.uk is a federation of many statistical observations and analysis from numerous government and public sector institutions throughout the United Kingdom. Education.data.gov.uk represents a moderately sized, public data source. This data source has been selected for examination due to its currency, size, internationality (non-Irish government data) and it characterizes a data provider with world-leading expertise in statistical data curation.
Type of data (e.g. federated, descriptive, longitudinal)	The UK government has released public data to help taxpayers understand how government works and how policies are made. There are over 9,000 datasets available, from all central government departments and a number of other public sector bodies and local authorities.
Location	The most current release of this dataset is made available for download from the following location: http://education.data.gov.uk
Size	6,630,934 triples File size unknown
Format	The data is available for access via a public-facing REST API and can be downloaded in a broad range of formats (CSV, HTML, JSON, RDF, TTL, Text and XML).

4.6. Assessment of Data

As outlined previously, there is a need to assess the datasets from both an objective and subjective perspective (Wang and Strong, 1996). The dimensions of provenance and licensing have been identified as demanding objective analysis, due to the requirement that they be assessed for the existence of specific attributes. The characteristics of reputation and believability will be examined subjectively as their assessment is based entirely on the subjective opinion of those interacting with the data. The final dimension of verifiability will be assessed both objectively and subjectively as recommended in the previous chapter. This is due to a requirement to objectively verify the usage of dedicated provenance ontologies but also to gain the subjective opinion from a community on the verifiability of a dataset.

4.6.1. Objective Assessment of Data

Provenance

In order to allow applications to be certain about the origin of data, as well as to enable them to assess the quality of data, data sources should publish provenance metadata together with the principal data. A widely deployed vocabulary for representing provenance information is Dublin Core (dc:creator, dc:publisher, dc:date). Alternative vocabularies that provide means for representing the data creation process in more detail include the W3C PROV-O vocabulary and the more specialized W3C PAV (Provenance, Authoring and Versioning) vocabulary.

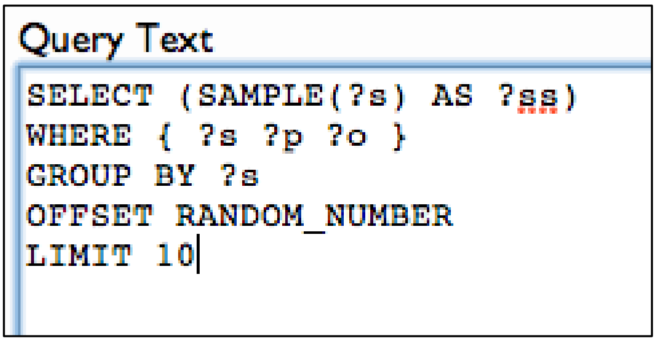
In addition to making individual object and resource data self-descriptive, it is also helpful that data publishers provide metadata that describes the characteristic of the entire dataset, for instance the topic of a dataset and more detailed information about the dataset. A vocabulary for representing such metadata is the VoID vocabulary.

There are a number of methods that can be utilized to objectively assess the provenance of a data source or dataset. Flemming (2010) suggests inspecting the dataset for the existence of basic provenance information, such as title, content and URI, within the dataset is one metric to assess provenance.

The existence of a VoID description file is also a metric that can be utilized. VoID is an RDF Schema vocabulary for expressing metadata about RDF datasets (Keith Alexander et al., 2011). The VoID file expresses access metadata, structural metadata, and links between datasets and for this reason is a highly useful resource when unfamiliar with the dataset. While it is considered best practice that every dataset should publish a VoID description (Berners-Lee, 2009), less than 30% of datasets on the LOD Cloud do so (Cyganiak, 2012).

The RFC 5758 (Dang, 2010) defines a mechanism for reserving 'well-known' URIs on any Web server. The URI /.well-known/void on any Web server is registered by this specification for a VoID description of any datasets hosted on that server. For example, on the host www.example.com, this URI would be *http://www.example.com/.well-known/void*. The VoID file accessible via the well-known URI should contain descriptions of all RDF datasets hosted on the server. This includes any datasets that have resolvable URIs, a SPARQL endpoint, a data dump, or any other access mechanism that maintains a URI on the server's hostname.

By examining randomly returned RDF records for provenance it is expected to get that it is possible to get a broad view of the data providers implementation of provenance standards, if any. This will be ascertained by running a SPARQL query on each resource to return a number of random resources for examination. By viewing the returned RDF for each of the subjects and objects it is possible to assess the contents with regard to provenance.



```
SELECT (SAMPLE(?s) AS ?ss)
WHERE { ?s ?p ?o }
GROUP BY ?s
OFFSET RANDOM_NUMBER
LIMIT 10
```

Figure 4.4 Sample SPARQL query to return 10 random subjects (Source: Author)

A SPARQL query such as that shown in Figure 4.4 will be used. This returns the subject, predicate and object of a number of triples, offset by a randomly generated number. It is expected that this will give an appropriate snapshot of the dataset.

Licensing

Web data should be self-descriptive concerning any restrictions that apply to its usage. A common way to express such restrictions is to attach a data license to published data. Doing so is essential to enable applications to use Web data on a secure legal basis. A common means to attach licenses to Linked Data is to use dc:rights links pointing at the license as document-level metadata.

Fleming and Hartig (2010) are strong advocates of this dimension of trusted data and suggest using five licensing conditions. Machine-readable and human-readable indications of a license, permission to use the dataset, attribution, and a CopyLeft or ShareAlike license if appropriate. A machine-readable license will be present within the metadata (e.g. cc:license or dc:license) of the resource whereas a human-readable license may be present on the main website of the resource. ShareAlike is a copyright licensing term used to describe works or licenses that require copies or adaptations of the work to be released under the same or similar license as the original. CopyLeft licenses are free content or free software licenses with a ShareAlike condition.

4.6.2. Subjective Assessment of Data

Reputation

It was argued by Flemming (2010) and Naumann (2002) that reputational trust often stems from the prominence of a source, rather than an objective assessment of the source. An emerging authoritative provider of high quality data may not receive any consumer trust for these reasons, despite it perhaps having met all other criteria for high quality data. Mendes *et al.*, (2012) suggest an approach that subjectively measures the reputation of a dataset.

In this experiment reputation will be examined by surveying a group of experts and parties with an active interest and involvement in the Linked Data field. The individuals will be prompted to provide their opinions on the reputation of a number of datasets, including but not limited to the datasets selected for examination. A similar approach will be taken in regard to the believability and verifiability of the data source.

Believability

In many instances, the terms *believability* and *accuracy* are used interchangeably. Gamble and Goble (2011) use believability as an intrinsic measure of trust, albeit a separate metric to accuracy. Naumann (2002) uses a metric of *reliability* to measure the likelihood of the data being correct. This is very different to his *accuracy* metric that objectively measures the correctness of the data. In this experiment the term *perceived accuracy* is used in reference to believability.

Bizer (2007) suggests that believability could be objectively measured by checking the data provider is contained within a list of trusted providers however an up-to-date register of this nature is not actively maintained.

In this experiment believability will be examined by surveying the same group of experts and parties within the Linked Data field. The individuals will be prompted to provide their opinions on their perception of the accuracy of a number of datasets, including but not limited to the datasets selected for examination.

4.6.3. Subjective and Objective Assessment of Data

Verifiability is described as "*the degree by which a data consumer can assess the correctness of a dataset and as a consequence its trustworthiness*" (Zaveri *et al.*, 2012). Thus it is a trust dimension that can at once be viewed subjectively and objectively. For this reason, the author will conduct two separate reviews of verifiability within the experiment.

Verifiability: Subjectively

Verifiability is a trust dimension that can be measured by subjectively examining the accuracy of the dataset by a trusted, impartial third party (Bizer, 2007). It is suggested that this subjective assessment, verifying the correctness of the dataset is beneficial. The experiment will survey a community of Linked Data experts and prompt them for their opinions on the verifiability of a number of datasets.

Verifiability: Objectively

Verifiability is described as “the degree by which a data consumer can assess the correctness of a dataset and as a consequence its trustworthiness” (Zaveri *et al.,* 2012). Verifiability is a trust dimension that can be measured objectively by examining for the presence of a digital signatures within the RDF of the dataset (Flemming, 2010). RDF with digital signatures is fundamental to building the "Web of Trust" for trusted linked data applications.

The usage of a provenance vocabulary is also considered to be a metric that leads to a measure of verifiability (Flemming and Hartig, 2010). The experiment will analyse data randomly for the usage of prominent provenance ontologies.

4.7. Conclusions

This chapter explored the datasets being used in this research. It began with a reminder of the architecture of the experiment. Next an overview of linked datasets in general was provided. Following this a series of criteria as to what represents a quality dataset was discussed, highlighting the importance of characteristics such as *Currency, Size* and *Internationality*. The next section listed potential datasets that were evaluated with respect to the criteria outlined in the previous section, and the best-fit linked datasets were identified: The OCLC WorldCat Dataset, The DBpedia Dataset, The International Monetary Fund (IMF) Dataset, The LinkedGeoData Dataset, The UK

Government Education Dataset. Each of those datasets was described in detail, and the final section described the five quality criteria (*Provenance*, *Licensing*, *Reputation*, *Believability*, and *Verifiability*) as either subjective, objective or both.

5. TECHNOLOGY DEPLOYMENT

5.1 Introduction

This chapter discusses the use of the SPARQL query language to interrogate the selected Linked Datasets. Section 5.2 provides an overview of the deployment of the Virtuoso SPARQL query service (which implements the SPARQL Protocol for RDF data). This has two main sub-sections, first looking at installing the Virtuoso SPARQL query service, and second loading the selected Linked Datasets into Virtuoso. Following this, Section 5.4 will discuss some of the technical limitations of the hardware used in this experiment and what issues that may cause.

5.2 Deploying the System

As discussed in Chapter 2, Linked Data datasets are interrogated using the SPARQL query language. This is similar to SQL querying of a relational database. A large number of data providers allow for their data to be queried directly by users by presenting a SPARQL endpoint to the public. Although not mandatory, implementing a SPARQL endpoint can result in your data becoming more accessible and therefore used by a greater number of individuals and computers. This has the added benefit of increasing exposure and thus *Reputation* and potentially, *Believability*. While many significant data providers do publish their data in this way, a large number do not. During this experiment, the data providers selected for examination who publish data dumps of their resource will have their data loaded as a graph in a local SPARQL endpoint.

5.2.1 Installing Virtuoso OpenSource (Vos)

Virtuoso was selected as the triplestore for this project as it was available in an open-source package. It was also widely deployed as the triplestore of the majority of datasets considered for evaluation (i.e. DBpedia, LinkedGeoData, Linked Logainm). A number of alternative SPARQL endpoints have a limit of 1 billion triples, whereas Virtuoso can process this volume with issue. Based on this decision, the operating system selected for the experiment was Ubuntu Server 12.04 LTS. This was installed on a HP Workstation with the following specification:

HP Z600 Workstation technical specifications	
CPU	4x Intel(R) Xeon(R) CPU E5530 @ 2.40GHz
Memory	4030MB (2795MB used)
Hard Disk	320GB ATA WDC WD3200AAJS-6
Operating System	Ubuntu 12.04.3 LTS

Once the operating system was installed it was a matter of installing the Virtuoso OpenSource server application. At the command line, enter the following commands to update the application repositories to access the latest versions of Ubuntu packages:

```
sudo apt-get update
```

Next, search the Ubuntu application repositories for all Virtuoso packages:

```
sudo apt-cache search ‘^virtuoso’
```

This results in the following output, listing all available Virtuoso packages:

```
virtuoso-nepomuk - OpenLink Virtuoso Open-Source Edition (OSE)

virtuoso-minimal - Virtuoso minimal Server (metapackage for latest version)

virtuoso-opensource - OpenLink Virtuoso Open-Source Edition (OSE)

virtuoso-opensource-6.1 - OpenLink Virtuoso Open-Source Edition - Server support files

virtuoso-opensource-6.1-bin - OpenLink Virtuoso Open-Source Edition - Server Binaries

virtuoso-opensource-6.1-common - OpenLink Virtuoso Open-Source Edition - Common Binaries

virtuoso-server - Virtuoso OSE Server (metapackage for latest version)
```

```
virtuoso-vad-bpel - OpenLink Virtuoso Open-Source Edition - BPEL

virtuoso-vad-conductor - OpenLink Virtuoso Open-Source Edition - Conductor

virtuoso-vad-demo - OpenLink Virtuoso Open-Source Edition - Demo

virtuoso-vad-doc - OpenLink Virtuoso Open-Source Edition - Documentation

virtuoso-vad-isparql - OpenLink Virtuoso Open-Source Edition - iSPARQL

virtuoso-vad-ods - OpenLink Virtuoso Open-Source Edition - Open Data Spaces

virtuoso-vad-rdfmappers - OpenLink Virtuoso Open-Source Edition - RDF Mappers

virtuoso-vad-sparqldemo - OpenLink Virtuoso Open-Source Edition - SPARQL Demo

virtuoso-vad-syncml - OpenLink Virtuoso Open-Source Edition - SyncML

virtuoso-vad-tutorial - OpenLink Virtuoso Open-Source Edition - Tutorial

virtuoso-vsp-startpage - OpenLink Virtuoso Open-Source Edition - Start Page

virtuosoconverter - converts nepomuk database to Virtuoso 6.1.0
```

The package *virtuoso-opensource* is the application that will be installed. The Virtuoso OpenSource server application is installed by issuing the following command through the command line.

```
sudo aptitude install virtuoso-opensource
```

Ubuntu lists all the ancillary application packages required by Virtuoso OpenSource that will also be installed.

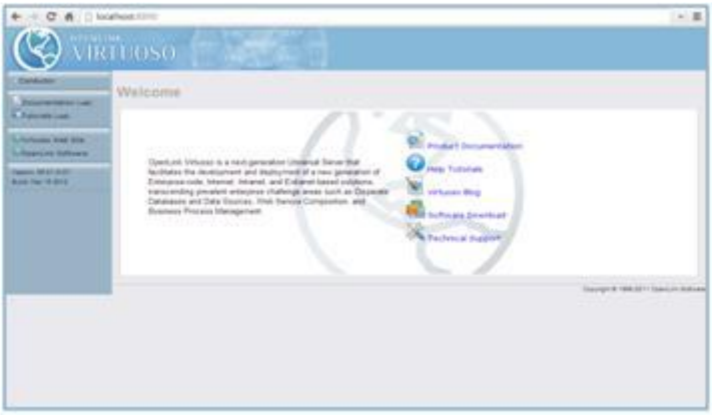
```
The following NEW packages will be installed:

   ghostscript{a} gsfonts{a} libavahi-client3{a} libavahi-common-data{a}            libavahi-
common3{a} libcups2{a} libcupsimage2{a} libgomp1{a} libgs8{a}    libice6{a} libjasper1{a}
libjpeg62{a} liblcms1{a} liblqr-1-0{a}    libltdl7{a} libmagickcore3{a} libmagickwand3{a}
libpaper-utils{a}    libpaper1{a} libreadline5{a} libsm6{a} libtiff4{a} libvirtodbc0{a}
libxt6{a} odbcinst{a} odbcinst1debian2{a} virtuoso-opensource    virtuoso-opensource-6.1{a}
virtuoso-opensource-6.1-bin{a} virtuoso-opensource-6.1-common{a} virtuoso-server{a}    virtuoso-
vad-conductor{a} virtuoso-vsp-startpage{a} x11-common{a}
```

```
0 packages upgraded, 34 newly installed, 0 to remove and 0 not upgraded. Need to get 19.8MB of
archives. After unpacking 63.4MB will be used.
```

As part of the installation, Ubuntu will prompt for passwords to use for the *dba* (main database administrator) and *dav* (WebDAV file system administrator) users. These must not be left blank or VOS will refuse to launch after installation.

```
Setting up libpaper-utils (1.1.24) ...
Setting up libreadline5 (5.2-7build1) ...
Setting up virtuoso-opensource-6.1-common (6.1.2+dfsg1-1ubuntu4) ...
Setting up virtuoso-opensource-6.1-bin (6.1.2+dfsg1-1ubuntu4) ...
Setting up odbcinst (2.2.14p2-1ubuntu1) ...
Setting up odbcinst1debian2 (2.2.14p2-1ubuntu1) ...
Setting up libvirtodbc0 (6.1.2+dfsg1-1ubuntu4) ...
Setting up virtuoso-opensource-6.1 (6.1.2+dfsg1-1ubuntu4) ...
* Starting Virtuoso OpenSource Edition 6.1  virtuoso-opensource-6.1   [ OK ]  Setting up virtuoso-
opensource (6.1.2+dfsg1-1ubuntu4) ...
Setting up virtuoso-vad-conductor (6.1.2+dfsg1-1ubuntu4) ...
Setting up virtuoso-vsp-startpage (6.1.2+dfsg1-1ubuntu4) ...
Setting up virtuoso-server (6.1.2+dfsg1-1ubuntu4) ...
Processing triggers for libc-bin ...
ldconfig deferred processing now taking place
peclarke@ubuntu:~$
```

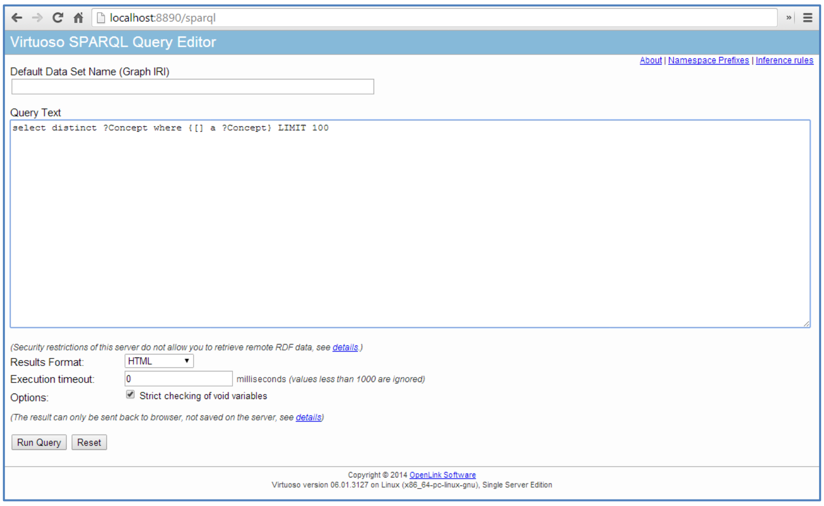
At this point Virtuoso OpenSource (VOS) is installed, running and accessible from *http://localhost:8890* as shown in Figure 5.1.



OpenLink Virtuoso is a next generation Universal Server that facilitates the development and deployment of a new generation of Enterprise-code, Internet, Intranet, and Extranet-based solutions, transcending prevalent enterprise challenge areas such as Disparate Databases and Data Sources, Web Service Composition, and Business Process Management.

Figure 5.1 Virtuoso OpenSource welcome screen (Source: author)

The SPARQL endpoint for the server is accessible from https://localhost:8890/sparql as shown in Figure 5.2



Virtuoso SPARQL Query Editor

Default Data Set Name (Graph IRI)

Query Text

```
select distinct ?Concept where {[] a ?Concept} LIMIT 100
```

(Security restrictions of this server do not allow you to retrieve remote RDF data, see details.)

Results Format: HTML

Execution timeout: 0 milliseconds (values less than 1000 are ignored)

Options: ☑ Strict checking of void variables

(The result can only be sent back to browser, not saved on the server, see details)

Copyright © 2014 OpenLink Software
Virtuoso version 06.01.3127 on Linux (x86_64-pc-linux-gnu), Single Server Edition

Figure 5.2 Virtuoso OpenSource SPARQL endpoint (Source: author)

At this moment, the SPARQL endpoint is installed and running but contains no data. The following section will outline the process involved in loading data into the triplestore.

5.2.2 Loading Data into the Triplestore

This section details the process of loading data into the Virtuoso OpenSource triplestore. It will utilise the OCLC WorldCat data as an example dataset in demonstrating the process. The following commands provide the user with root access and create a directory to store the dataset to be downloaded:

```
sudo -i
mkdir -p /usr/local/data/datasets/worldcat
cd /usr/local/data/datasets/worldcat
```

The WorldCat data dump can be downloaded to the current directory by issuing the following command:

```
wget http://purl.oclc.org/dataset/WorldCat/datadumps/WorldCatMostHighlyHeld-2012-05-
15.nt.gz
```

Pre-processing involves unzipping the data dump to create the .nt file. To unzip the file, issue the following command.

```
gunzip WorldCatMostHighlyHeld-2012-05-15.nt.gz
```

The following command will provide the user with an SQL interface with which to perform transactions on the SPARQL database:

```
isql-vt
```

To register the files to be loaded into the triplestore, issue the following command, providing the location of the file(s), the file types to load and the named graph to assign to the dataset.

```
ld dir all('/usr/local/data/datasets/worldcat', '*.*', 'http://www.oclc.org');
```

The output to this command should resemble the following:

```
SQL> ld dir all('/usr/local/data/worldcat/', '*.nt', 'http://www.oclc.org');
Connected to OpenLink Virtuoso
Driver: 06.01.3127 OpenLink Virtuoso ODBC Driver

Done. -- 2 msec.
SQL>
```

To verify the data that will be loaded into the triplestore, issue the following command:

```
SELECT * FROM DB.DBA.LOAD LIST;
```

The output should resemble the following:

```
SQL> select * from DB.DBA.LOAD LIST;
ll file
ll graph
ll state     ll started              ll done                  ll host       ll work time
ll error
VARCHAR                                    NOT                                      NULL
VARCHAR
INTEGER     TIMESTAMP            TIMESTAMP            INTEGER     INTEGER     VARCHAR
```

```
/usr/local/data/worldcat//WorldCatMostHighlyHeld-2012-05-15.nt
http://www.oclc.org                                                                     0
NULL                 NULL                 NULL        NULL        NULL


1 Rows. -- 1 msec.
SQL>
```

Once the files have been successfully registered, they can be added to the triplestore with the following command:

```
rdf loader run();
```

On the workstation used for this experiment, this process took just under 16 hours to completely load the WorldCat dataset. The output of this command was as follows:

```
SQL> rdf loader run();
Done. -- 57251399 msec.
```

By issuing the Select statement from above, the timestamps from process can be verified.

```
SQL> select * from DB.DBA.LOAD LIST;
ll file
ll graph
ll state     ll started              ll done                  ll host      ll work time
ll error
VARCHAR                                     NOT                                      NULL
VARCHAR
INTEGER     TIMESTAMP            TIMESTAMP            INTEGER     INTEGER     VARCHAR



/usr/local/data/worldcat//WorldCatMostHighlyHeld-2012-05-15.nt
http://www.oclc.org                                                                     2
2014.2.7 8:45.17 0   2014.2.8 0:39.20 0   0           NULL        NULL


1 Rows. -- 15 msec.
```

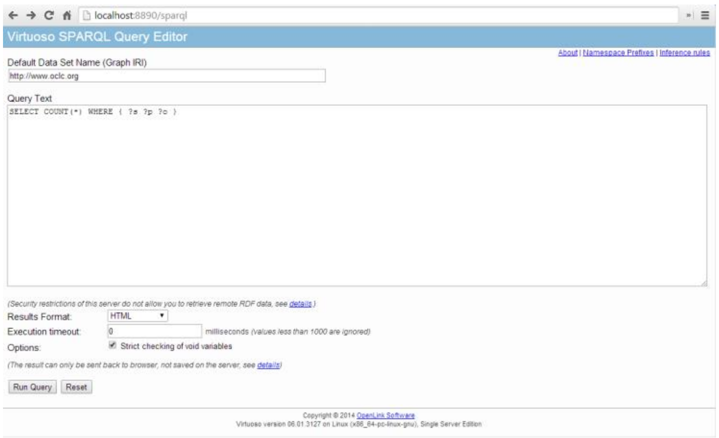
Once the dataset has been loaded, it is recommended to commit this work and create a database checkpoint. This creates a rollback position should corruption occur in the database.

```
commit work;
Done. -- 38 msec.


SQL> checkpoint;
Done. -- 526 msec.
SQL> quit;
```

The process has been completed but it is advised to consult the log file located at */var/lib/virtuoso/db/virtuoso.log* for errors. It is possible that the data has loaded into the triplestore but errors may have arisen and the data could be incomplete.

At this stage in the process it is possible to visit the SPARQL endpoint at localhost:8890/sparql and conduct queries on the data. The following query will provide a count of all triples in the default graph.

```
SELECT COUNT(*) WHERE { ?s ?p ?o }
```



localhost:8890/sparql

Virtuoso SPARQL Query Editor

Default Data Set Name (Graph IRI): http://www.oclc.org

Query Text:

```
SELECT COUNT(*) WHERE { ?s ?p ?o }
```

(Security restrictions of this server do not allow you to retrieve remote RDF data, see details.)

Results Format: HTML

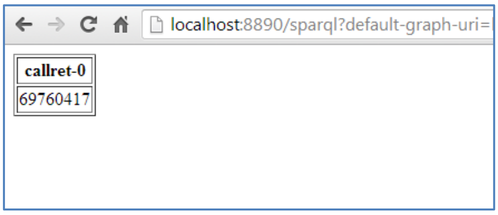
Execution timeout: 0 milliseconds (values less than 1000 are ignored)

Options: ☑ Strict checking of void variables

(The result can only be sent back to browser, not saved on the server, see details)

Copyright © 2014 OpenLink Software
Virtuoso version 06.01.3127 on Linux (x86_64-pc-linux-gnu), Single Server Edition

This process will take a number of minutes and then returns the following output:



localhost:8890/sparql?default-graph-uri=

callret-0
69760417

It is recommended at this stage to stop Virtuoso to back up the dataset. This process is performed using the following commands:

```
sudo -i
cd /
/etc/init.d/virtuoso-opensource                                                stop &&
tar -cvf - /var/lib/virtuoso | gzip --fast > virtuoso-6.1.6-dev-DBDUMP-dbpedia-3.7-
en de-$(date '+%F').tar.gz &&
/etc/init.d/virtuoso-opensource start
```

This section details the process involved in loading the WorldCat dataset into the Virtuoso triplestore. This process was repeated for all subsequent datasets within the experiment.

5.3 Limitations with Technical Aspects of the Experiment

This experiment required a considerable number of days to perform. One significant limitation of the experiment was a result of the selection of the host computer on which the experiment was conducted. In production environments, where timeliness of query responses is a consideration, a number of high-end servers would be employed to host the triplestore. The machine selected for the experiment was adequate generally, but limitations of the hard disk capacity necessitated the larger datasets to be loaded separately. The disk capacity precluded the DBpedia and LinkedGeoData graphs being loaded simultaneously on the machine. This prolonged the experiment but did not impact on the results.

5.4 Conclusions

This chapter discussed the set-up of the Virtuoso SPARQL query service to explore the selected Linked Datasets. First the installation of the Virtuoso SPARQL query service was discussed and next the process of loading the selected Linked Datasets into Virtuoso was discussed. Finally, some of the technical limitations of the hardware used in this experiment were mentioned as well as the impacts of those limitations.

6. PEOPLE-ORIENTATED ASSESSMENT

6.1 Introduction

This chapter discusses the questionnaire deployed to assess the more subjective characteristics of the measurement of the quality of the linked datasets. Section 6.2 outlines in detail the purpose of the survey. Section 6.3 discusses each question of the survey in detail; outlining the purpose of each question, how that question ties back to the main research question, and a summary of the results of that question. Finally Section 6.4 highlights the key findings of the survey.

6.2 Survey

As outlined in the previous chapters, it is recommended that data is evaluated using both objective and subjective measures (Wang and Strong, 1996). The dimensions of provenance and licensing have been identified as needing objective analysis, due to the requirement that they be assessed for the existence of specific attributes. The characteristics of reputation and believability will be examined subjectively as their assessment is based entirely on the subjective opinion of those interacting with the data. The final dimension of verifiability will be assessed both objectively and subjectively as recommended in the previous chapter. This is due to a requirement to objectively verify the usage of dedicated provenance ontologies but also to gain the subjective opinion from a community on the verifiability of a dataset.

The following section outlines the subjective, people-oriented element of the experiment. It details the questions that were posed to the survey cohort and the reasons these questions were posed, with an explanation of how it relates to the research question. Finally, the responses to the survey are identified and remarked upon.

6.3 Survey Questions and Results

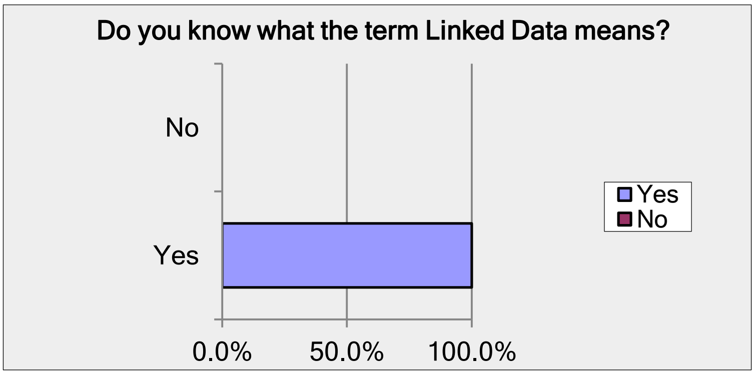
In this section the survey will be examined in detail providing an explanation of the intention of each of the questions and a description of how the question relates to the overall research question. There will also be a discussion of the results achieved for each question and an analysis of the overall questionnaire.

The survey was created following a detailed literature review on the topic of trust in Linked Data. Questions were compiled over a number of days and reflected upon for suitability. The final questionnaire was deployed using SurveyMonkey and emailed to a broad cohort; including colleagues, Linked Data researchers, computer professionals and fellow students. Over the course of 8 days, 35 replies were received, of which 32 were fully completed. The remaining three responses were eliminated from the results as they were incomplete.

Question 1: Do you know what the term Linked Data means?

This was a YES/NO question whose goal was to discern if the respondent is suitably comfortable with the concept to participate in the survey. This question allowed for an assessment of how familiar the respondent is with the concept of Linked Data.

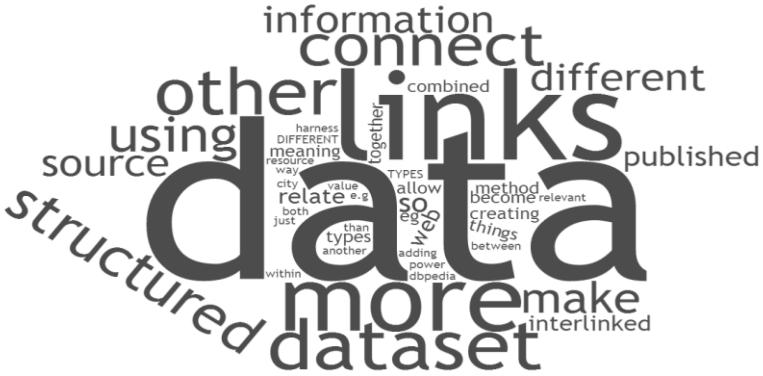
All 32 participants responded that they were familiar with the concept of Linked Data.



Question 2: If "Yes", how would you explain the concept to a non-technical person?

In order to verify the answer to the previous question, participants were asked to provide a brief explanation of what they understood the term to mean. Describing the topic in non-technical terms removes the potential for misleading concepts and terminology being used. This question further clarifies the experience of the participant with regard to Linked Data. It provides a more detailed insight into the participants understanding of the concept.

The majority of the answers correctly related to publishing structured data on the web and linking this to other structured data sets. The following image depicts a word cloud of all responses to the survey.

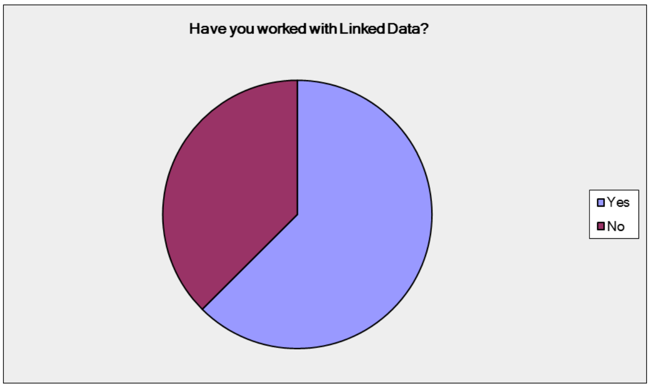


The majority of respondents defined linked data as a method of publishing structured data that could be linked to other data to become more useful. One such reply suggested that Linked Data involved "*attaching more meaning to data by connecting to other datasets with relevance*".

Question 3: Have you worked with Linked Data?

Having identified in questions One and Two whether the participant has knowledge of the topic the questionnaire now attempts to discern what level of experience they have with Linked Data. This question further clarifies the experience of the participant with regard to the Linked Data. It provides a more detailed insight into the participants experience with Linked Data.

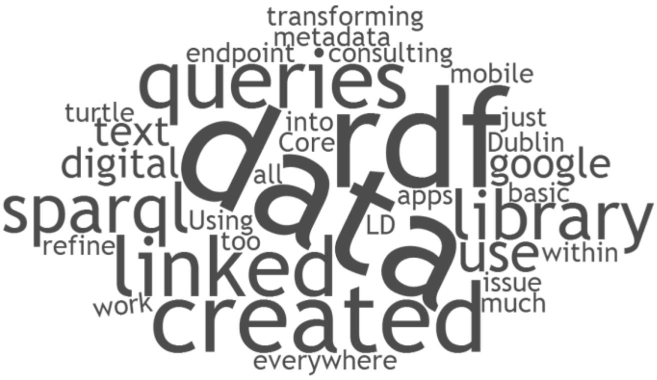
All participants responded to this question with 62.5% declaring that they had worked directly with Linked Data.



Question 4: If "Yes", what is your experience with Linked Data?

In order to assess exactly what experience the participant has with the concept of Linked Data, they have been asked to provide specific examples of their experience with Linked Data. This question serves to further clarify the experience of the participant with regard to the Linked Data. It provides a more detailed insight into the participants experience with Linked Data.

All 20 of the respondents that had replied 'Yes' to the previous question provided a response to this question. The responses indicate that many of the participants had experience in transforming legacy data, in text and CSV formats, into RDF. The following word cloud depicts responses to this question.

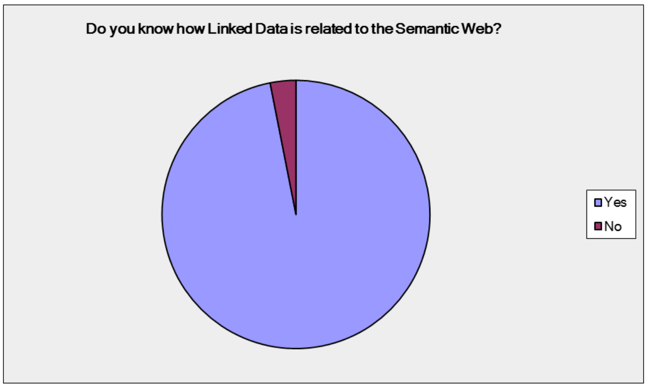


A sample of the answers to this question includes "*Creating RDF for library collections*" and "*Creating structured data, SPARQL queries*".

Question 5: Do you know how Linked Data is related to the Semantic Web?

This question serves to discern that the respondent is suitably comfortable with the concept of the Semantic Web and can discern its difference and relationship to Linked Data. This question further clarifies the experience of the participant with regard to the Linked Data and the Semantic Web. It provides a more detailed insight into the participants understanding of the concept of the Semantic Web.

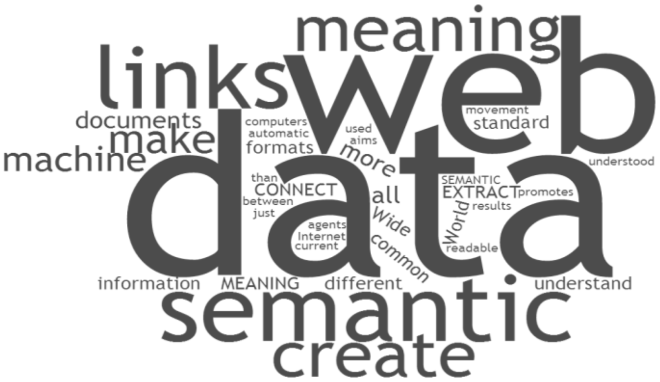
All but one of the respondents answered 'Yes' to this question, indicating that the participants consider themselves familiar with both concepts.



Question 6: If "Yes", how would you explain the concept to a non-technical person?

In order to verify the answer to the previous question, participants were asked to provide a brief explanation of what they understood the term to mean. Describing the topic in non-technical terms removes the potential for misleading concepts and terminology being used. This question further clarifies the experience of the participant with regard to the Semantic Web. It provides a more detailed insight into the participants understanding of the concept.

Of the 31 positive responses to the previous question, 29 elaborated on their answer. The most prominent responses centred on creating meaning from diverse sources of data and enabling a web of data that is understood by computers. The following word cloud gives an overview of the most prominent terms.

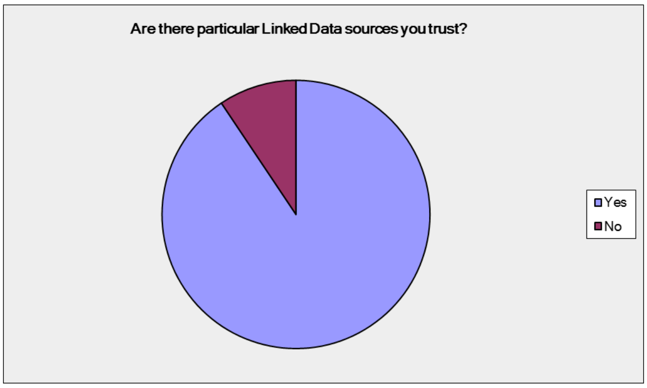


A sample of the answers to this question includes "*Linked Data are the links that create the web. It is making a web of data which is consumable by machines*" and "*The Semantic Web aims to interlink structured data following descriptive standards (metadata).*"

Question 7: Are there particular Linked Data sources you trust?

This was a YES/NO question whose goal was discern whether the respondent is trusting Semantic Web resources. This question further clarifies the experience of the participant with regard to the Semantic Web. It demands that the user reflect on their trust with regard to Semantic Web resources.

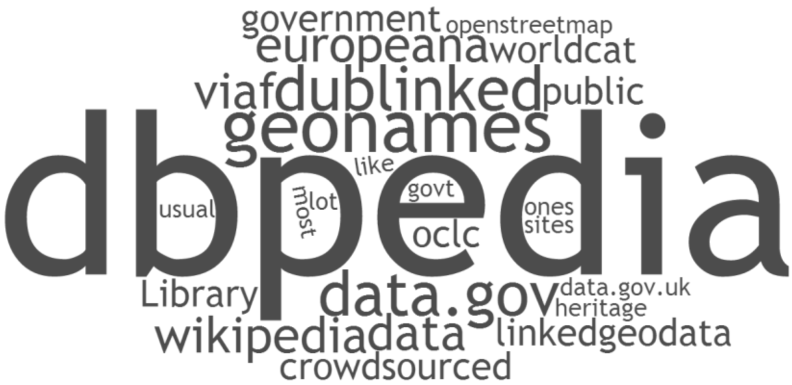
Only three participants answered this question negatively, indicating that over 90% of those surveyed are trusting resources on the Semantic Web.



Question 8: If "Yes", please list some below

In order to verify the answer to the previous question, participants were asked to provide examples of Semantic Web resources they trusted. This question further clarifies the experience of the participant with regard to their trust of Semantic Web resources. It provides a more detailed insight into the participant's experience of the technologies.

All but two of the participants that answered 'Yes' to the previous question provided further information to this question. Perhaps surprisingly, the resource with the most overwhelming levels of trust was DBpedia. The following word cloud provides further insight into the responses to this question.

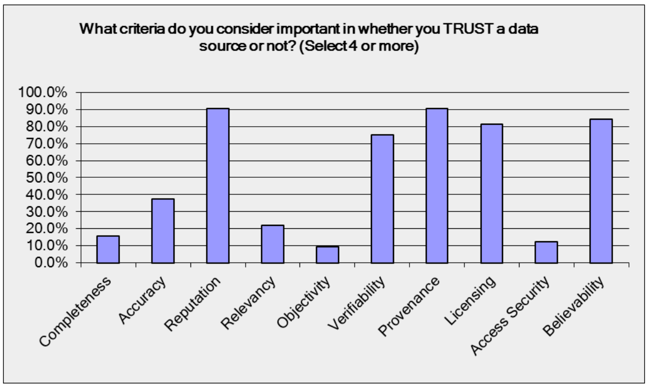


As can be seen above, other prominent sites include VIAF, WorldCat and Europeana. These are library-centric resources and highlight that a significant interest or familiarity with the field of library science. A trust of geospatial data sources can also be identified with the prevalence of Geonames, OpenStreetmap and LinkedGeoData.

Question 9: What criteria do you consider important in whether you TRUST a data source or not? (Select 4 or more)

In order to verify the answer to the previous question, participants were asked to select criteria they considered contributed to their trust in a resource. This question further clarifies the experience of the participant with regard to their trust of Semantic Web resources. It provides a more detailed insight into what qualities the participants expect trusted resources to exhibit.

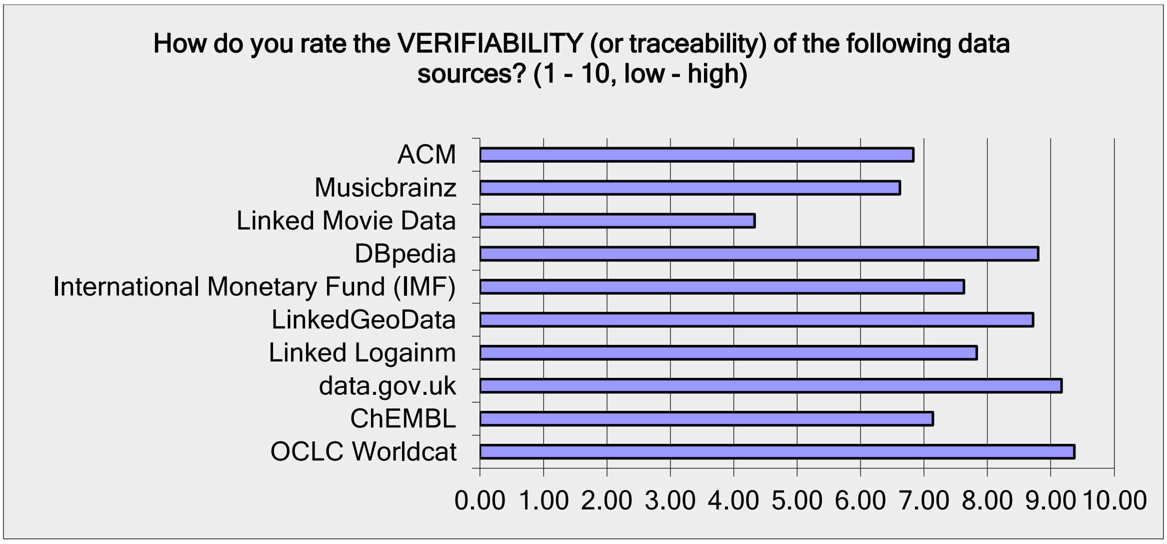
The results indicate that reputation, provenance, licensing, verifiability and believability were the qualities most important to the participants in this survey. Over 90% of responses included reputation and provenance as criteria for trust. The top five responses were made up of all five of the trust characteristics from Chapter 3.



Question 10: How do you rate the VERIFIABILITY (or traceability) of the following data sources? (1 - 10, low - high) Verifiability is described as "the degree by which a data consumer can assess the correctness of a dataset and as a consequence its trustworthiness"

This question serves to clarify the experience of the participant with regard to a number of Semantic Web data sources in terms of verifiability. It provides a more comprehensive insight into the participants experience with these resources by prompting for their rating of the resource in terms of verifiability. The purpose of this question is to gauge the participant's experience of the resource with regard to verifiability. With these responses it is expected that a comparison can be made of all of the resources listed rating them in terms of this metric.

The responses show that a number of resources were deemed to exhibit high verifiability. Namely, DBpedia, LinkedGeoData, data.gov.uk and OCLC WorldCat scored over 80%. This is an interesting result as 50% of these resources are crowd-sourced and the other 50% are highly curated sources of data. The resource that was deemed the least verifiable has not been updated since 2008.



Question 11: Please provide some explanation of your previous answer. Helpful and useful information would include your experience (if any) with the resources listed and your opinions of their verifiability.

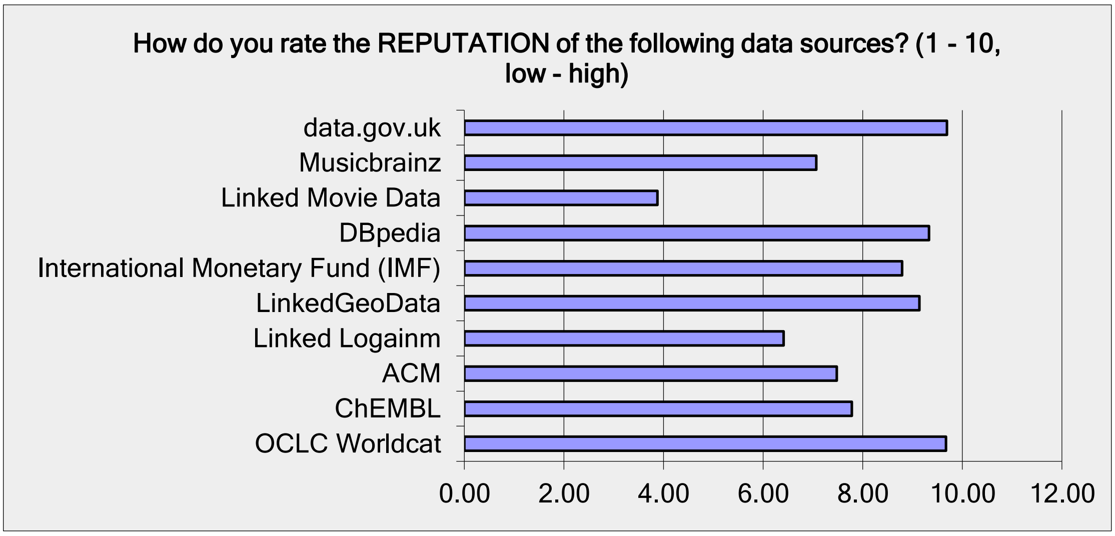
This question serves to further clarify the experience of the participant with regard to a number of Semantic Web data sources in terms of verifiability. It provides a more comprehensive insight into the participants experience with these resources by prompting for their rating of the resource in terms of verifiability. The purpose of this question is to gauge the participant's experience of the resource with regard to verifiability. The participant was asked to elaborate on their answers from the previous question.

The responses for this question were quite poor, with only a 25% response rate. Perhaps it is the case that the previous question called for too much detail or surveyed too many resources. It was however felt that making the response to the question mandatory would jeopardise the overall response to the questionnaire. Of those that provided information, over half of the responses suggested that provenance was partly a source of verifiability. One such respondent indicated that they believed that verifiability was *"based on clear provenance and overall reputation of the source organisation. Gave higher score to sources which clearly list the source of the data, contact details, funding details, creators etc."* Reponses also indicated that the reputation of the resource played a part in their opinion of the verifiability of the resource and that data from government and prominent organisations was rated higher for verifiability. One response indicated "*Scored sources from government or well established, internationally recognised organisations higher. Likely sustainability of the source also considered. A lack of easily found information about the data from an otherwise reputable organisation knocked points off (e.g. ACM)."*

Question 12: How do you rate the REPUTATION of the following data sources? (1 - 10, low - high) Reputation is defined as "a judgment made by a user to determine the integrity of a source. It is mainly associated with a data published, a person, organization, group of people or community of practice rather than being a characteristic of a dataset. The data publisher should be identifiable for a certain (part of) a dataset"

This question serves to clarify the experience of the participant with regard to a number of Semantic Web data sources in terms of reputation. It provides a more comprehensive insight into the participants experience with these resources by prompting for their rating of the resource in relation to reputation. The purpose of this question is to gauge the participant's experience of the resource with regard to reputation. With these responses it is expected that a comparison can be made of all of the resources listed rating them in terms of this metric.

The responses show that a number of resources were deemed to exhibit high reputation. Namely, data.gov.uk, DBpedia, LinkedGeoData, IMF and OCLC WorldCat scored over 80%. This is an interesting result as it demonstrates that the reputation of entities hosting crowd-sourced data is of parity with those highly curated sources of data. Again, the resource that was deemed to possess the lowest reputation has not been updated since 2008. This demonstrates that the respondents provided truthful and accurate answers.



Question 13: Please provide some explanation of your previous answer. Helpful and useful information would include your experience (if any) with the resources listed and your opinions of their reputation.

This question serves to further clarify the experience of the participant with regard to a number of Semantic Web data sources in terms of reputation. It provides a more comprehensive insight into the participants experience with these resources by prompting for their rating of the resource in terms of reputation. The purpose of this question is to gauge the participant’s experience of the resource with regard to reputation. The participant was asked to elaborate on their answers from the previous question.

The responses for this question were quite poor, with a response rate of just over 33%. It may be the case that provided multiple-choice sample answers would have yielded a greater response. Of those that provided information, over half of the responses suggested that they considered most Linked Data sources to have a high reputation.

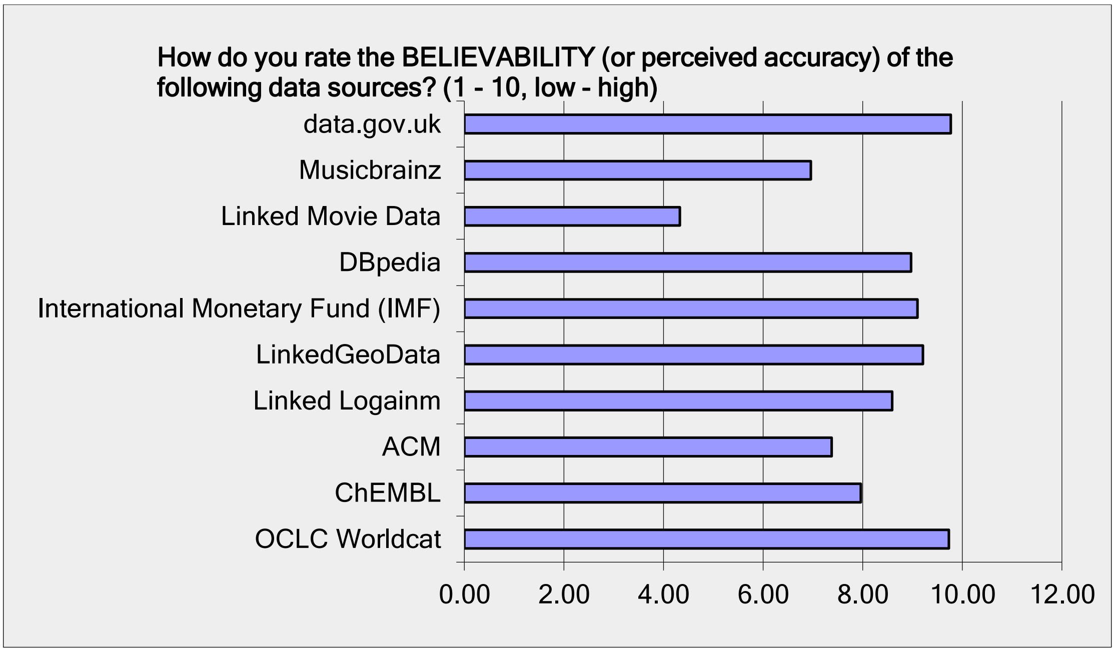
One such response reported that the resources they were familiar with "*are mostly trustworthy*". The 'Linked Movie Data' resource was identified a number of participants as being unknown or non-existent. A response indicated that reputation does not stem from the data but rather a holistic view of the reputation of the body publishing the data. They "*reputation isn't coming from their data, more of a general opinion on the reputation of the organisation."* Data from government or highly familiar sources was recognised possessing a higher reputation.

Question 14: How do you rate the BELIEVABILITY (or perceived accuracy) of the following data sources? (1 - 10, low - high) Believability is defined here as "the extent to which information is regarded as true and credible" and can be considered as 'perceived accuracy'.

Why is this question being asked?

This question serves to clarify the experience of the participant with regard to a number of Semantic Web data sources in terms of believability. It provides a more comprehensive insight into the participants experience with these resources by prompting for their rating of the resource in relation to believability. The purpose of this question is to gauge the participant's experience of the resource with regard to believability. With these responses it is expected that a comparison can be made of all of the resources listed rating them in terms of this metric.

The responses show that a number of resources were deemed to exhibit high believability. Namely, data.gov.uk, DBpedia, LinkedGeoData, Linked Logainm, IMF and OCLC WorldCat scored over 80%. This is an interesting result as it demonstrates that the majority of the data sources selected were deemed to be believable or perceived to be accurate and correct. Once again, the resource that was deemed to possess the lowest reputation has not seen updates since 2008. This demonstrates that the respondents provided truthful and accurate answers.



Question 15: Please provide some explanation of your previous answer. Helpful and useful information would include your experience (if any) with the resources listed and your opinions of their believability (perceived accuracy).

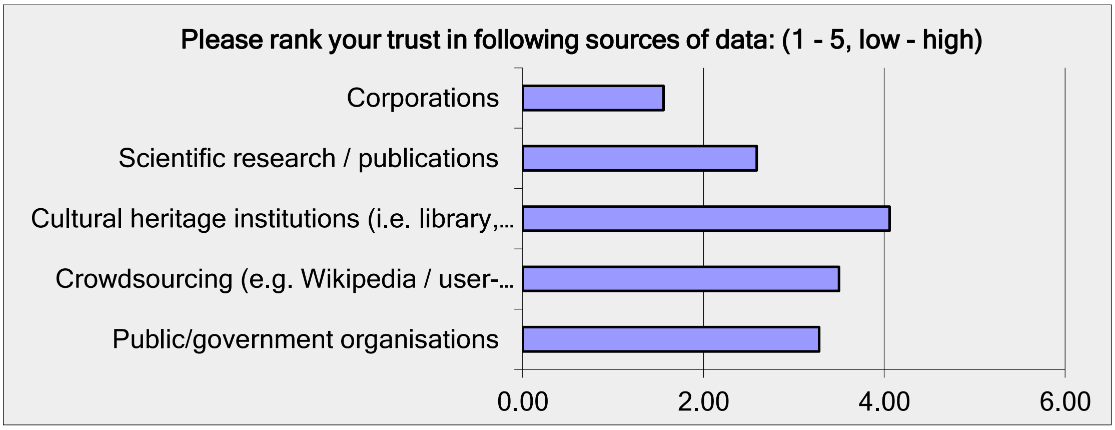
This question serves to further clarify the experience of the participant with regard to a number of Semantic Web data sources in terms of believability. It provides a more comprehensive insight into the participants experience with these resources by prompting for their rating of the resource in terms of believability. The purpose of this question is to gauge the participant's experience of the resource with regard to believability. The participant was asked to elaborate on their answers from the previous question.

The responses for this question were quite poor, with a response rate of just below 33%. These three questions suffered from a low response rate overall. This is perceived to be a result of the question calling for a more descriptive answer but also due to the fact that the response was optional. A mandatory selection of multiple-choice sample answers may have yielded a more positive response. Of those that provided information, the majority of the responses considered most Linked Data sources to be highly believable and accurate. Many of the indicated that "*most of these are believable.*" A number of responses suggested that due to the fact that many of the resources are highly specialized and there are no alternative sources of the information, there is no choice but to believe and use these resources. Two such answers were "*...you have to believe most as usually it's the only source of a particular piece of information e.g. how many other dbpedias are there*?" and "*we have to trust these as there often aren't alternative sources of information."* The 'Linked Movie Data' resource was identified by a number of participants as being unknown or non-existent. A number of responses also indicated that believability is a factor of the general reputation of the organisation publishing the data. One such answer specified "*the high ratings are for organisations where I know the data providers would be credible and 'believable.'*"

Question 16: Please rank your trust in following sources of data: (1 - 5, low - high)

This question serves to further clarify the experience of the respondent with regard to trust of organisations generally. It allows for the previous responses to be verified in terms of organisation. This response allows for an assessment of the trust in organisations generally. The answers from questions 10, 12 and 14 can be compared to the response to this question.

These results are in line with the responses from the previous questions. Cultural heritage institutions, public bodies and crowd-sourced data providers scored highest throughout the survey.



Question 17: Any additional comments related to the survey?

This question allows the participant to provide any additional information they deem to be relevant to the survey or elaborate on any question or topic they choose to. This question enables the respondent to elaborate on their experience in the field of Linked Data and the Semantic Web and to offer further knowledge or opinion on the topic.

There were no meaningful responses to this question. It was remarked by one of the participants that the survey was quite long, so perhaps it was the case that participants were '*burned out*' at this stage. It may also be the case that they considered the questionnaire to be comprehensive. Many of the responses indicated that this was an interesting are of research and that it deserves further examination. From the 32 responses there were 20 participants who sought to be informed of the results and outcomes of the experiment.

6.4 Key Findings of Survey

The survey demonstrated that the Linked Data knowledge of the respondents was high, overall. None of the responses were entirely incorrect with regard to the concept terminology thus indicating that the concepts of Linked Data and the Semantic Web are widely understood.

The results of Question 9 confirmed that the five main characteristics of trust are commonly viewed to be provenance, reputation, verifiability, believability and licensing, as outlined in the previous chapters of this study.

The datasets chosen for analysis correspond to the highest-rated resources in Questions 10, 12 and 14. This demonstrates that there is broad acceptance of the datasets and that attributes such as size, technology, currency and internationality of the dataset are not considered factors that negatively impact on the overall suitability of the data source.

The answers to Questions 10 through 15 indicate that there are strong relationships between the characteristics of provenance, verifiability, reputation and believability.

This supports the authors view as outlined in the Trust Assessment Model for Linked Data from Chapter 3, section 3.5.

6.5 Conclusions

This chapter discussed the questionnaire deployed to assess the more subjective characteristics of the measurement of the quality of the linked datasets. Firstly the purpose of the survey was discussed in detail. Next each question of the survey was discussed in detail; outlining the purpose of each question, how that question ties back to the main research question, and a summary of the results of that question. Finally the key findings of the survey were outlined.

7. TECHNOLOGY-ORIENTATED ASSESSMENT

7.1 Introduction

This chapter discusses the technology-based exploration of the selected Linked Datasets using the Virtuoso SPARQL query service. Section 7.2 reviews the findings of the previous chapter to highlight what has been uncovered, and what has yet to be uncovered, specifically the tangible objective metrics of the quality of the Linked Datasets. Section 7.3 details the objective exploration of each of the five datasets under the heading of *Provenance, Verifiability*, and *Licensing*. Provenance will look at both the VoID description and the Provenance Metadata. The Verifiability will be explored using the Provenance Ontologies and Digital Signatures. The Licensing, if present, is available in the VoID description. Finally Section 7.4 presents the key findings of this evaluation.

7.2 Findings from the People-Oriented Assessment

As discussed in the previous chapter, the results of the people-oriented assessment provided support for the notion that the five main characteristics of trustable Linked Data are; *Provenance*, *Reputation*, *Verifiability*, *Believability* and *Licensing*, as also outlined in the previous chapters of this study. The datasets chosen for analysis outlined in Chapter 4 correspond exactly with the highest-rated resources in which the survey prompted users to rate data sources by three of these factors (*Reputation*, *Verifiability*, and *Believability*). This demonstrates that there is broad acceptance of the datasets and that attributes such as size, technology, currency and internationality of the dataset are not considered factors that negatively impact on the overall suitability of the data source. The results of the survey also indicate that there are strong relationships between the characteristics of *Provenance, Reputation*, *Verifiability*, and *Believability*. This supports the proposed Trust Assessment Model for Linked Data from Section 3.5.

This survey provides significant insight into how trust in data and data sources is established but cannot provide tangible metrics towards objectively assessing the trustworthiness of data. To this end, a technology-oriented assessment of the data sources is required to examine the development of those tangible metrics.

7.3 Technology-Oriented Assessment

As outlined in the previous chapter, it is recommended that researchers evaluate data using both objective and subjective measures (Wang and Strong, 1996). The dimensions of *Provenance* and *Licensing* have been identified as demanding objective analysis, due to the requirement that they be assessed for the existence of certain attributes. The characteristic of *Verifiability* has already been assessed in the survey but will also be examined in the technology-oriented assessment. This is due to a requirement to objectively verify the usage of dedicated *Provenance* ontologies within the dataset. The following section outlines the objective, technology-oriented element of the experiment. It details the dimensions being examined alongside the metrics for that dimension and provides a summary of the steps taken to assess each metric. Finally, a synopsis of the results is provided for each dataset.

7.3.1 The OCLC WorldCat dataset

There is no publicly accessible SPARQL endpoint for this data source. In order to query this data it was necessary to create a local SPARQL endpoint, download the data dump from OCLC and load this into the application for querying. As this is quite a large dataset this process took just under 16 hours. A query to return 20 random records from the dataset for examination returned the following results:

```
http://www.worldcat.org/isbn/9780585030005

http://www.worldcat.org/isbn/9780585030357

http://www.worldcat.org/isbn/9780585030463

http://www.worldcat.org/isbn/9780585030838
```

```
http://www.worldcat.org/isbn/9780585030951

http://www.worldcat.org/isbn/9780585031149

http://www.worldcat.org/isbn/9780585031156

http://www.worldcat.org/isbn/9780585031262

http://www.worldcat.org/isbn/9780585031989

http://www.worldcat.org/isbn/9780585032061

http://www.worldcat.org/isbn/9780585032436

http://www.worldcat.org/isbn/9780585032443

http://www.worldcat.org/isbn/9780585032788

http://www.worldcat.org/isbn/9780585032818

http://www.worldcat.org/isbn/9780585032894

http://www.worldcat.org/isbn/9780585032924

http://www.worldcat.org/isbn/9780585033099

http://www.worldcat.org/isbn/9780585033129

http://www.worldcat.org/isbn/9780585033235

http://www.worldcat.org/isbn/9780585033242
```

7.3.1.1 Provenance

VoID Description

OCLC states that in cases where the organization specifically publishes linked datasets, it will provide a Vocabulary of Interlinked Datasets (VoID) description of each dataset. The VoID will reference the specific license applicable to the dataset and provide guidance on how to satisfy the attribution requirements, if any. A VoID description exists for this dataset but is not accessible at the web root nor via the *well-known* convention described as best practice. The following VoID description was returned from a Google search and is the description used throughout the dataset by OCLC.

```
http://purl.oclc.org/dataset/WorldCat
```

Provenance Metadata

As would be expected from an organization highly proficient with metadata, provenance metadata is present in both the VoID description and the sample RDF files returned for examination. These use a combination of Dublin Core and schema.org elements and attributes rather than a dedicated provenance vocabulary such as PROV, OPMV or PAV. The following provides an example of provenance information contained in the RDF for the following resource http://www.worldcat.org/isbn/9780585030005

```
<http://www.worldcat.org/title/-/oclc/42854417>
        a               gen-ont:ContentTypeGenericResource ;
        dct:created     "1989-08-30" ;
        dct:source      <http://orhddb01dxdu.dev.oclc.org:9006/worldcat/42854417> ;
        void:inDataset  <http://purl.oclc.org/dataset/WorldCat> ;
        schema:about    oclc:42854417 .
```

7.3.1.2 Verifiability

Provenance Ontology

OCLC does not employ dedicated provenance ontologies within its dataset but rather utilizes provenance-related elements from Dublin Core and schema.org, as shown in the previous section.

Digital Signature

There were no instances of digital signatures observed within any of the Linked Data files consulted during this part of the experiment.

7.3.1.3 Licensing

The VoID description contains all the licensing information available for this dataset. This Linked Data release of WorldCat.org is made available by OCLC under the Open Data Commons Attribution License (ODC-BY). ODC-BY is designed to “allow users to freely share, modify, and use” a database while giving attribution to the source of the data. The license does not place any restrictions on use, including commercial use, beyond the attribution requirement (Infotoday, 2012).

```
<dcterms:license rdf:resource="http://opendatacommons.org/licenses/by/1.0/"/>

<cc:attributionURL rdf:resource="http://www.worldcat.org/"/>

<cc:morePermissions rdf:resource="mailto:data@oclc.org"/>

<cc:attributionName>WorldCat</cc:attributionName>

<cc:useGuidelines>
```

7.3.1.4 Synopsis

Dimension	Description	Exists
Provenance	VoID description	Yes
	Provenance metadata	Yes
Verifiability	Provenance ontology	No
	Digital signature	No
Licensing	Machine-readable license metadata	Yes
	Human-readable license	Yes
	Permission metadata	Yes
	Attribution metadata	Yes
	CopyLeft/ShareAlike conditions	Yes

7.3.2 The International Monetary Fund (IMF) dataset

There is no official IMF SPARQL endpoint available for this data source. Rather, the IMF publishes a REST API that can be utilized to query the data. The following website is hosted by DERI and contains a copy of the dataset scraped by Capadisli using SDMX (Capadisli *et al.*, 2013). This data is available as Linked Data for consumption by the general public.

```
http://imf.270a.info/
```

While there is a publicly accessible SPARQL endpoint for this data source available, it was deemed worthwhile by the author to download and deploy a local copy of this data. This dataset is of a moderate size and was loaded into the local instance of Virtuoso in 6 hours. The random query results returned for examination were as follows:

```
<http://imf.270a.info/dataset/MBLD/IAP/MX/CPIS/A/176/PUB/L M/2012>
<http://imf.270a.info/dataset/MBLD/IAP/MX/CPIS/A/178/PUB/L M/2003>
<http://imf.270a.info/dataset/MBLD/IAP/MX/CPIS/A/178/PUB/L M/2004>
<http://imf.270a.info/dataset/MBLD/IAP/MX/CPIS/A/178/PUB/L M/2005>
<http://imf.270a.info/dataset/MBLD/IAP/MX/CPIS/A/178/PUB/L M/2006>
<http://imf.270a.info/dataset/MBLD/IAP/MX/CPIS/A/178/PUB/L M/2007>
<http://imf.270a.info/dataset/MBLD/IAPA/MX/CPIS/A/528/PUB/L M/2012>
<http://imf.270a.info/dataset/MBLD/IAPA/MX/CPIS/A/542/PUB/L M/2012>
<http://imf.270a.info/dataset/MBLD/IAPA/MX/CPIS/A/548/PUB/L M/2012>
<http://imf.270a.info/dataset/MBLD/IAPA/MX/CPIS/A/922/PUB/L M/2012>
<http://imf.270a.info/dataset/MBLD/IAPA/MX/CPIS/A/924/PUB/L M/2012>
<http://imf.270a.info/dataset/MBLD/IAPDG/IT/CPIS/A/853/PUB/L M/2005>
<http://imf.270a.info/dataset/MBLD/IAPDG/IT/CPIS/A/853/PUB/L M/2007>
<http://imf.270a.info/dataset/MBLD/IAPDG/IT/CPIS/A/853/PUB/L M/2010>
<http://imf.270a.info/dataset/MBLD/IAPDG/IT/CPIS/A/853/PUB/L M/2011>
<http://imf.270a.info/dataset/MBLD/IAPDG/IT/CPIS/A/856/PUB/L M/2002>
<http://imf.270a.info/dataset/MBLD/IAPDOF S/IT/CPIS/A/113/PUB/L M/2011>
<http://imf.270a.info/dataset/MBLD/IAPDOF S/IT/CPIS/A/113/PUB/L M/2012>
<http://imf.270a.info/dataset/MBLD/IAPDOFM S/GB/CPIS/A/582/PUB/L M/2005>
<http://imf.270a.info/dataset/MBLD/IAPDOFM_S/GB/CPIS/A/582/PUB/L_M/2006>
```

7.3.2.1 Provenance

VoID Description

A VoID description for this dataset is accessible at the root directory as is considered a best practice.

```
http://imf.270a.info/void.ttl
```

Provenance Metadata

As can be observed from the VoID description, the IMF dataset utilizes the PROV provenance ontology. The entire provenance of the dataset, detailing the dates created, transactions processed and methods employed is visible from the VoID description. There is basic provenance metadata held for the dataset, expressed using dcterms elements. For example:

```
dcterms:title        "International Monetary Fund datasets"@en ;
dcterms:creator      <http://csarven.ca/#i> ;
dcterms:modified     "2014-03-05"^^xsd:date ;
dcterms:publisher    <http://csarven.ca/#i> ;
dcterms:source       <http://www.imf.org/> ;
```

As this dataset utilizes the PROV ontology it is possible to consult the database for more detailed provenance information. By querying the dataset for random triples that contain the *prov#Activity* object it was possible to get a clearer picture of the provenance metadata within the dataset.

```
select ?s where
{?s ?p <http://www.w3.org/ns/prov#Activity>}
offset RANDOM NUMBER
limit 20
```

```
<http://imf.270a.info/provenance/activity/20140305081506>
<http://imf.270a.info/provenance/activity/20140304051457>
<http://imf.270a.info/provenance/activity/20140305081934>
<http://imf.270a.info/provenance/activity/20140304051841>
<http://imf.270a.info/provenance/activity/20140305081449>
<http://imf.270a.info/provenance/activity/20140304051439>
<http://imf.270a.info/provenance/activity/20140305082004>
<http://imf.270a.info/provenance/activity/20140304051906>
<http://imf.270a.info/provenance/activity/20140305081918>
<http://imf.270a.info/provenance/activity/20140304051828>
<http://imf.270a.info/provenance/activity/20140305081636>
<http://imf.270a.info/provenance/activity/20140304051622>
<http://imf.270a.info/provenance/activity/20140305081819>
<http://imf.270a.info/provenance/activity/20140304051735>
<http://imf.270a.info/provenance/activity/20140305081609>
<http://imf.270a.info/provenance/activity/20140304051555>
<http://imf.270a.info/provenance/activity/20140305081857>
<http://imf.270a.info/provenance/activity/20140304051807>
<http://imf.270a.info/provenance/activity/20140305081735>
<http://imf.270a.info/provenance/activity/20140304051708>
```

The following is a snippet of the provenance information maintained for the resource <http://imf.270a.info/provenance/activity/20140305081506>

```
prov:generated <http://imf.270a.info/dataset/MCORE> ;
prov:qualifiedAssociation <http://csarven.ca/linked-sdmx-data> ;
prov:qualifiedUsage  :mor53296e9947bd4 ,  :mor53296e9947c22 ;
prov:startedAtTime   "2014-03-05T08:15:06Z"^^<http://www.w3.org/2001/XMLSchema#dateTime>
;
prov:used                  <http://imf.270a.info/data/MCORE.AU.xml>                  ,
<http://imf.270a.info/data/MCORE.Structure.xml> ;
prov:wasAssociatedWith <https://github.com/csarven/linked-sdmx> ;
prov:wasInformedBy      <http://imf.270a.info/provenance/activity/20140304051457>      ,
<http://imf.270a.info/provenance/activity/20140304045019> ;
prov:wasStartedBy <http://csarven.ca/#i> .
```

7.3.2.2 Verifiability

Provenance Ontology

The IMF dataset utilizes the PROV ontology to detail transformations of the data. There are significant provenance records maintained for this resource as the previous section details.

Digital Signature

There were no instances of digital signatures observed within any of the Linked Data files consulted during this part of the experiment.

7.3.2.3 Licensing

This Linked Data release of IMF dataset is made available via a Creative Commons license. In this instance a CC0 1.0 Universal license is applied. By applying this license, all rights in the content are waived and the objects can be used by anyone without any restrictions. This dataset contains a human-readable license and machine-readable license metadata and grants rights to freely use the dataset even for commercial purposes, without seeking permission or demanding attribution.

```
imf-dataset:imf  a           void:Dataset ;
        dcterms:license      <http://creativecommons.org/publicdomain/zero/1.0/> ;
        dcterms:modified     "2014-03-05"^^xsd:date ;
        dcterms:publisher    <http://csarven.ca/#i> ;
        dcterms:source       <http://www.imf.org/> ;
```

7.3.2.4 Synopsis

Dimension	Description	Exists
Provenance	VoID description	Yes
	Provenance metadata	Yes
Verifiability	Provenance ontology	Yes
	Digital signature	No
Licensing	Machine-readable license metadata	Yes
	Human-readable license	Yes
	Permission metadata	Yes
	Attribution metadata	Yes
	CopyLeft/ShareAlike conditions	Yes

7.3.3 The DBpedia dataset

While there is a publicly accessible SPARQL endpoint for this data source available at http://dbpedia.org/sparql, it was deemed worthwhile by the author to download and deploy a local copy of this data. This dataset is of significant size and took over two days to load into the local Virtuoso instance. The random query results returned for examination were as follows:

```
http://dbpedia.org/resource/Long Lake (Englehart River)

http://dbpedia.org/resource/Long Walk to Freedom

http://dbpedia.org/resource/Salman F Rahman

http://dbpedia.org/resource/New Caledonia cricket team

http://dbpedia.org/resource/The Ambidextrous Universe

http://dbpedia.org/resource/Tiger Mask

http://dbpedia.org/resource/Gladys Taylor

http://dbpedia.org/resource/356 Liguria

http://dbpedia.org/resource/Popstars series

http://dbpedia.org/resource/Let the Good Times Roll (film)

http://dbpedia.org/resource/R%C3%ADo de las Vacas

http://dbpedia.org/resource/Microsoft_PhotoDraw
```

```
http://dbpedia.org/resource/USS Flusser (DD-20)

http://dbpedia.org/resource/International Day of Zero Tolerance to Female Genital Mutilation

http://dbpedia.org/resource/Cowtail Pine

http://dbpedia.org/resource/United Nations Security Council Resolution 1862

http://dbpedia.org/resource/Francis Williams

http://dbpedia.org/resource/Old Appleton, Missouri

http://dbpedia.org/resource/Colvin Run Mill

http://dbpedia.org/resource/1981%E2%80%9382_Detroit_Pistons_season
```

7.3.3.1 Provenance

VoID Description

A VoID description for the dataset is available at the following address but does not comply with the conventions expected with regards to the *well-known* or root directories.

```
http://dbpedia.org/void/page/Dataset
```

Provenance Metadata

There is minimal provenance metadata provided for resources. Many of the traditional provenance elements of *author*, *title*, *publisher* are expressed using the *dbprop* properties (DBpedia, 2012). There is also nominal usage of the PROV provenance ontology where every DBpedia resource records the *wasDerivedFrom* relationship with its original Wikipedia resource page.

```
http://dbpedia.org/resource/Long Lake (Englehart River)


<http://dbpedia.org/resource/Long Lake (Englehart River)>       ns16:wasDerivedFrom

          <http://en.wikipedia.org/wiki/Long Lake (Englehart River)?oldid=466316841> .
```

7.3.3.2 Verifiability

Provenance Ontology

As discussed in the previous section, DBpedia makes use of the PROV ontology to detail the resource from which each DBpedia resource was created. The *wasDerivedFrom* property is the only PROV ontology property used within DBpedia but in the current dataset release this is referred to 12,461,335 times.

Digital Signature

There were no instances of digital signatures observed within any of the Linked Data files consulted during this part of the experiment.

7.3.3.3 Licensing

DBpedia is derived from Wikipedia and is distributed under the same licensing terms as Wikipedia itself. In 2009, with the release of version 3.4, DBpedia moved to a dual-licensing policy to match the licensing policies of Wikipedia. DBpedia data is licensed under the terms of the Creative Commons Attribution-ShareAlike 3.0 license and the GNU Free Documentation License.

DBpedia encourages that attribution is made via DBpedia URIs. By making these URIs visible and active through @href, <link />, or “Link:" tags. When live links are impossible (e.g., in print), a text-based attribution is acceptable to DBpedia.

7.3.3.4 Synopsis

Dimension	Description	Exists
Provenance	VoID description	Yes
	Provenance metadata	Yes
Verifiability	Provenance ontology	Yes
	Digital signature	No
Licensing	Machine-readable license metadata	Yes
	Human-readable license	Yes
	Permission metadata	Yes
	Attribution metadata	Yes
	CopyLeft/ShareAlike conditions	Yes

7.3.4 The LinkedGeoData dataset

While there is a publicly accessible SPARQL endpoint for this data source available at http://linkedgeodata.org/sparql, it proved to be unreliable even in the early stages of the experiment. It was deemed worthwhile by the author to download and deploy a local copy of this data. This dataset is of a significant size and was loaded into the local instance of Virtuoso in 3 days. Due to the nature of the dataset, many of the resources contain simply a resource containing a latitude and longitude variable. Place information was deemed by the author to be a more worthwhile source of metadata. To this end, a query to return random places was employed on this occasion. The random query results returned for examination were as follows:

```
prefix lgd:<http://linkedgeodata.org/>
prefix lgdo:<http://linkedgeodata.org/ontology/>
SELECT DISTINCT ?place
FROM <http://linkedgeodata.org>
WHERE
{
   ?place a lgdo:Place .
   ?place rdfs:label ?label .
}
OFFSET RANDOM NUMBER
LIMIT 20


http://linkedgeodata.org/triplify/node1022149544
http://linkedgeodata.org/triplify/node1045409334
http://linkedgeodata.org/triplify/node1046599231
http://linkedgeodata.org/triplify/node105572715
http://linkedgeodata.org/triplify/node1060913331
http://linkedgeodata.org/triplify/node1068954332
http://linkedgeodata.org/triplify/node1082381714
http://linkedgeodata.org/triplify/node1082799173
http://linkedgeodata.org/triplify/node1082955125
http://linkedgeodata.org/triplify/node1088741426
http://linkedgeodata.org/triplify/node1094802543
http://linkedgeodata.org/triplify/node1094802574
http://linkedgeodata.org/triplify/node1097279235
http://linkedgeodata.org/triplify/node1102505658
http://linkedgeodata.org/triplify/node1098739058
http://linkedgeodata.org/triplify/node1106183318
http://linkedgeodata.org/triplify/node1106614970
http://linkedgeodata.org/triplify/node1110615112
http://linkedgeodata.org/triplify/node1117284869
http://linkedgeodata.org/triplify/node1126197369
```

7.3.4.1 Provenance

VoID Description

There is no VoID description available for this dataset. This makes it difficult for new users of the dataset to gain a comprehensive understanding of the data. Despite this, a REST API and detailed documentation is available which assists new users in getting familiar with the dataset and retrieving their desired query results.

Provenance Metadata

There was no provenance metadata available in any of the sample resources returned by the query.

7.3.4.2 Verifiability

Provenance Ontology

There was no evidence of any common provenance ontology being employed by the dataset. In querying the existence of common provenance properties there were no resources returned.

Digital Signature

There were no instances of digital signatures observed within any of the RDF consulted.

7.3.4.3 Licensing

The LinkedGeoData database is made available under the Open Database License (ODbL) (OpenDataCommons, 2014). This license implores the user to attribute any public use of the database, or works produced from the database, in the manner specified in the ODbL. For any use or redistribution of the database, or works produced from it, the license of the database must be made clear to others and any notices on the original database must be kept intact. The ShareAlike policy incorporated into this license requires any adapted versions of this database, or works produced from an adapted database, to also be offered under an ODbL license.

7.3.4.4 Synopsis

Dimension	Description	Exists
Provenance	VoID description	No
	Provenance metadata	No
Verifiability	Provenance ontology	No
	Digital signature	No
Licensing	Machine-readable license metadata	No
	Human-readable license	Yes
	Permission metadata	No
	Attribution metadata	No
	CopyLeft/ShareAlike conditions	No

7.3.5 The education.data.gov.uk Dataset

There is no publicly available data dump download for this data so it was not possible to create a local instance of this dataset for querying. There is a publicly accessible SPARQL endpoint for this data source made available, which was used to perform the general query as per the previous elements of this experiment.

```
http://openuplabs.tso.co.uk/sparql/gov-education
```

It should be noted that this data source, while not hosted by data.gov.uk, is widely referenced by UK government publications and thus can be deemed to be official. It is also worth noting that although this data is from 2009, it was last updated in 2013 (Datahub.io, 2014).

The random query results returned for examination were as follows:

```
<http://education.data.gov.uk/id/school/536153>

<http://education.data.gov.uk/id/school/census/119714>
```

```
<http://education.data.gov.uk/id/school/census/120972>
<http://education.data.gov.uk/id/school/535297>
<http://education.data.gov.uk/id/school/census/106138>
<http://education.data.gov.uk/id/school/521044>
<http://education.data.gov.uk/id/school/119020>
<http://education.data.gov.uk/id/school/census/100295>
<http://education.data.gov.uk/id/school/102947>
<http://education.data.gov.uk/id/school/135112>
<http://education.data.gov.uk/id/school/census/104657>
<http://education.data.gov.uk/id/school/census/111539>
<http://education.data.gov.uk/id/school/census/102847>
<http://education.data.gov.uk/id/school/census/113137>
<http://education.data.gov.uk/id/school/census/121845>
<http://education.data.gov.uk/id/school/census/121925>
<http://education.data.gov.uk/id/school/census/114822>
<http://education.data.gov.uk/id/school/census/120688>
<http://education.data.gov.uk/id/school/census/107518>
<http://education.data.gov.uk/id/school/census/119308>
```

7.3.5.1 Provenance

VoID Description

There is no VoID description available for this dataset. This makes it difficult for new users of the dataset to gain a comprehensive understanding of the data. Despite this, a REST API and supporting documentation is available which assists new users in getting familiar with the dataset and retrieving their desired query results.

Provenance Metadata

There was no provenance metadata available in any of the sample resources returned by the query.

7.3.5.2 Verifiability

Provenance Ontology

There was no evidence of any common provenance ontology being employed by the dataset. In querying the existence of common provenance properties there were no resources returned.

Digital Signature

There were no instances of digital signatures observed within any of the RDF consulted.

7.3.5.3 Licensing

Data.gov.uk cites that data and information provided by www.data.gov.uk are available under terms described in the “license” or "constraints" field of individual dataset metadata (2014). In the records returned by the random query there were no instances of such license metadata. All dataset metadata published on www.data.gov.uk are licensed under the Open Government Licence (The National Archives, 2010). Users are free to copy, publish, distribute and adapt the data both non-commercially or for commercial gain. However, it is stipulated that the specified attribution statement must be included with the data together with a link to the license, where possible.

7.3.5.4 Synopsis

Dimension	Description	Exists
Provenance	VoiD description	No
	Provenance metadata	No
Verifiability	Provenance ontology	No
	Digital signature	No
Licensing	Machine	No
	Human	Yes
	Permission	Yes
	Attribution	Yes
	CopyLeft/ShareAlike	Yes

7.4 Key Findings of the Evaluation

The evaluation finds that there is significant disparity between the implementation of measures associated with the characteristics of trusted data and the perception of the resource as a trusted dataset. Despite there being broad agreement and standards regarding the publication of dataset VoID descriptions (Keith Alexander *et al.*, 2011), 40% of the datasets reviewed failed to publish this information. This makes it difficult for users unfamiliar with the dataset to gain a full understanding of the dataset and the data contained within. Where VoID descriptions were present, only one resource, the IMF dataset, presented this metadata along accepted best practice. The publication of VoID descriptions is presently not mandatory and is considered to be a courtesy to the user. The author has the view that when computers are potentially the prime audience of these datasets it has to become a mandatory requirement for data providers to implement VoID descriptions.

Dedicated provenance ontologies, such as PROV and OPMV, were not widely employed amongst the evaluated datasets. Only 60% of those resources assessed published provenance metadata to any degree. The resource with the most detailed provenance data, the IMF dataset, did not come from the field of library science, as could be expected. As demonstrated, the usage of digital signatures to provide automated verifiability was non-existent in the review sample. The author suspects that either the Linked Data community do not feel that this is the best method to certify verifiability or that perhaps the overheads associated with implementing such solutions are too great. The legal implications of not implementing a licensing policy are great. This ensures that licenses are widely deployed by the sample data sources, albeit to varying degrees. LinkedGeoData was the sole data provider to publish no licensing metadata, within the data sampled.

It has been discovered during the course of the experiment that a number of the resources identified as trustworthy have failed to adequately implement the technology-oriented trust characteristics. LinkedGeoData and education.data.gov.uk, failed to meet any of the provenance or verifiability measures and did not completely meet the licensing requirements. The OCLC WorldCat, IMF and DBpedia datasets represent exemplars for trust in Linked Data, having adequately met most of the trust metrics identified.

It is clear that there is further work required in this field and that the establishment of best practice and focused procedural instruction would contribute greatly to the Linked Data community and trust in data, generally. With this in mind, instructional materials will be created that can be consulted by future data publishers seeking to strongly embed the characteristics of provenance, verifiability, reputation, believability and licensing within their data.

7.5 Conclusions

This chapter discussed the technology-based exploration of the selected Linked Datasets. First a review of the findings of the previous chapter was presented to highlight what had been uncovered; and the lack of tangible objective metrics that these did not provide. Following this, an objective exploration of each of the five datasets under the heading of *Provenance, Verifiability*, and *Licensing* was presented. Finally the key findings of this evaluation were summarised and presented.

Based on the outcomes of this chapter and the previous chapter it has been possible to identify some of the key deficits, both perceived and actual, that prevent more widespread development and tagging of Linked Datasets. In the following chapter an instructional artefact will be developed to help address some of these deficits.

8. PROCESS-ORIENTATED ASSESSMENT

8.1 Introduction

This chapter presents the development of a set of instructional materials that represent the outcomes of the experimental work of the previous two chapters as well as some of the key issues uncovered in the literature exploration. In Section 8.2 the development of the instructional materials is discussed, with its focus on issues that appear to be missing in the development of quality Linked Datasets. Next Section 8.3 outlines the evaluation of these materials by three participants took part in a final knowledge-based focus group, whose feedback is presented here.

8.2 Creation of Instructional Materials

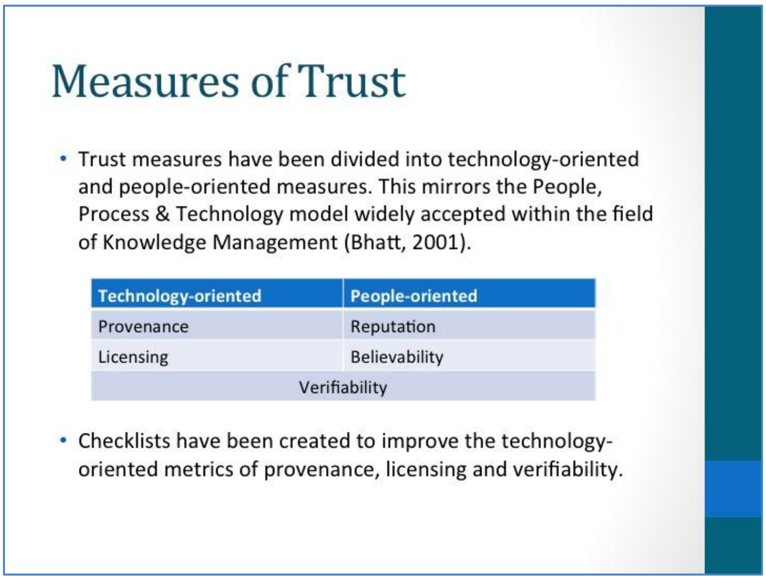
As discussed by Bhatt (2001), successful knowledge management is dependent upon the harmonious relationship between the people, processes and technologies. The following section outlines the process-oriented element of the experiment. Following on from both Chapter 6 (the people-oriented assessment) and Chapter 7 (the technology-oriented assessment), this chapter presents the development of an instructional document to highlight the key learnings in this experiment, and focuses on the key “knowledge gaps” that exist in the creation of quality Linked Datasets. This learning material was presented and discussed with a number of participants and a process-oriented assessment of the measures of trust in Linked Data was executed. This section details the features of that learning material and itemises the responses, questions and opinions voiced by the participants. Finally, the responses to the survey are identified and remarked upon.

In order to conduct the process-oriented element of the experiment, three short presentations were held, lasting between 15 and 20 minutes each, which presented the instructional document to individuals who are experts in the field of Linked Data. Two of the participants had previously completed the people-oriented assessment questionnaire from Chapter 6.

The first two slides shape the field of research and seek to compound the participant's comprehension of the concepts by providing definitions to both Linked Data and the Semantic Web.

The next slide introduces the concept of Data Quality and then provides an overview of Zaveri's data quality dimensions (Zaveri *et al.*, 2012). At this stage, the participants were asked to provide their opinions on trust. Participant #1 (P1) responded that they viewed trust as "*an unwritten agreement between two parties that they will do no harm*". Participant #2 (P2) answered that they considered trust to be based upon a common "*understanding of confidence in a relationship*". Participant #3 (P3) provided a similar response to P2.

The following slide introduces the topic of Trust and elaborates on its position on the Semantic Web technology stack. Slide 8 (Figure 8.1) presents the common measures of trust, separating the five characteristics into Technology-oriented (objective) and People-oriented (subjective) measures.



Measures of Trust

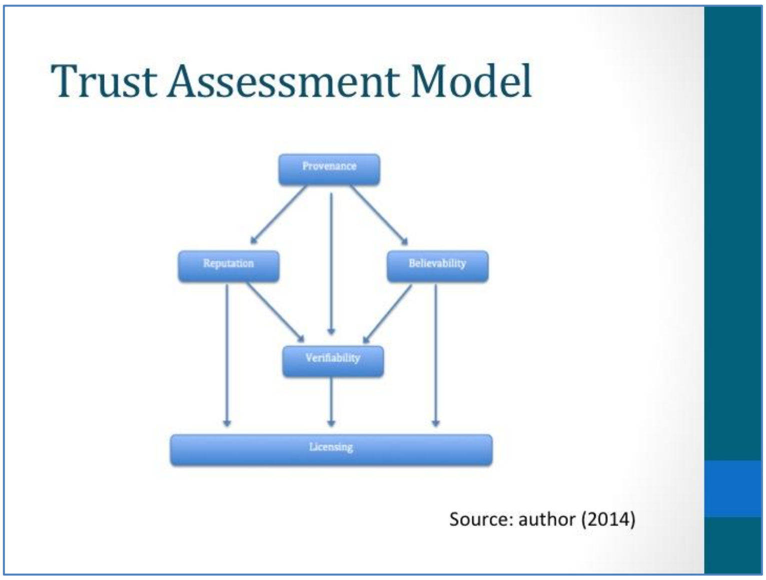
- Trust measures have been divided into technology-oriented and people-oriented measures. This mirrors the People, Process & Technology model widely accepted within the field of Knowledge Management (Bhatt, 2001).

Technology-oriented	People-oriented
Provenance	Reputation
Licensing	Believability
Verifiability	

- Checklists have been created to improve the technology-oriented metrics of provenance, licensing and verifiability.

Figure 8.1 Measures of Trust slide (Source: author)

The *Trust Assessment Model* from Chapter 3 is then introduced (Figure 8.2) with the relationships between the five measures elaborated on.

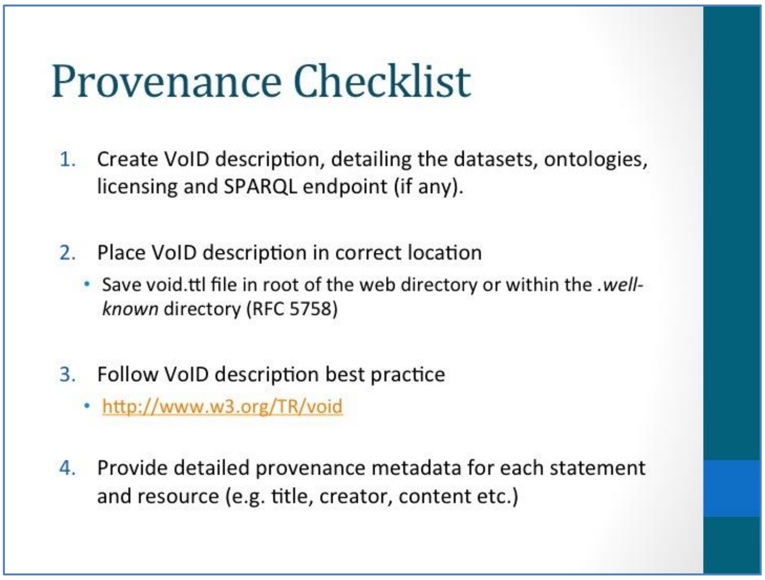


Trust Assessment Model

Source: author (2014)

Figure 8.2 Trust Assessment Model slide (Source: author)

Each of the three technology-oriented measures is then discussed in turn. Provenance is introduced, with its primary features, in this context, identified. A provenance checklist is presented (Figure 8.3), outlining the steps that should be taken to increase trust in the data being created.

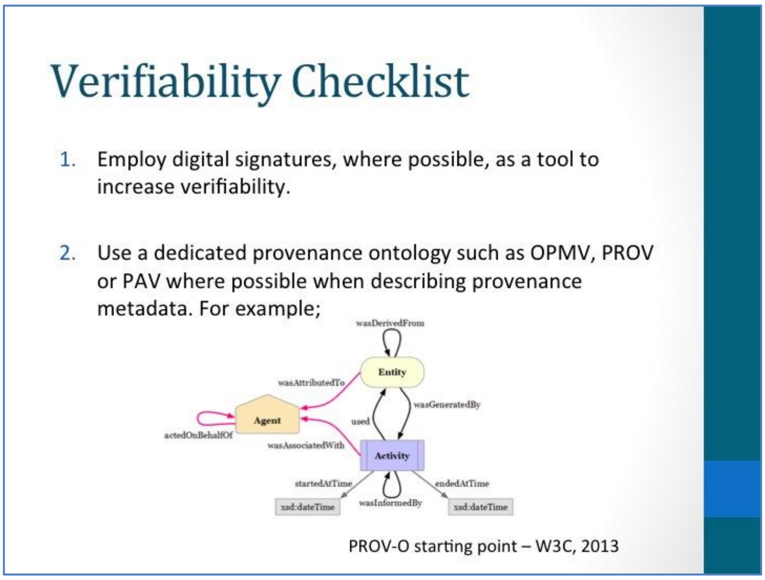


Provenance Checklist

1. Create VoID description, detailing the datasets, ontologies, licensing and SPARQL endpoint (if any).
2. Place VoID description in correct location
   - Save void.ttl file in root of the web directory or within the *.well-known* directory (RFC 5758)
3. Follow VoID description best practice
   - http://www.w3.org/TR/void
4. Provide detailed provenance metadata for each statement and resource (e.g. title, creator, content etc.)

Figure 8.3 Provenance Checklist slide (Source: author)

Verifiability is then introduced and its predominant features are discussed, with a focus on trust. A verifiability checklist is outlined (Figure 8.4), defining the steps that should be taken to increase trust in the data being created. At this point, P1 remarked that they "*don't see a difference between Provenance and Verifiability.*"



Verifiability Checklist

1. Employ digital signatures, where possible, as a tool to increase verifiability.

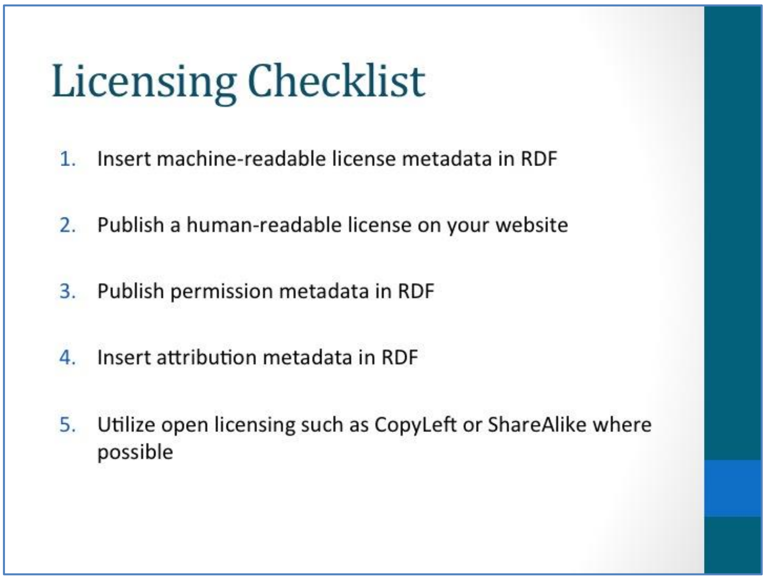
2. Use a dedicated provenance ontology such as OPMV, PROV or PAV where possible when describing provenance metadata. For example;

wasDerivedFrom
Entity
wasAttributedTo
wasGeneratedBy
Agent
used
actedOnBehalfOf
wasAssociatedWith
Activity
startedAtTime
endedAtTime
xsd:dateTime
wasInformedBy
xsd:dateTime

PROV-O starting point – W3C, 2013

Figure 8.3 Verifiability Checklist slide (Source: author)

The final measure, licensing, is then reviewed, providing an introduction and definition to the measure. A licensing checklist is defined, detailing the steps that should be taken to increase trust in the data being created and providing an example from the *OCLC WorldCat* dataset.



Licensing Checklist

1. Insert machine-readable license metadata in RDF

2. Publish a human-readable license on your website

3. Publish permission metadata in RDF

4. Insert attribution metadata in RDF

5. Utilize open licensing such as CopyLeft or ShareAlike where possible

Figure 8.3 Licensing Checklist slide (Source: author)

Lastly, the topic is reviewed and the participants are provided time to ask any questions they may have on the topic.

8.3 Response to Instructional Material

This section serves to summarise the responses and questions raised by the participants in the process-oriented element of the experiment. Three participants took part in a final knowledge-based focus group. Two of the participants had previously completed the people-oriented element of the experiment.

Participant #1 (P1) had not completed the survey. During the presentation they provided their view that *Provenance* and *Verifiability* were too close in context to warrant being separate measures of trust and that they could perhaps be amalgamated. The participant raised the question regarding provenance; "*How does provenance really 'prove' anything? Can we trust that provenance is actually the truth?*" The question whether provenance be backdated or begun from the present moment was raised also. P1 suggested that a check sheet or RDF template could be created that would benefit those creating RDF from legacy data files such as spreadsheets and metadata. Their response to the material was positive on the whole with the individual stating *"Overall, I understood the metrics and found the checklists helpful. This would make an excellent poster presentation at a conference related to the field of information and data science."*

Participant #2 (P2) had completed the survey. At the verifiability phase of the presentation the individual raised the question "*Why are digital signatures still part of the measures if they aren't used by data providers, generally?*" Their opinion was that this should be uniform throughout the field, possibly built into RDF tools, as there is no benefit in having sporadic implementations of the technology. Participant #2 also questioned whether this was, in fact, simply a variation of Tim Berners-Lee "Oh yeah?" button (Bizer et al., 2009) that could be "*stuck onto web resources that would provide a response or rating with regard to trust*". The question of whether penalties could be applied to data providers of untrustworthy data also arose. When prompted for their opinion on the document the individual remarked, "*The instructional document is very good. It provides a strong overview of that is required for creating data that can be trusted*". However, they did not feel "*that this guarantees that the data will be trusted, there are other criteria that can outweigh the metrics of provenance, verifiability and licensing, e.g. reputation*". They finally questioned whether a weighting for the five measures could be devised and applied.

Participant #3 (P3) had also completed the survey previously. This respondent voiced very few questions but was attentive throughout and understood the material. They questioned "*Are there not, or why are there no, W3C standards for this kind of thing already?*" Following this the individual remarked upon the possibility of creating "*a 'Trust Standard' that allows for a data source to be given a score rating. This measure could be included into the RDF of the dataset of the VoID description*" and/or displayed on the website of the provider. When prompted for a general view of the learning material, they answered that they considered it to be "*a great introduction to an interesting subject. I had never thought about a lot of this until the survey asked the questions. Can this be applied to data generally?"*

8.3.1 Summary of Feedback

Based on the feedback received from the three participants, it is clear that the concepts of Linked Data and the Semantic Web are well understood. The measures of trust require further clarification and standardisation as questions were raised regarding the definitions and the precise metrics. Each of the participants presented their own suggestions of potential extensions to research or prospective solutions to the problem. This suggests that the topic is both widely considered by the community and also broadly understood. There was general agreement amongst the participants that there is a need for a trust framework and a consensus that the instructional material above represents the first steps towards developing such a framework.

8.4 Conclusions

This chapter discussed the development and evaluation of a set of instructional materials that represent the key learnings that have been accrued during the various experimental procedures undertaken during the course of this research. The instructional materials represent the best practice “*knowledge gaps*” that appear to be missing in the development of quality Linked Datasets. Three participants took part in a final knowledge-based focus group, to evaluate these materials. It was strongly agreed that there is a for a trust framework.

9. CONCLUSIONS AND FUTURE WORK

9.1 Introduction

This chapter presents a review of the key findings of this research and includes some suggestions for future research directions. The key goal of the research was to explore what represents quality and trustworthiness in Linked Datasets. The five quality criteria that were used to explore this issue were *Provenance, Licensing, Reputation, Believability,* and *Verifiability,* and these were explored as either being subjective, objective or both.

9.2 Conclusions

The primary area of research in this dissertation focused on the topic of trust on the Semantic Web. It specifically attempted to create a trust framework by which data could be created and assessed.

This research began by conducting a literature review on the topics of the Semantic Web and Knowledge Management. The review focused on clarifying the often-misunderstood concepts and identifying the technologies utilised within the field. The Linked Data landscape was assessed and proven to be developing at rapid pace.

The findings from the literature review were interpreted to develop a suitable method to assess the trustworthiness of Linked Data and a framework for the creation of trustworthy Linked Data. A number of trust characteristics were identified within the literature review from which a Trust Assessment Model was created.

The trust characteristics also contributed to the creation of a questionnaire from which a number of Linked Data researchers, library professionals and technologists were evaluated.

9.3 Contribution to the Body of Knowledge

A Trust Assessment Model has been created by the author as part of this research that outlines the key characteristics of trust of Linked Data sources. The model demonstrates the inter-connectedness and dependencies of each of these metrics and provides an insight into the criteria by which trust is assessed.

Although the topic of trust on the Semantic Web has been widely discussed, to date, there has been no analysis of this subject using a people-process-technology approach.

Instructional material has been created which can assist those creating Linked Data in the future or assessing existing Linked Data from a technology-oriented perspective.

9.4 Key Findings

The literature review demonstrated that the trust aspects of data quality should be evaluated using objective and subjective assessments. It was found that these correspond with technology-oriented and people-oriented assessments, as prescribed by Bhatt (2001).

The literature review identified a number of data quality characteristics that apply to trust. Through analysis of existing research the dependencies that exist between each of these characteristics were identified.

The findings of the people-oriented assessment element of the experiment indicated that the knowledge of the Semantic Web was high overall. There was broad agreement on the metrics identified within he literature review. The survey also confirmed the strong relationships that exist between the characteristics of trust.

The findings of the technology-oriented assessment show that there is a disparity between the data providers considered trustworthy and the implementation many of the technical measures considered to ensure the resource will be trusted. This demonstrated the power of reputation.

The process-oriented assessment demonstrated that within the survey cohort there is broad understanding of, and agreement on, the concepts related to Semantic technologies. The instructional material was found to adequately explain the measures to put in place to create Linked Data that conforms to the characteristics of trust.

9.4.1 Key Outcomes Achieved

1. Performed a literature review of the Semantic Web and Knowledge Management.
2. Performed a literature review of trust, identifying the key characteristics and measures.
3. Assessed trust using a people-oriented approach.
4. Assessed trust with a technology-oriented approach.
5. Assessed trust using a process-oriented approach.
6. Created a trust assessment model and framework that can be applied to Linked Datasets.

9.5 Future Work

There exists an opportunity to automate many of the metrics determined within this research. It would be possible to create a tool that would embed the technology-oriented aspects of trust into the creation of future Linked Data. Should this become mandatory, or best practice, the overall quality and trust of Linked Data would grow.

The automation of the people-oriented aspects of the experiment could also be examined for future study. Many of these measures could be derived from the relationships that exist with other datasets. Undergoing a trust evaluation process before linking to a prominent, trusted dataset could ensure that these links were trustworthy. This would in turn improve the overall verifiability, believability and reputation of the dataset.

The instructional material should be developed further. It is hoped that through conducting further instruction sessions, the process-oriented aspects could be refined and contribute to the Linked Data community.

The domain of the Linked Datasets may have an impact on which quality characteristics are of most importance. Therefore, another experiment where datasets from a very specific domain, e.g. medical services, might reveal a requirement for special emphasis on one of the existing criteria, or the need for a new set of additional quality criteria.

Repeating the experiment with a larger set of datasets over a longer period of time may also have uncovered additional requirements. Also of interest would be to examine the same dataset over a number of distinct time periods, e.g. 2008, 2010 and 2012, and to examine whether that dataset was adhering more closely to the quality criteria of venturing further away.

BIBLIOGRAPHY

Antoniou, G., Van Harmelen, F., 2004. A semantic web primer. MIT press.

Artz, D., Gil, Y., 2007. A survey of trust in computer science and the Semantic Web. Web Semantics: Science, Services and Agents on the World Wide Web 5, 58–71. doi:10.1016/j.websem.2007.03.002

Barney, J.B., Hansen, M.H., 1994. Trustworthiness as a source of competitive advantage. Strategic management journal 15, 175–190.

Bellinger, G., Castro, D., Mills, A., 2006. Data, Information, Knowledge, and Wisdom (2004). Available at: www. systems-thinking. org/dikw/dikw. htm (accessed: 5 February 2006).

Berners-Lee, T., 2000. Semantic Web - XML2000 - slide “Architecture” [WWW Document]. URL http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html (accessed 1.19.14).

Berners-Lee, T., 2009. Linked Data - Design Issues [WWW Document]. URL http://www.w3.org/DesignIssues/LinkedData.html (accessed 7.1.12).

Berners-Lee, T., 2011. Tim Berners-Lee’s FOAF file [WWW Document]. URL http://dig.csail.mit.edu/2008/webdav/timbl/foaf.rdf (accessed 2.2.14).

Bhatt, G.D., 2001. Knowledge management in organizations: examining the interaction between technologies, techniques, and people. Journal of Knowledge Management 5, 68–75. doi:10.1108/13673270110384419

Bizer, C., 2007. Quality Driven Information Filtering: In the Context of Web Based Information Systems. VDM Publishing.

Bizer, C., Cyganiak, R., 2011. Quality-driven information filtering using the WIQA policy framework. Web Semantics: Science, Services and Agents on the World Wide Web 7.

Bizer, C., Heath, T., Berners-Lee, T., 2009. Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems 5, 1–22. doi:10.4018/jswis.2009081901

Bizer, C., Jentzsch, A., Cyganiak, R., 2010. State of the LOD Cloud (Preliminary Release No. Version 0.1).

Bizer, C., Jentzsch, A., Cyganiak, R., 2011. State of the LOD Cloud (Preliminary Release No. Version 0.1).

Bruce, T.R., Hillmann, D.I., 2004. The Continuum of Metadata Quality: Defining, Expressing, Exploiting. ALA Editions.

Capadisli, S., Auer, S., Ngonga Ngomo, A.-C., 2013. Linked SDMX Data. Semantic Web.

Casebourne, I., Davies, C., Fernandes, M., Norman, N., 2012. Assessing the accuracy and quality of Wikipedia entries compared to popular online encyclopaedias [WWW Document]. URL https://meta.wikimedia.org/wiki/Research:Accuracy_and_quality_of_Wikipedia_entries (accessed 2.15.14).

Cayzer, S., 2004. Semantic blogging and decentralized knowledge management. Communications of the ACM 47, 47–52.

Chapman, A., 2005. Principles of data quality.

Cyganiak, R., 2011. How to Publish Open Data.

Cyganiak, R., 2012. VoID: Metadata for RDF Datasets.

Dai, C., Lin, D., Bertino, E., Kantarcioglu, M., 2008. An approach to evaluate data trustworthiness based on data provenance, in: Secure Data Management. Springer, pp. 82–98.

Dang, Q., 2010. Internet X.509 Public Key Infrastructure: Additional Algorithms and Identifiers for DSA and ECDSA [WWW Document]. URL https://tools.ietf.org/html/rfc5758 (accessed 2.21.14).

data.gov.uk, 2014. Terms and conditions | data.gov.uk [WWW Document]. URL http://data.gov.uk/terms-and-conditions (accessed 2.12.14).

Datahub.io, 2014. education.data.gov.uk - Datahub.io [WWW Document]. URL http://datahub.io/dataset/education-data-gov-uk (accessed 2.11.14).

Davenport, T.H., Prusak, L., 2000. Working knowledge: how organizations manage what they know. Harvard Business School Press, Boston, Mass.

Davis, R., Shrobe, H., Szolowits, P., 1993. What is a Knowledge Representation?

DBpedia, 2012. DBpedia Data Set Properties [WWW Document]. URL http://wiki.dbpedia.org/Datasets/Properties (accessed 2.19.14).

Decker, S., Mitra, P., Melnik, S., 2000. Framework for the semantic Web: an RDF tutorial. IEEE Internet Computing 4, 68–73. doi:10.1109/4236.895018

Dishongj, 2012. OCLC adds linked data to WorldCat.org [WWW Document]. URL http://www.oclc.org/news/releases/2012/201238.htm (accessed 8.25.12).

DuCharme, B., 2011. Learning SPARQL, 1st ed. O'Reilly Media.

Fensel, D., 2003. Ontologies:: A Silver Bullet for Knowledge Management and Electronic Commerce. Springer.

Flemming, A., 2010. Quality characteristics of linked data publishing datasources (Master’s thesis). Humboldt-Universität zu Berlin.

Flemming, A., Hartig, O., 2010. Quality Criteria for Linked Data sources.

Gamble, M., Goble, C., 2011. Quality, trust, and utility of scientific data on the web: Towards a joint model, in: Proceedings of the 3rd International Web Science Conference. ACM, p. 15.

Golbeck, J., Hendler, J., 2004. Inferring reputation on the semantic web, in: Proceedings of the 13th International World Wide Web Conference. Citeseer.

Golbeck, J., Mannes, A., 2006. Using Trust and Provenance for Content Filtering on the Semantic Web., in: MTW.

Grandison, T., Sloman, M., 2000. A survey of trust in internet applications. Communications Surveys & Tutorials, IEEE 3, 2–16.

Gruber, T., 1993. What is an Ontology.

Hartig, O., 2010. Towards a data-centric notion of trust in the semantic web, in: 2nd Workshop on Trust and Privacy on the Social and Semantic Web SPOT2010, Heraklion (Greece).

Hartig, O., Zhao, J., 2009. Using Web Data Provenance for Quality Assessment. SWPM 526.

Hartig, O., Zhao, J., 2010. Publishing and Consuming Provenance Metadata on the Web of Linked Data, in: McGuinness, D.L., Michaelis, J.R., Moreau, L. (Eds.), Provenance and Annotation of Data and Processes, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 78–90.

Hausenblas, M., 2009. Exploiting Linked Data to Build Web Applications. IEEE Internet Computing 13, 68–73.

Hausenblas, M., Karnstedt, M., 2010. Understanding Linked Open Data as a Web-Scale Database, in: Advances in Databases Knowledge and Data Applications (DBKDA), 2010 Second International Conference on. pp. 56 –61. doi:10.1109/DBKDA.2010.23

Heath, T., Bizer, C., 2011. Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web: Theory and Technology 1, 1–136. doi:10.2200/S00334ED1V01Y201102WBE001

Hebeler, J., Fisher, M., Blace, R., Perez-Lopez, A., 2011. Semantic Web Programming. John Wiley & Sons.

Hislop, D., 2013. Knowledge Management in Organizations: A Critical Introduction. Oxford University Press.

Hogan, A., Umbrich, J., Harth, A., Cyganiak, R., Polleres, A., Decker, S., 2012. An empirical survey of Linked Data conformance. Web Semantics: Science, Services and Agents on the World Wide Web 14, 14–44.

Infotoday, 2012. OCLC Recommends ODC-BY for WorldCat Data [WWW Document]. URL http://newsbreaks.infotoday.com/NewsBreaks/OCLC-Recommends-ODCBY-for-WorldCat-Data-84389.asp (accessed 2.19.14).

Jacobi, I., Kagal, L., Khandelwal, A., 2011. Rule-Based Trust Assessment on the Semantic Web, in: Bassiliades, N., Governatori, G., Paschke, A. (Eds.), Rule-Based Reasoning, Programming, and Applications, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 227–241.

Jurisica, I., Mylopoulos, J., Yu, E., 2004. Ontologies for Knowledge Management: An Information Systems Perspective. Knowl. Inf. Syst. 6, 380–401. doi:10.1007/s10115-003-0135-4

Keith Alexander, Richard Cyganiak, Michael Hausenblas, Jun Zhao, 2011. Describing Linked Datasets with the VoID Vocabulary [WWW Document]. URL http://www.w3.org/TR/void/ (accessed 2.21.14).

Library of Congress, 2012. Library of Congress Launches Beta Release of Linked Data Classification « In Custodia Legis: Law Librarians of Congress [WWW Document]. URL http://blogs.loc.gov/law/2012/07/library-of-congress-launches-beta-release-of-linked-data-classification/ (accessed 8.25.12).

Mendes, P.N., Bryl, V., Bizer, C., 2014. Sieve - Linked Data Quality Assessment and Fusion [WWW Document]. URL http://sieve.wbsg.de/ (accessed 2.20.14).

Mendes, P.N., Mühleisen, H., Bizer, C., 2012. Sieve: linked data quality assessment and fusion, in: Proceedings of the 2012 Joint EDBT/ICDT Workshops. ACM, pp. 116–123.

Möller, K., Hausenblas, M., Cyganiak, R., Grimnes, G., Handschuh, S., 2010. Learning from Linked Open Data Usage: Patterns & Metrics [WWW Document]. URL http://journal.webscience.org/302/ (accessed 9.1.12).

Mui, L., Mohtashemi, M., Halberstadt, A., 2002. A computational model of trust and reputation, in: System Sciences, 2002. HICSS. Proceedings of the 35th Annual Hawaii International Conference on. IEEE, pp. 2431–2439.

Naumann, F., 2002. Quality-driven query answering for integrated information systems. Springer.

Nonaka, I., Takeuchi, H., 1997. The knowledge-creating company. 1995.

Open Data Institute, 2014. Guide to Open Data Licensing [WWW Document]. URL http://theodi.org/guides/publishers-guide-open-data-licensing (accessed 3.20.14).

OpenDataCommons, 2014. ODC Open Database License (ODbL) Summary | Open Data Commons [WWW Document]. URL http://opendatacommons.org/licenses/odbl/summary/ (accessed 2.19.14).

Park, J.-R., 2009. Metadata Quality in Digital Repositories: A Survey of the Current State of the Art. Cataloging & Classification Quarterly 47, 213–228. doi:10.1080/01639370902737240

Pedrinaci, C., Domingue, J., 2011. Linked Services and the Future Internet.

Pipino, L., Wang, R., Kopcso, D., Rybold, W., 2005. Developing measurement scales for data-quality dimensions. ME Sharpe, New York.

Pipino, L.L., Lee, Y.W., Wang, R.Y., 2002. Data quality assessment. Commun. ACM 45, 211–218. doi:10.1145/505248.506010

Rajabi, E., Kahani, M., Sicilia, M.-A., 2012. Trustworthiness of linked data using pki, in: Proceedings of the World Wide Web Conference (WWW2012).

Rolland, N., Chauvel, D., 2000. Knowledge horizons: the present and the promise of knowledge management. Butterworth-Heinemann, Boston.

semanticweb.com, 2011. Quality Indicators for Linked Data Datasets.

Stallings, W., Brown, L., Bauer, M.D., Howard, M., 2008. Computer security: principles and practice. Prentice Hall, Upper Saddle River, N.J.

Stroka, S., 2010. Knowledge Representation Technologies in the Semantic Web.

Tan, W.C., 2007. Provenance in Databases: Past, Current, and Future. IEEE Data Eng. Bull. 30, 3–12.

The National Archives, 2010. Open Government Licence [WWW Document]. URL http://www.nationalarchives.gov.uk/doc/open-government-licence/version/2/ (accessed 2.11.14).

Villazón-Terrazas, B., Vilches-Blázquez, L.M., Corcho, O., Gómez-Pérez, A., 2011. Methodological Guidelines for Publishing Government Linked Data, in: Wood, D. (Ed.), Linking Government Data. Springer New York, pp. 27–49.

w3.org, 2002. Resource Description Framework (RDF): Concepts and Abstract Data Model.

W3C, 2013. PROV-O: The PROV Ontology [WWW Document]. URL http://www.w3.org/TR/prov-o/

Wang, R.Y., Strong, D.M., 1996. Beyond accuracy: what data quality means to data consumers. J. Manage. Inf. Syst. 12, 5–33.

Wang, X.H., Zhang, D.Q., Gu, T., Pung, H.K., 2004. Ontology based context modeling and reasoning using OWL, in: Pervasive Computing and Communications Workshops, 2004. Proceedings of the Second IEEE Annual Conference on. IEEE, pp. 18–22.

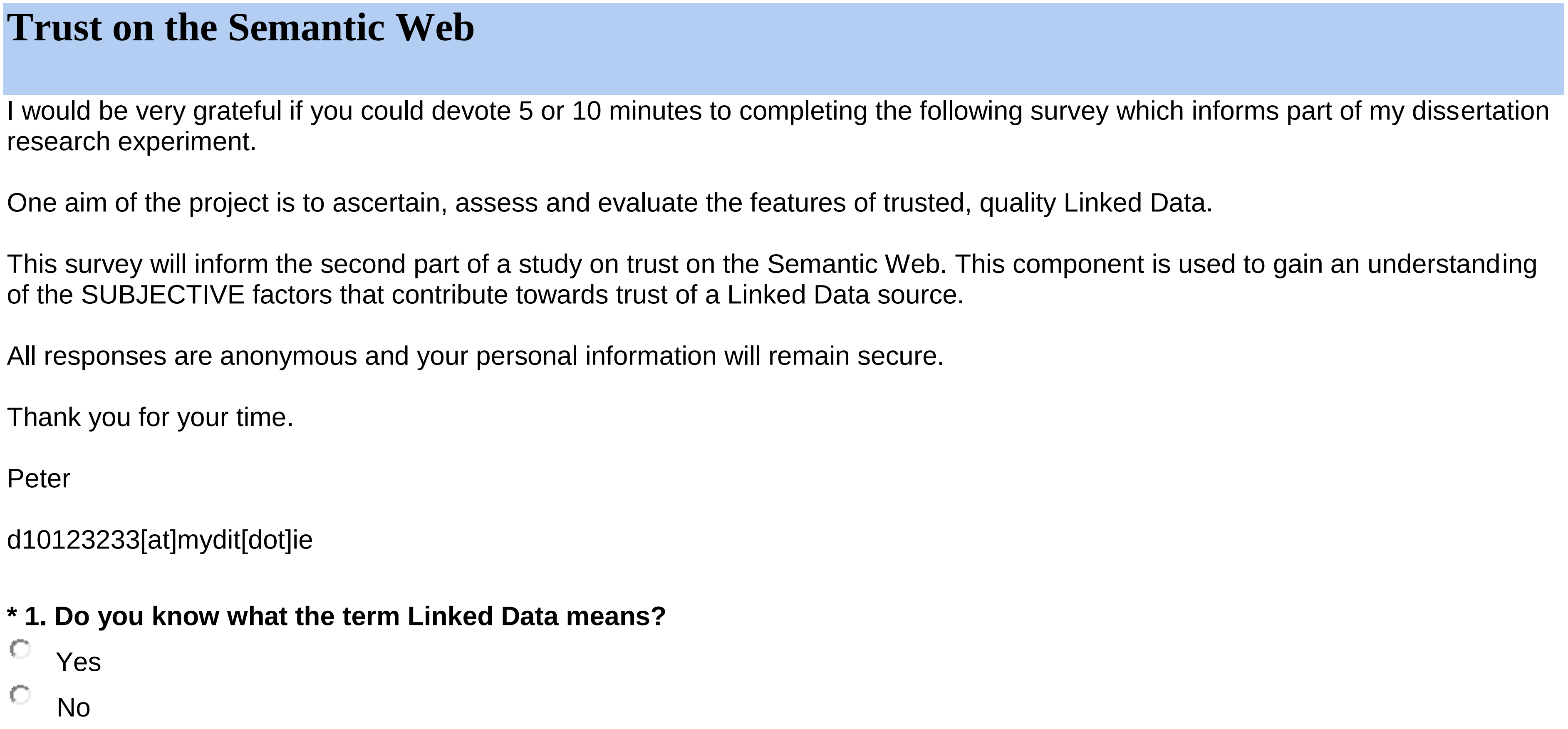
Wikipedia, 2013. Linked data. Wikipedia, the free encyclopedia.

Zaveri, A., Kontokostas, D., Sherif, M.A., Bühmann, L., Morsey, M., Auer, S., Lehmann, J., 2013. User-driven Quality Evaluation of DBpedia, in: To Appear in Proceedings of 9th International Conference on Semantic Systems, I-SEMANTICS ’13, Graz, Austria, September 4-6, 2013. ACM.

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S., 2012. Quality Assessment Methodologies for Linked Open Data | www.semantic-web-journal.net [WWW Document]. URL http://www.semantic-web-journal.net/content/quality-assessment-methodologies-linked-open-data (accessed 8.2.13).

Zhao, J., Hartig, O., 2012. Towards Interoperable Provenance Publication on the Linked Data Web., in: LDOW.

APPENDIX A – People-Oriented Assessment



Trust on the Semantic Web

I would be very grateful if you could devote 5 or 10 minutes to completing the following survey which informs part of my dissertation research experiment.

One aim of the project is to ascertain, assess and evaluate the features of trusted, quality Linked Data.

This survey will inform the second part of a study on trust on the Semantic Web. This component is used to gain an understanding of the SUBJECTIVE factors that contribute towards trust of a Linked Data source.

All responses are anonymous and your personal information will remain secure.

Thank you for your time.

Peter

d10123233[at]mydit[dot]ie

*** 1. Do you know what the term Linked Data means?**

○ Yes

○ No

2. If "Yes", how would you explain the concept to a non-technical person?

*** 3. Have you worked with Linked Data?**

○ Yes

○ No

4. If "Yes", what is your experience with linked data

*** 5. Do you know how Linked Data is related to the Semantic Web?**

○ Yes

○ No

6. If "Yes", how would you explain the concept to a non-technical person

*** 7. Are there particular Linked Data sources you trust?**

○ Are there particular Linked Data sources you trust? Yes

○ No

8. If "Yes", please list some below

*** 9. What criteria do you consider important in whether you TRUST a data source or not? (Select 4 or more)**

□ Believability

□ Reputation

□ Provenance

□ Verifiability

□ Licensing

□ Relevancy

□ Accuracy

□ Objectivity

□ Completeness

□ Access Security

*** 10. How do you rate the VERIFIABILITY (or traceability) of the following data sources? (1 - 10, low - high)**

Verifiability is described as “the degree by which a data consumer can assess the correctness of a dataset and as a consequence its trustworthiness”

	1	2	3	4	5	6	7	8	9	10	Not sure
ChEMBL	○	○	○	○	○	○	○	○	○	○	○
data.gov.uk	○	○	○	○	○	○	○	○	○	○	○
Musicbrainz	○	○	○	○	○	○	○	○	○	○	○
International Monetary Fund (IMF)	○	○	○	○	○	○	○	○	○	○	○
LinkedGeoData	○	○	○	○	○	○	○	○	○	○	○
OCLC Worldcat	○	○	○	○	○	○	○	○	○	○	○
DBpedia	○	○	○	○	○	○	○	○	○	○	○
Linked Movie Data	○	○	○	○	○	○	○	○	○	○	○
Linked Logainm	○	○	○	○	○	○	○	○	○	○	○
ACM	○	○	○	○	○	○	○	○	○	○	○

11. Please provide some explanation of your previous answer.
Helpful and useful information would include your experience (if any) with the resources listed and your opinions of their verifiability.

*** 12. How do you rate the REPUTATION of the following data sources? (1 - 10, low - high)**

Reputation is defined as “a judgment made by a user to determine the integrity of a source. It is mainly associated with a data published, a person, organization, group of people or community of practice rather than being a characteristic of a dataset. The data publisher should be identifiable for a certain (part of) a dataset"

	1	2	3	4	5	6	7	8	9	10	Not sure
ChEMBL	○	○	○	○	○	○	○	○	○	○	○
Linked Logainm	○	○	○	○	○	○	○	○	○	○	○
Linked Movie Data	○	○	○	○	○	○	○	○	○	○	○
DBpedia	○	○	○	○	○	○	○	○	○	○	○
Musicbrainz	○	○	○	○	○	○	○	○	○	○	○
OCLC Worldcat	○	○	○	○	○	○	○	○	○	○	○
ACM	○	○	○	○	○	○	○	○	○	○	○
International Monetary Fund (IMF)	○	○	○	○	○	○	○	○	○	○	○

	1	2	3	4	5	6	7	8	9	10	Not sure
LinkedGeoData	○	○	○	○	○	○	○	○	○	○	○
data.gov.uk	○	○	○	○	○	○	○	○	○	○	○

13. Please provide some explanation of your previous answer.
Helpful and useful information would include your experience (if any) with the resources listed and your opinions of their reputation.

*** 14. How do you rate the BELIEVABILITY (or perceived accuracy) of the following data sources? (1 - 10, low - high)**

Believability is defined here as "the extent to which information is regarded as true and credible” and can be considered as 'perceived accuracy'.

	1	2	3	4	5	6	7	8	9	10	Not sure
ChEMBL	○	○	○	○	○	○	○	○	○	○	○
OCLC Worldcat	○	○	○	○	○	○	○	○	○	○	○
LinkedGeoData	○	○	○	○	○	○	○	○	○	○	○
International Monetary Fund (IMF)	○	○	○	○	○	○	○	○	○	○	○
Linked Logainm	○	○	○	○	○	○	○	○	○	○	○

	1	2	3	4	5	6	7	8	9	10	Not sure
DBpedia	○	○	○	○	○	○	○	○	○	○	○
Linked Movie Data	○	○	○	○	○	○	○	○	○	○	○
Musicbrainz	○	○	○	○	○	○	○	○	○	○	○
ACM	○	○	○	○	○	○	○	○	○	○	○
data.gov.uk	○	○	○	○	○	○	○	○	○	○	○

15. Please provide some explanation of your previous answer.
Helpful and useful information would include your experience (if any) with the resources listed and your opinions of their believability (perceived accuracy).

*** 16. Please rank your trust in following sources of data: (1 - 5, low - high)**

	1	2	3	4	5
Public/government organisations	○	○	○	○	○
Crowdsourcing (e.g. Wikipedia / user-generated)	○	○	○	○	○
Corporations	○	○	○	○	○

	1	2	3	4	5
Cultural heritage institutions (i.e. library, museum)	○	○	○	○	○
Scientific research / publications	○	○	○	○	○

17. Any additional comments related to the survey?

18. Please add your email address if you would like the results of this work to be shared with you

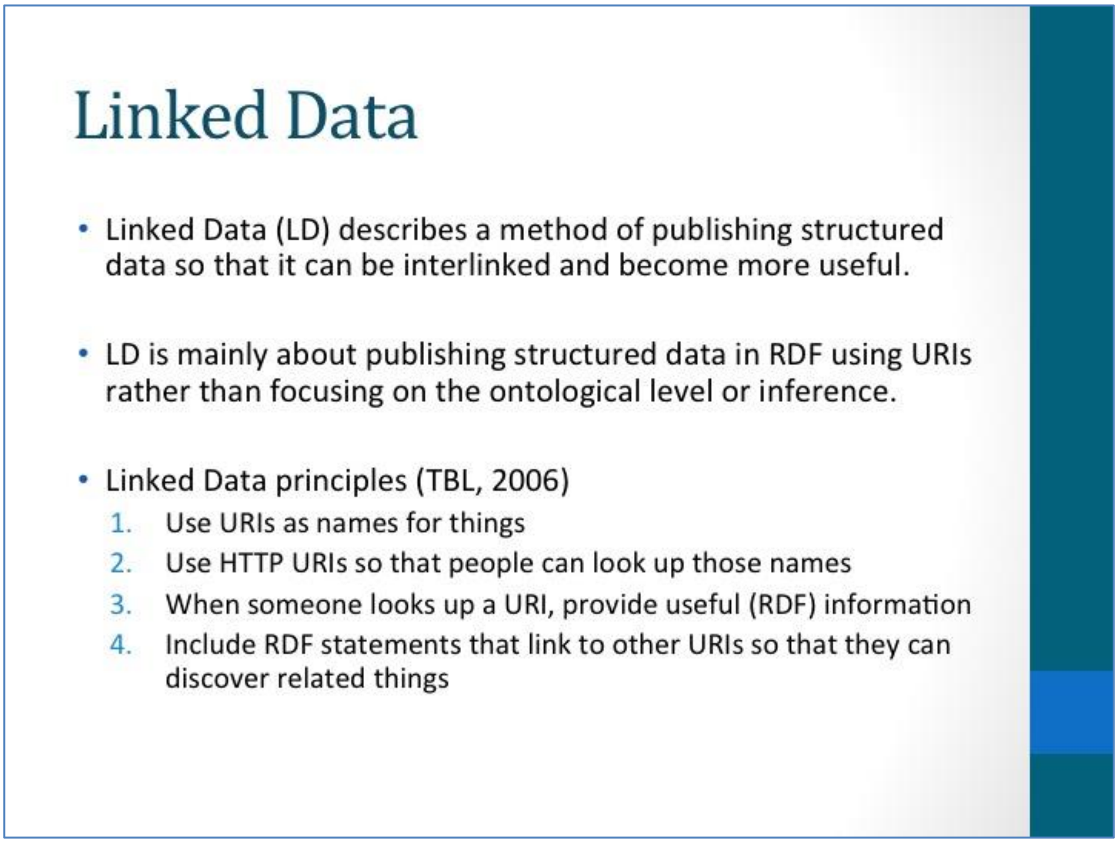
That's it! Please submit your survey answers by clicking DONE.

Again, thank you for your time.

Peter

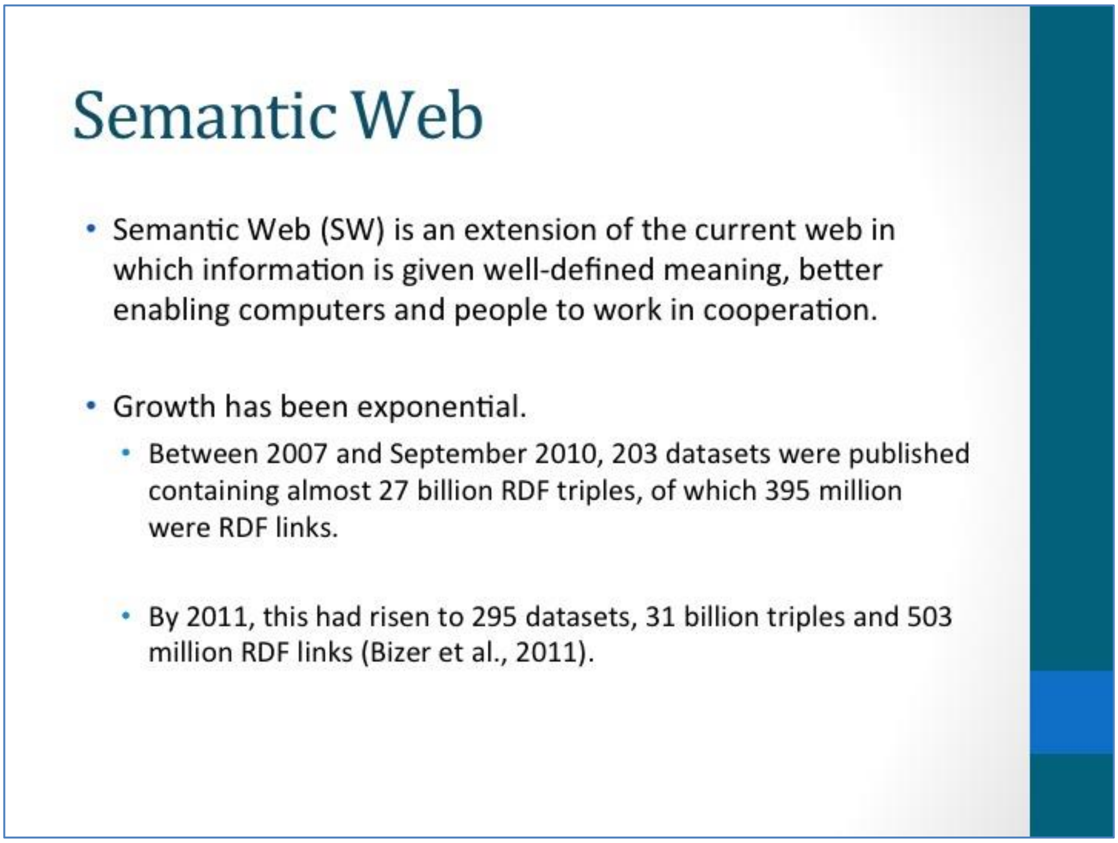
Trust in Linked Data

Measures towards creating trusted data on the Semantic Web



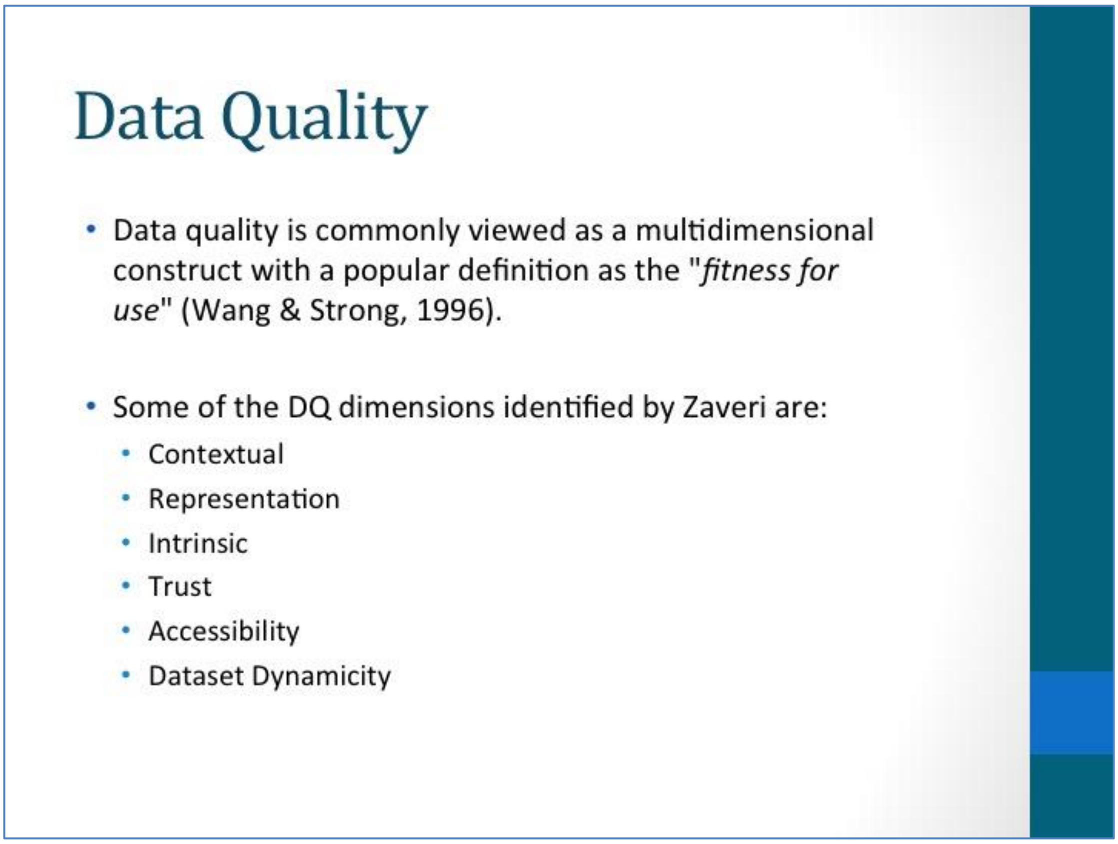
Linked Data

- Linked Data (LD) describes a method of publishing structured data so that it can be interlinked and become more useful.
- LD is mainly about publishing structured data in RDF using URIs rather than focusing on the ontological level or inference.
- Linked Data principles (TBL, 2006)
  1. Use URIs as names for things
  2. Use HTTP URIs so that people can look up those names
  3. When someone looks up a URI, provide useful (RDF) information
  4. Include RDF statements that link to other URIs so that they can discover related things



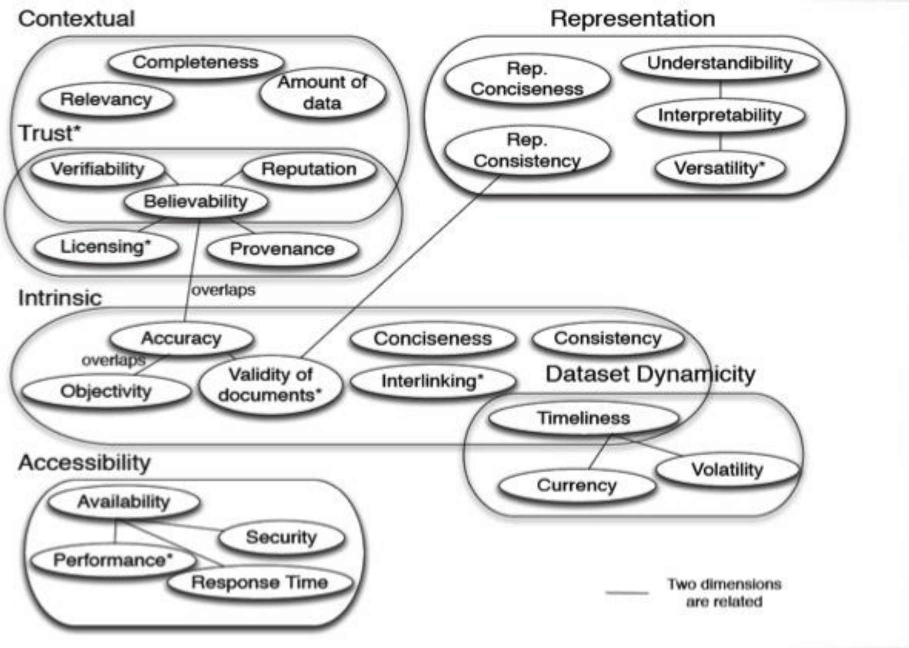
Semantic Web

- Semantic Web (SW) is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation.
- Growth has been exponential.
  - Between 2007 and September 2010, 203 datasets were published containing almost 27 billion RDF triples, of which 395 million were RDF links.
  - By 2011, this had risen to 295 datasets, 31 billion triples and 503 million RDF links (Bizer et al., 2011).



Data Quality

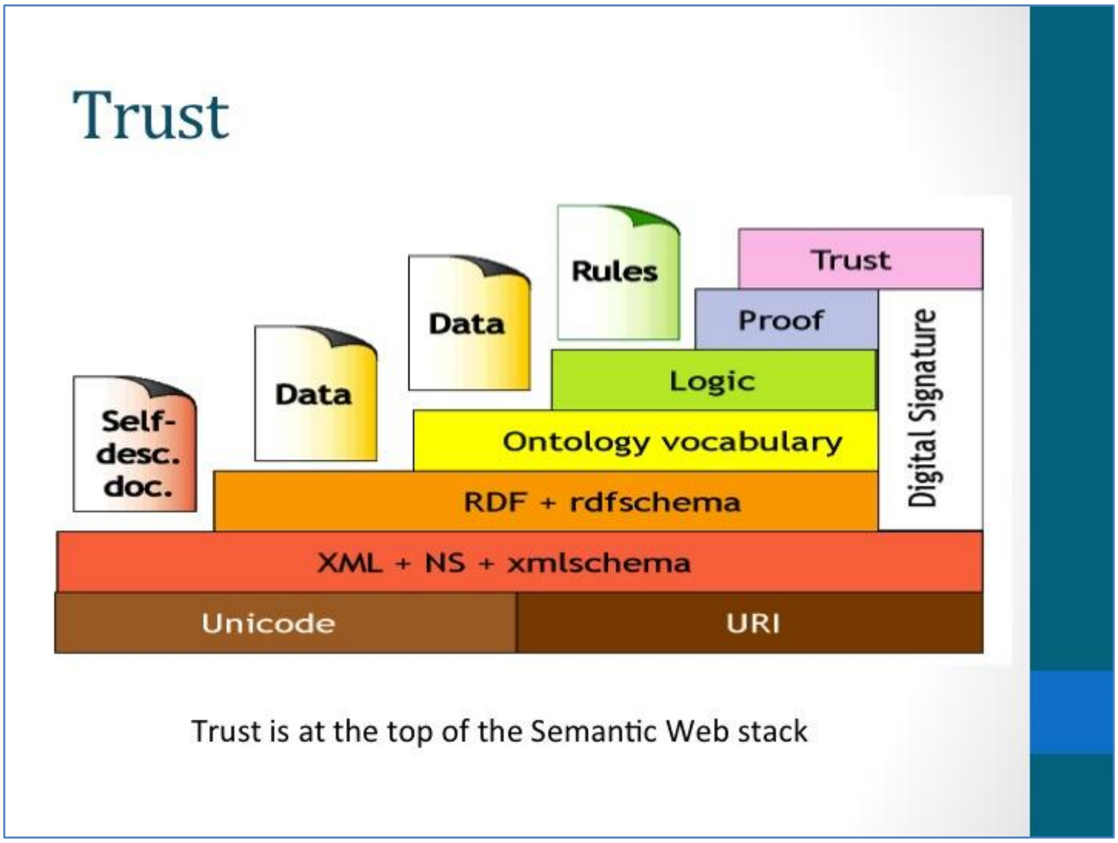
- Data quality is commonly viewed as a multidimensional construct with a popular definition as the "*fitness for use*" (Wang & Strong, 1996).
- Some of the DQ dimensions identified by Zaveri are:
  - Contextual
  - Representation
  - Intrinsic
  - Trust
  - Accessibility
  - Dataset Dynamicity



Linked Data quality dimensions - Zaveri (2012)

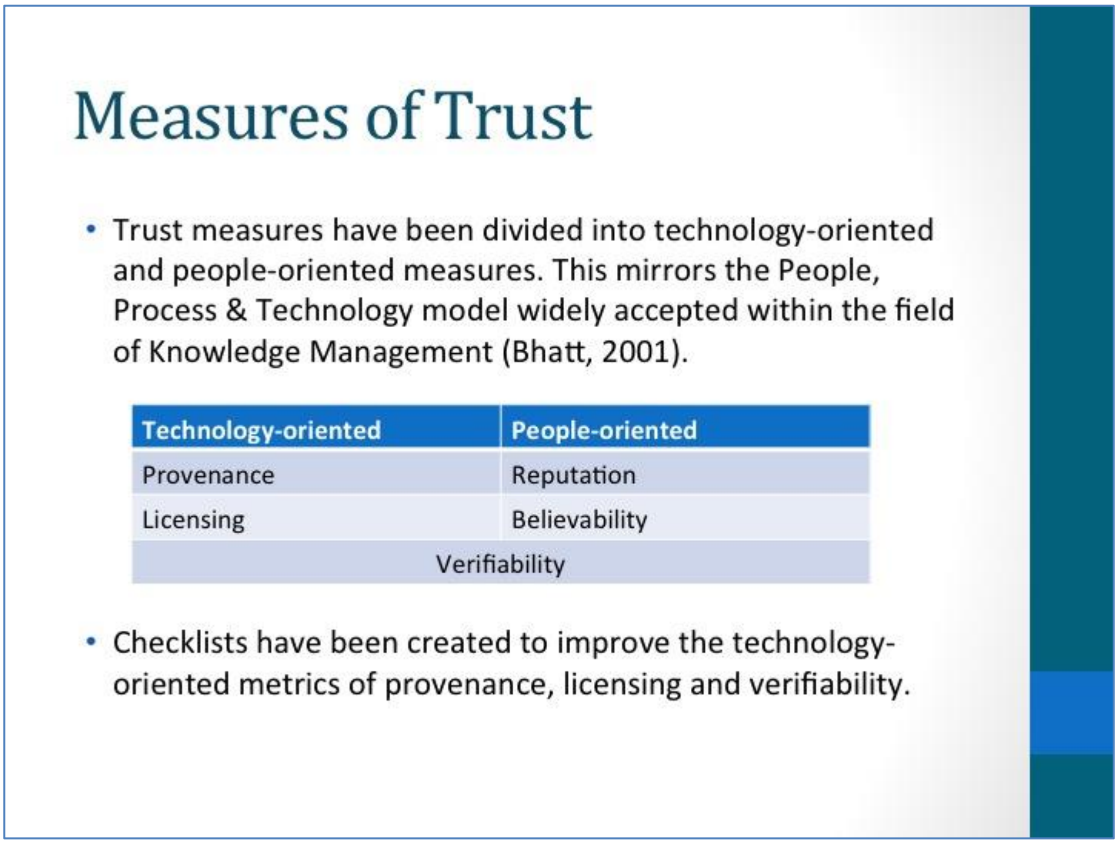
Trust

- *"Trust is the firm belief in the competence of an entity to act dependably, securely, and reliably within a specified context"* (Grandison and Sloman, 2000).
- Trust is an essential component of the initial Semantic Web vision, described by Berners-Lee (2000).
- The Trust layer of the SW stack has remained relatively untouched since the outset of the Semantic Web.



Trust

Trust is at the top of the Semantic Web stack

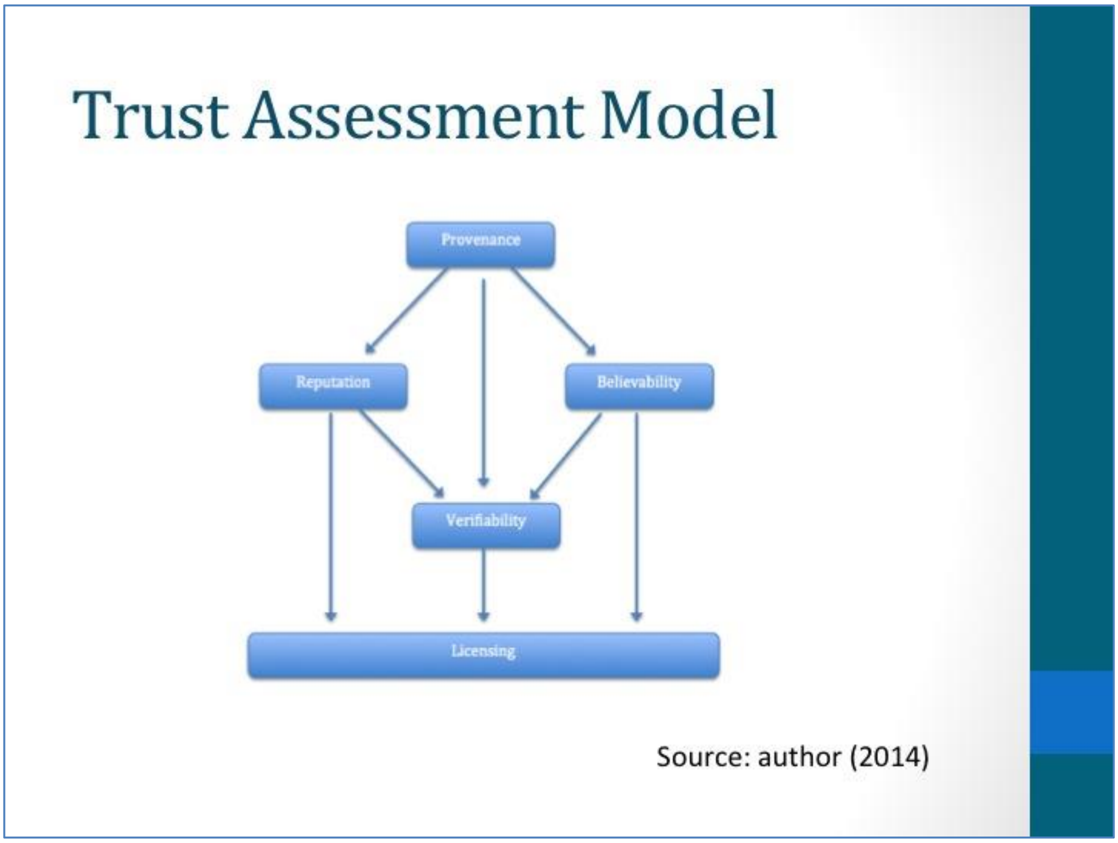


Measures of Trust

- Trust measures have been divided into technology-oriented and people-oriented measures. This mirrors the People, Process & Technology model widely accepted within the field of Knowledge Management (Bhatt, 2001).

Technology-oriented	People-oriented
Provenance	Reputation
Licensing	Believability
Verifiability	

- Checklists have been created to improve the technology-oriented metrics of provenance, licensing and verifiability.



Trust Assessment Model

Provenance

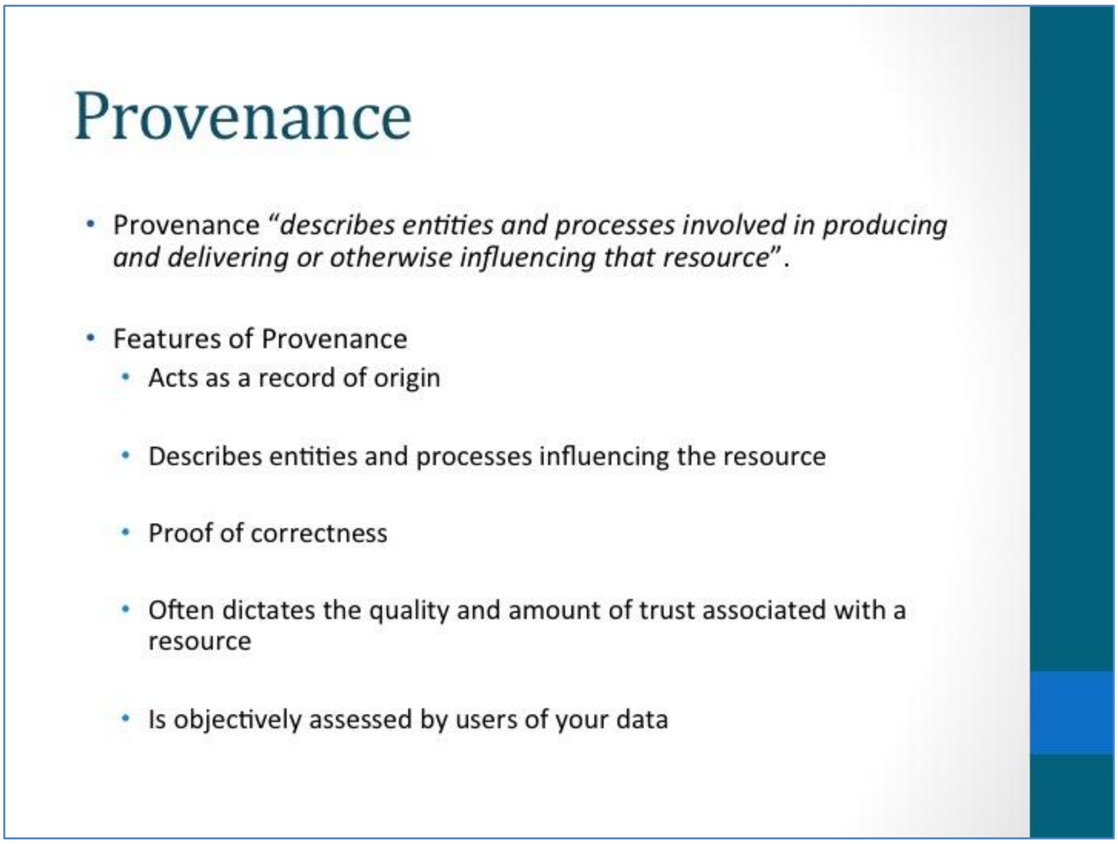
Reputation

Believability

Verifiability

Licensing

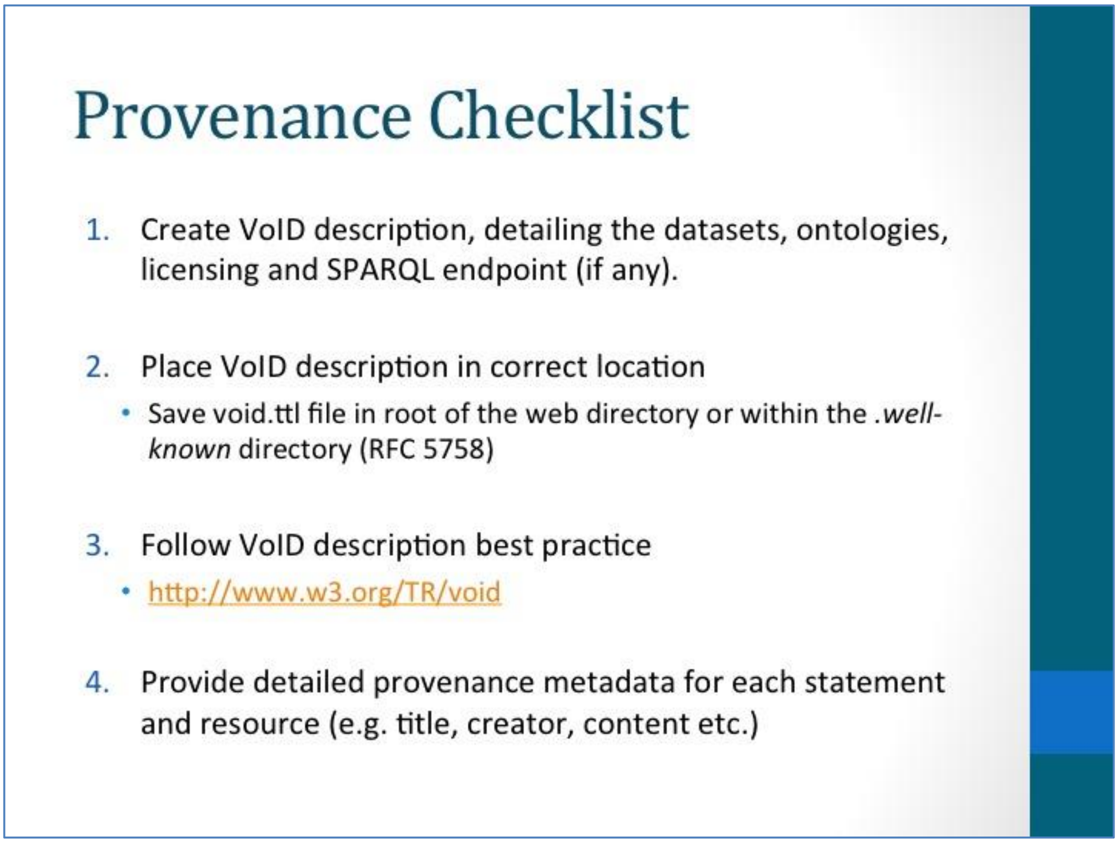
Source: author (2014)



Provenance

- Provenance *"describes entities and processes involved in producing and delivering or otherwise influencing that resource"*.

- Features of Provenance
  - Acts as a record of origin
  - Describes entities and processes influencing the resource
  - Proof of correctness
  - Often dictates the quality and amount of trust associated with a resource
  - Is objectively assessed by users of your data



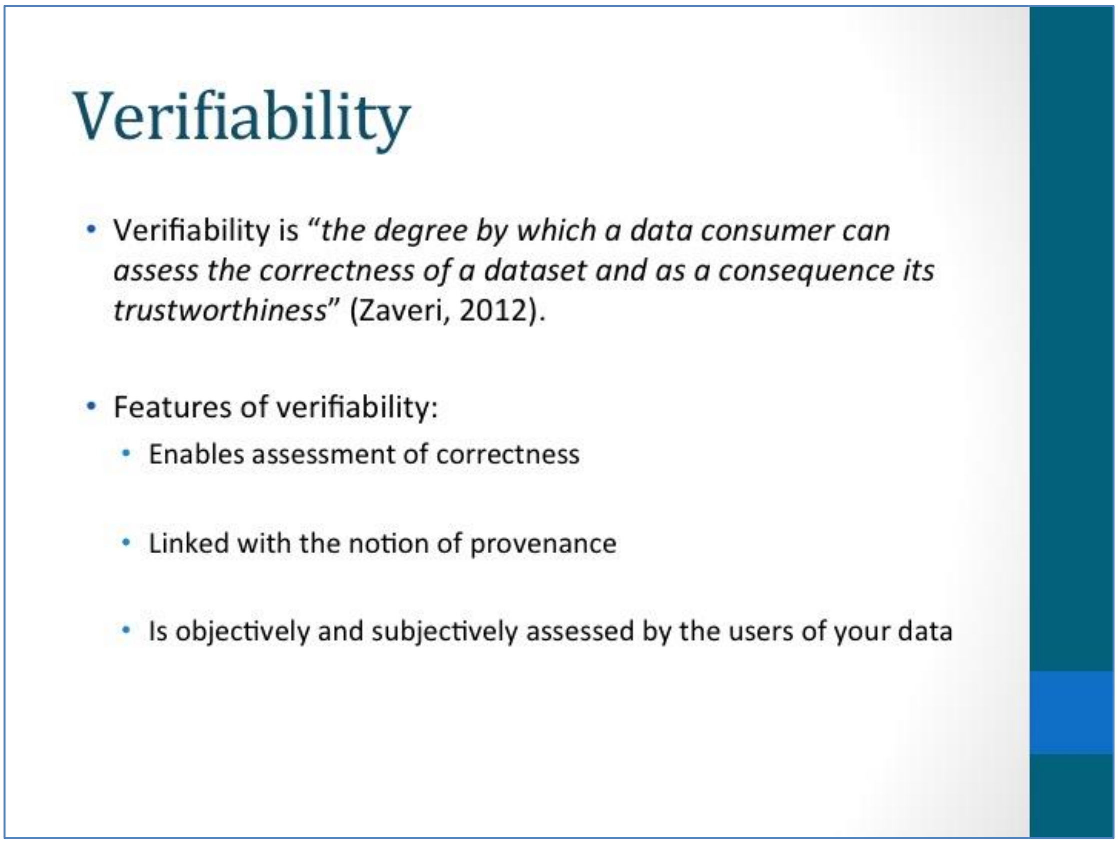
Provenance Checklist

1. Create VoID description, detailing the datasets, ontologies, licensing and SPARQL endpoint (if any).

2. Place VoID description in correct location
   - Save void.ttl file in root of the web directory or within the *.well-known* directory (RFC 5758)

3. Follow VoID description best practice
   - http://www.w3.org/TR/void

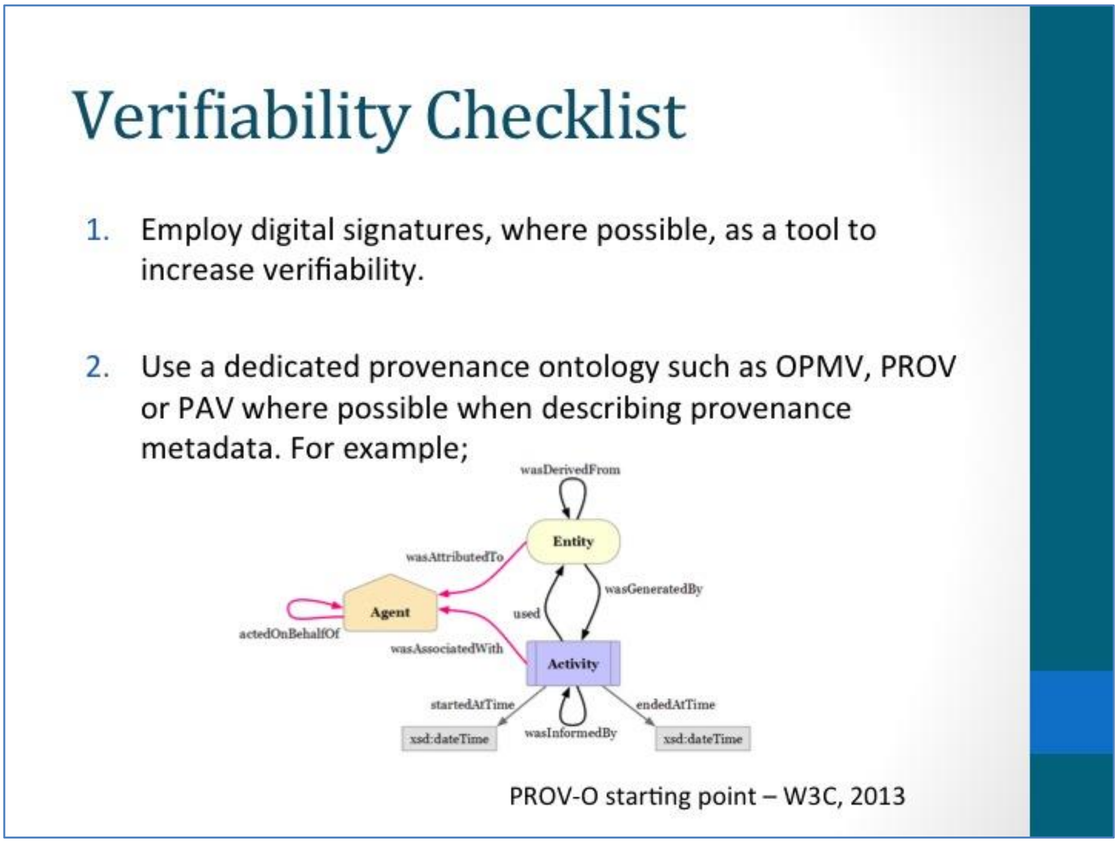
4. Provide detailed provenance metadata for each statement and resource (e.g. title, creator, content etc.)



Verifiability

- Verifiability is "*the degree by which a data consumer can assess the correctness of a dataset and as a consequence its trustworthiness*" (Zaveri, 2012).

- Features of verifiability:
  - Enables assessment of correctness
  - Linked with the notion of provenance
  - Is objectively and subjectively assessed by the users of your data

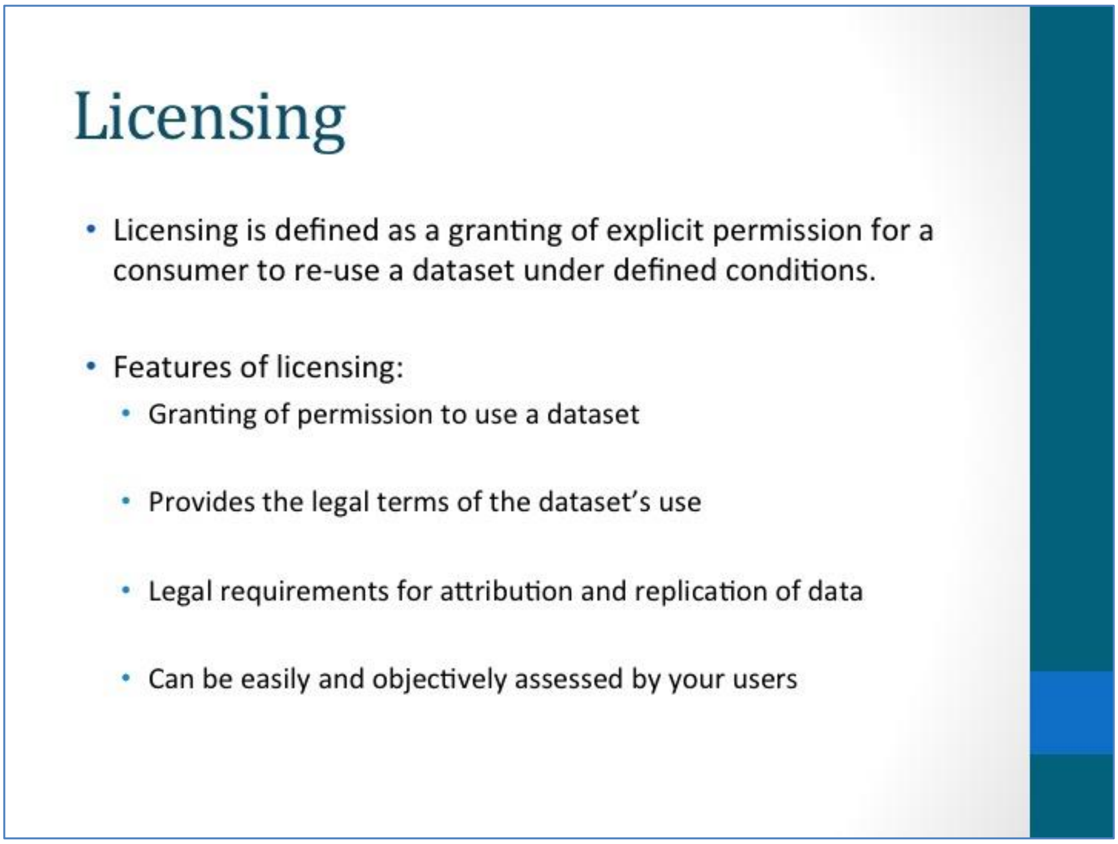


Verifiability Checklist

1. Employ digital signatures, where possible, as a tool to increase verifiability.

2. Use a dedicated provenance ontology such as OPMV, PROV or PAV where possible when describing provenance metadata. For example;

PROV-O starting point – W3C, 2013



Licensing

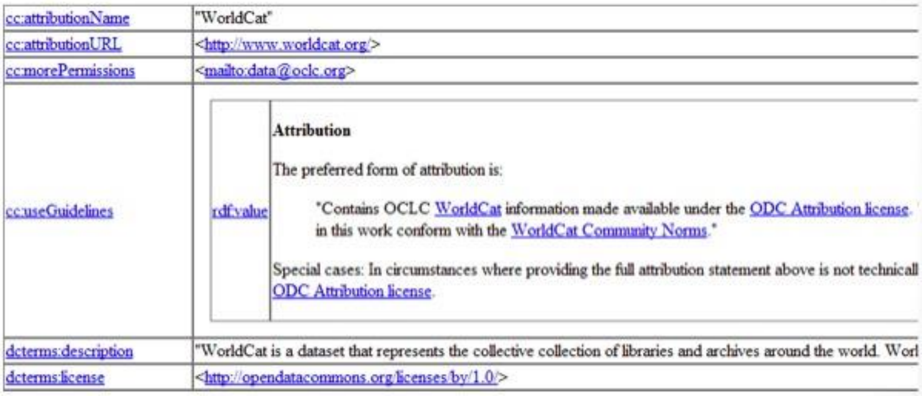
- Licensing is defined as a granting of explicit permission for a consumer to re-use a dataset under defined conditions.

- Features of licensing:
  - Granting of permission to use a dataset
  - Provides the legal terms of the dataset’s use
  - Legal requirements for attribution and replication of data
  - Can be easily and objectively assessed by your users

Licensing Checklist

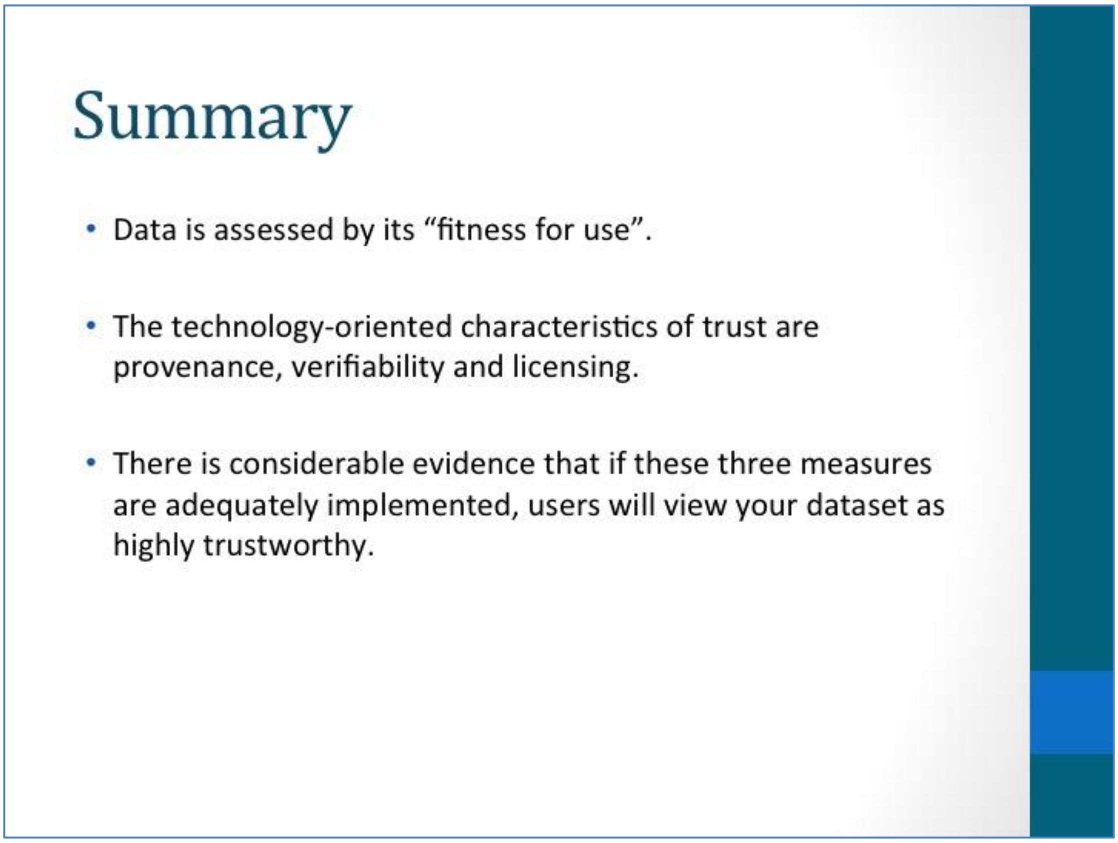
1. Insert machine-readable license metadata in RDF
2. Publish a human-readable license on your website
3. Publish permission metadata in RDF
4. Insert attribution metadata in RDF
5. Utilize open licensing such as CopyLeft or ShareAlike where possible

Licensing Checklist



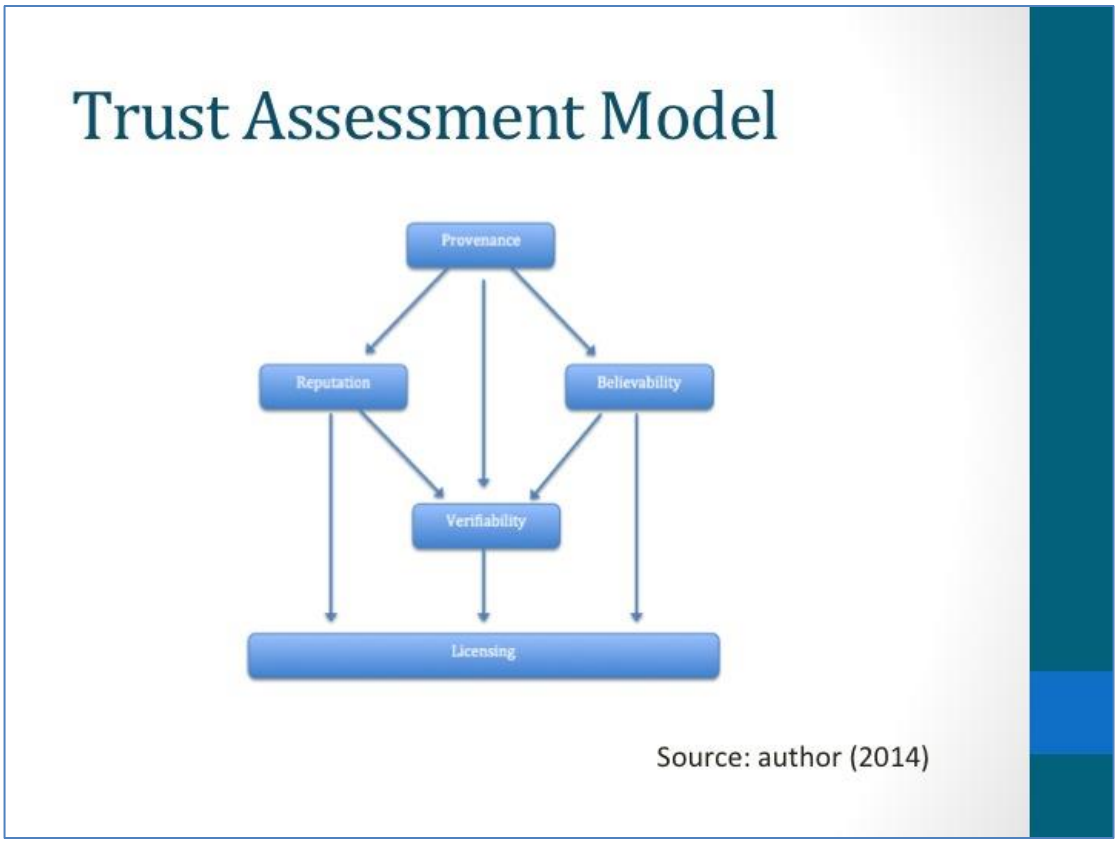
cc:attributionName	"WorldCat"
cc:attributionURL	<http://www.worldcat.org/>
cc:morePermissions	<mailto:data@oclc.org>
cc:useGuidelines	rdf:value | **Attribution** The preferred form of attribution is: "Contains OCLC WorldCat information made available under the ODC Attribution license. in this work conform with the WorldCat Community Norms." Special cases: In circumstances where providing the full attribution statement above is not technicall ODC Attribution license.
dcterms:description	"WorldCat is a dataset that represents the collective collection of libraries and archives around the world. Worl
dcterms:license	<http://opendatacommons.org/licenses/by/1.0/>

http://purl.oclc.org/dataset/WorldCat



Summary

- Data is assessed by its "fitness for use".
- The technology-oriented characteristics of trust are provenance, verifiability and licensing.
- There is considerable evidence that if these three measures are adequately implemented, users will view your dataset as highly trustworthy.



Trust Assessment Model

Source: author (2014)