

University of Southampton Research Repository ePrints Soton

Copyright © and Moral Rights for this thesis are retained by the author and/or other copyright owners. A copy can be downloaded for personal non-commercial research or study, without prior permission or charge. This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the copyright holder/s. The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the copyright holders.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given e.g.

AUTHOR (year of submission) "Full thesis title", University of Southampton, name of the University School or Department, PhD Thesis, pagination

UNIVERSITY OF SOUTHAMPTON

On the Derivation of value from Geospatial Linked Data

by

Jennifer L. Black

A thesis submitted in partial fulfillment for the
degree of Doctor of Engineering

in the

Faculty of Physical and Applied Sciences
Department of Electronics and Computer Science

August 2013

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF PHYSICAL AND APPLIED SCIENCES
DEPARTMENT OF ELECTRONICS AND COMPUTER SCIENCE

Doctor of Engineering

by Jennifer L. Black

[Linked Data \(LD\)](#) is a set of best practices for publishing and connecting structured data on the web. [LD](#) and [Linked Open Data \(LOD\)](#) are often conflated to the point where there is an expectation that [LD](#) will be free and unrestricted. The current research looks at deriving commercial value from [LD](#). When there is both free and paid for data available the issue arises of how users will react to a situation where two or more options are provided. The current research examines the factors that would affect choices made by users, and subsequently created prototypes for users to interact with, in order to understand how consumers reacted to each of the different options. Our examination of commercial providers of [LD](#) uses [Ordnance Survey \(OS\)](#) (the UK national mapping agency) as a case study by studying their requirements for and experiences of publishing [LD](#), and we further extrapolate from this by comparing the [OS](#) to other potential commercial publishers of [LD](#).

Our research looks at the business case for [LD](#) and introduces the concept of [LOD](#) and [Linked Closed Data \(LCD\)](#). We also determine that there are two types of [LD](#) users; non-commercial users and commercial users and as such, two types of use of [LD](#); [LD](#) as a raw commodity and [LD](#) as an application. Our experiments aim to identify the issues users would find whereby [LD](#) is accessed via an application. Our first investigation brought together technical users and users of Geographic Information (GI). With the idea of [LOD](#) and [LCD](#) we asked users what factors would affect their view of data quality. We found 3 different types of buying behaviour on the web. We also found that context actively affected the users decision, i.e. users were willing to pay when the data was to make a professional decision but not for leisure use.

To enable us to observe the behaviour of consumers whilst using data online, we built a working prototype of a [LD](#) application that would enable potential users of the system to experience the data and give us feedback about how they would behave in a [LD](#) environment. This was then extended into a second [LD](#) application to find if the same principles held true if actual capital was involved and they had to make a conscious decision regarding payment. With this in mind we proposed a potential architecture for the consumption of [LD](#) on the web.

We determined potential issues which affect a consumers willingness to pay for data which surround quality factors. This supported our hypothesis that context affects a consumers willingness to pay and that willingness to pay is related to a requirement to reduce search times. We also found that a consumers perception of value and criticality of purpose also affected their willingness to pay.

Finally we outlined an architecture to enable users to use [LD](#) where different scenarios may be involved which may have potential payment restrictions. This work is our contribution to the issue of the business case for [LD](#) on the web and is a starting point for further research regarding the pricing of [LD](#) on the web.

Contents

Declaration of Authorship	ix
Acknowledgements	x
1 Introduction	1
1.1 Motivation	1
1.2 Research Hypotheses	3
1.3 Approach	4
1.4 Thesis Structure	5
2 Architecture of the World Wide Web	6
2.1 Key Technologies	8
2.1.1 URI	9
2.1.2 HTTP	10
2.1.3 HTML	11
2.1.4 XML	11
2.1.4.1 OAuth	12
2.1.5 Limitations of data on the Web	15
2.2 The Semantic Web	16
2.2.1 What is the Semantic Web?	16
2.2.2 The Technologies	16
2.2.2.1 Ontologies	17
2.3 Linked Data	17
2.3.1 What is Linked Data?	17
2.3.2 Datasets in the Linked Data Cloud	22
2.3.3 W3C and Data Publishing	24
2.3.4 The Application of LD	25
2.3.5 The Technologies	25
2.3.5.1 RDF	26
2.3.5.2 RDFS	30
2.3.5.3 SPARQL	31
2.3.5.4 Tools for Linked Data	32
2.3.5.5 APIs	33
2.4 Conclusion	34
3 PSI and Geographic Information	35
3.1 Public Sector Information	35
3.2 PSI Data Providers	38

3.3	PSI and LD	39
3.4	Geographic Information	41
3.4.1	What is GI?	41
3.4.2	Why is it expensive/costly?	41
3.4.3	Key Technologies	42
3.4.3.1	Vector and Raster	42
3.5	GI Data Providers	44
3.5.1	Ordnance Survey	44
3.5.1.1	Ordnance Survey Data Products	45
3.5.2	User Generated GI	46
3.5.2.1	Open Street Map	46
3.5.3	User generated GI vs Traditional Mapping	47
3.6	Conclusion	51
4	Ordnance Survey Use Case	52
4.1	Ordnance Survey Vector Data Products	52
4.2	Ordnance Survey Raster Data Products	55
4.3	Ordnance Survey OpenSpace	56
4.4	OS OpenSpace Pro	56
4.5	Ordnance Survey Linked Data Products	58
4.6	Licenses	62
4.7	OS MasterMap Pricing	63
4.8	Pricing changes since 2008	66
4.9	Conclusion	67
5	The Business of Linked Data	68
5.1	Excludable vs Non Excludable, Rivalrous and Non Rivalrous	69
5.2	Public, Private, Club and Common Goods	70
5.3	Information Goods	71
5.3.1	Information Value and the Value added by Linked Data	72
5.3.2	Affordances	74
5.3.3	Information Quality	76
5.3.4	The Economics of Linked Data	77
5.4	Revenue Models	78
5.4.1	Revenue Models for Digital Goods	80
5.4.1.1	The Advertising Model	80
5.4.1.2	Sponsorship	81
5.4.1.3	The Transaction Fee Model - Micro-payments	81
5.4.1.4	Volume	82
5.4.1.5	The Subscription Model	82
5.4.1.6	Free	82
5.4.1.7	The Freemium Model	85
5.4.2	Willingness to Pay for Information Goods (Online)	87
5.4.3	Trust	88
5.5	Digital Content Industries	90
5.5.1	Newspaper Industry	90
5.5.2	Music Industry	93

5.5.3	Software	96
5.6	Experiences With Linked Data	98
5.7	Licensing	101
5.7.1	Content licenses	102
5.7.2	Creative Commons Licensing	103
5.7.3	Click-Use Licenses	105
5.7.4	OS OpenData License	106
5.7.5	Open Government License	106
5.7.6	Public Service Mapping Agreement	107
5.7.7	Comparison of the old Click-Use licenses versus Creative Commons Licenses	107
5.7.8	Open Data Commons Licenses	107
5.7.9	Open Street Map Licensing	108
5.7.10	Licenses used on Data from the Linked Data Cloud	108
5.8	Summary	109
6	Requirements Elicitation	111
6.1	Ordnance Survey Open Space Investigation	111
6.1.1	Results	112
6.1.2	Discussion	114
6.2	Terra Future - Forging Links Seminar	116
6.2.1	The Approach	116
6.2.2	Conclusion	117
6.3	Investigation - Information Quality Criteria Questionnaire	118
6.3.1	Experimental Design and Methodology	118
6.4	Participants	120
6.5	Results and Statistical Analysis	120
6.6	Discussion	121
6.7	Conclusion	122
6.8	Summary of Research Contributions	123
7	Linked Data Investigations	126
7.1	Linked Data Experiment 1	126
7.1.1	Experimental Design and Methodology	126
7.1.2	Participant Interaction with the Study	129
7.1.3	Participants	129
7.1.4	Results	129
7.1.5	Discussion	131
7.2	Linked Data Investigation 2	133
7.2.1	Experimental Design and Methodology	134
7.2.2	Participants	135
7.2.3	Results of Linked Data Simulation	135
7.2.4	Results of the Post Simulation Questionnaire	137
7.2.5	Participant discussion	137
7.2.6	Discussion and Conclusion	139
8	An Ordnance Survey Case Study Using Linked Geospatial Data	143

8.1	Introduction	143
8.2	Ordnance Survey Case Study	144
8.2.1	Background	144
8.3	Analysis and Requirements	152
8.4	Actor Description	153
8.4.1	Context - How does Addressing apply to our suggested architecture?	155
8.4.2	SPRITE	157
8.4.3	Application - Why use Linked Data for this?	158
8.5	Technologies	159
8.5.1	Data Format	159
8.5.2	Query Language	159
8.5.3	Authentication	160
8.5.3.1	API Keys	160
8.5.3.2	OpenID	160
8.5.4	Licensing	161
8.6	Protocols	161
8.7	Other Considerations	165
8.7.1	Trust and Reputation	165
8.7.2	Willingness to pay for data	166
8.8	Conclusion	166
9	Conclusions	168
9.1	User Requirements	169
9.2	User Interaction with Linked Data	170
9.3	Architecture for Linked Geospatial Data	170
9.4	Future Work and Research Directions	171
9.4.1	Implementation of the suggested technical architecture	171
9.4.2	Usability and HCI	171
9.4.3	Return on Investment	172
9.4.4	Test for speed of data discovery	172
9.4.5	Pricing	172
9.4.6	New Classes of Data	173
9.4.7	Awareness of the Potential of LD	173
	Appendix A Terra Future Invite	174
	Appendix B Parking Experiment	177
	Appendix C Information Quality Questionnaire	182
	Appendix D Information Quality Questionnaire Significance Results	188
	Appendix E Linked Data Study Screen Shots	190
	Appendix F Linked Data Questionnaire - Post Study	194
	Bibliography	197

List of Figures

2.1	An example of an early web page taken from the W3C from 1992 http://www.w3.org/History/1992/hypertext/hypertext/WWW/TheProject.html	7
2.2	OAuth Authentication Flow Diagram from http://oauth.googlecode.com/svn/spec/core/1.0/	11
2.3	Linking Open Data cloud diagram, as of September 2011, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/	20
2.4	Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/	21
2.5	A graph to show the data model of the example RDF	29
3.1	An area and its approximation by a polygon taken from Longley et al. (2005)	43
3.2	Raster Representation taken from Longley et al. (2005)	43
3.3	Comparison of OSM vs OS - This image show a snapshot of an OS map	49
3.4	Comparison of OSM vs OS - This image shows a snapshot of an OSM map and illustrates the same area as shown on the OS map above	50
6.1	125
7.1	Data type selection.	132
7.2	Graph to show proportion of users who chose each payment type	136
8.1	Illustration of the Addressing Example	151
8.2	Actors directly involved in the framework and the transactions between them	154
8.3	Stage one through the framework	156
8.4	Possible outcomes from SPARQL Query	164

List of Tables

2.1	Table to show the successful and further action required http status codes	10
2.2	Table to show the client and server http codes	11
2.3	Key Linked Datasets from the OS Cloud in Figure 2.2	23
2.4	Table to show the triples in the RDF example	28
4.1	OS MasterMap Products	53
4.2	OS Vector Data Products	54
4.3	OS Raster Data Products	55
4.4	OpenSpace Pro Pricing and Terms	57
4.5	Linked Datasets which use OS Data	61
4.6	Product Pricing	64
4.7	OnDemand Pricing	65
4.8	AddressPoint Pricing	65
4.9	Codepoint Pricing	66
5.1	Summary of Types of Goods	71
5.2	Categories and Dimensions of Data Quality - Adapted from Strong (1996)	77
5.3	Comparison of Google Maps API vs Google Maps API for Business taken from https://developers.google.com/maps/licensing	83
5.4	Comparison of The Different 'Free' Models	83
5.5	Comparison of Click use licenses versus Creative Commons Licenses	107
5.6	Key Linked Datasets and their Licenses	109
6.1	Classification Of Open Space API Users	113
6.2	Classification Of Open Space API Use	113
6.3	Results of Information Quality Questionnaire	120
6.4	Results of Statistical Analysis	121
6.5	Significant Association of Results	121
6.6	Preferred Quality Criteria	122
7.1	Proportion of participants in each category	130
7.2	Proportion of participants who chose each option	130
8.1	Status Codes and Their Uses	162

Declaration of Authorship

I, Jennifer Louise Black, declare that the thesis entitled on the Derivation of Value from Geospatial Linked Data and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- parts of this work have been published in a conference paper

Signed:

Date:

Acknowledgements

I would like to express my very great appreciation to my parents, Rachel and Andrew and my sister Claire for their continued patience and understanding throughout the past four years, who without their continuing reassurance and support I wouldn't have made it. To my friends who have stood by me through the ups and downs! With special thanks to Phil, Davina (and the Louissons!) and Rachel for being so supportive through all my quandaries!

I would like to offer thanks to my friends and colleagues in ECS, especially Kate and Marcus for their ongoing support and guidance throughout the EngD.

I thank my academic supervisors Professor Nigel Shadbolt and Dr Nicholas Gibbins for sharing their expertise in my chosen research area.

I wish to acknowledge the help and support provided by everyone at Ordnance Survey, my industrial sponsor. I would especially like to thank Glen Hart and Dr John Goodwin who have provided strong support throughout and Anne for encouraging me, especially through the last few months!

To my family, for their unreserved support and understanding

Acronyms

API Application Programming Interface. [27](#), [67](#), [119](#)

CC Creative Commons. [79–81](#), [83](#), [84](#)

CIE Common Information Environment. [79](#)

CSS Cascading Style Sheet. [5](#)

GB Great Britain. [40–42](#)

GI Geographic Information. [i](#), [1](#), [3](#), [29](#), [30](#), [32](#), [33](#), [35–39](#), [45–47](#), [50–52](#), [55](#), [56](#), [58](#), [71](#), [75](#), [79](#), [84](#), [91](#), [92](#), [106](#), [116](#), [130](#), [131](#), [135](#)

GIS Geographic Information System. [37](#)

GPL General Public License. [47](#), [83](#)

GPS Global Positioning System. [40](#)

HTML Hyper Text Markup Language. [4](#), [6](#), [9](#), [10](#), [108](#)

HTTP Hyper Text Transfer Protocol. [4](#), [6–8](#)

IP Intellectual Property. [77](#)

IQ Information Quality. [93](#)

JSON JavaScript Object Notation. [24](#), [28](#)

LCD Linked Closed Data. [18](#), [35](#), [54](#), [60](#), [61](#), [67](#), [69](#), [77](#), [79](#), [90](#), [98](#), [100](#), [116](#), [117](#), [122](#), [126](#), [131](#), [132](#), [134](#)

LD Linked Data. [i](#), [1–4](#), [6](#), [8](#), [11](#), [14](#), [15](#), [18](#), [19](#), [26–36](#), [38](#), [39](#), [44–46](#), [50](#), [52–54](#), [56–79](#), [82–84](#), [86](#), [89–92](#), [97](#), [98](#), [100–103](#), [107](#), [108](#), [113](#), [116](#), [117](#), [121](#), [122](#), [124–126](#), [130–135](#), [139](#)

LOD Linked Open Data. [3](#), [18](#), [35](#), [54](#), [60](#), [77](#), [79](#), [98](#), [116](#), [117](#), [122](#), [126](#), [131](#), [132](#), [134](#)

- ODbL** Open Database License. 47, 84
- ODC** Open Data Commons. 47, 83, 84
- OPSI** Office of Public Sector Information. 82, 83
- OS** Ordnance Survey. i, 3, 13, 18, 19, 31, 33, 35–41, 44–48, 50, 52, 53, 55, 56, 58, 61, 64, 79, 84, 86, 90, 92, 98, 100, 117, 125, 130, 131, 133, 135
- OSAPR** Ordnance Survey Address Point Reference. 41
- OSM** Open Street Map. 37, 46–48, 50, 114
- OWL** Web Ontology Language. 1, 6, 13
- PDDL** Public Domain Dedication and License. 84
- PSGI** Public Sector Geographic Information. 35
- PSI** Public Sector Information. 1–3, 29–35, 50, 54, 84, 116, 130, 132, 135
- PSIH** Public Sector Information Holder. 31, 32
- PSMA** Public Sector Mapping Agreement. 83, 89, 90
- RDF** Resource Description Framework. 1, 6, 10, 13, 14, 18–21, 24–27, 46, 68, 80, 108, 118
- RDFS** Resource Description Framework Schema. 19, 24, 71
- RRI** Road Routing Information. 41
- SEO** Search Engine Optimisation. 19
- SP** Service Provider. 120
- SPARQL** Sparql Protocol and RDF Query Language. 1, 18, 25, 26, 80
- SW** Semantic Web. i, 1, 3, 4, 6, 11–14, 20, 26, 27, 29, 30, 52, 58, 86, 116, 130
- SWT** Semantic Web Technologies. 13, 14, 29, 34–36, 80
- TOID** Topographic Identifier. 40, 44
- UK** United Kingdom. 33, 36, 39, 50
- URI** Uniform Resource Identifier. 4, 6, 7, 11, 12, 20, 28, 45, 58, 65
- URL** Uniform Resource Location. 4, 7
- USA** United States of America. 34, 36, 50

WWW World Wide Web. [6](#)

XML Extensible Markup Language. [4](#), [6](#), [9](#), [10](#), [24](#), [80](#)

Chapter 1

Introduction

In this document, we look at three research areas: The [Semantic Web \(SW\)](#) and specifically [LD](#), [Geographic Information \(GI\)](#) and the business case for linked [GI](#).

The W3C's intent in developing the [SW](#) and introducing a range of [SW](#) technologies including [Resource Description Framework \(RDF\)](#), [Sparql Protocol and RDF Query Language \(SPARQL\)](#) and [Web Ontology Language \(OWL\)](#) has been to extend the current web of documents to encompass structured data in addition to human-readable text.

However, the scope of W3C's work has been almost entirely technical, rather than social or economic: these areas need to be investigated and understood if the [SW](#) is to realise its full potential. With a new culture of data publishing on the web, we investigate the opportunities to derive value from linked data that is explicitly geographical in nature and we consider three themes within this: the technology used to create links, [GI](#) and the business and economics of a Linked Data Web. More specifically we will investigate the opportunities to derive value from [LD](#) that is explicitly geographical in nature.

1.1 Motivation

The Open Data White Paper ([The Stationery Office, 2012](#)) explores how the government aims to unlock the potential of Open Data. Now that this movement has gained momentum we are aware of the need to investigate the factors which will contribute to a world where data is a raw commodity.

There is a great deal of research being carried out looking at the value of [Public Sector Information \(PSI\)](#) and over the past two years the availability of more [PSI](#) has meant that consumers of data are now able to begin to understand the possibilities for open data.

It is not just the inherent value from **PSI** alone, the recent emergence of the Open Data Institute¹ has demonstrated the potential for public and private sectors to merge, thus producing value. Two key concerns for organisations that hold large quantities of information are determining the value of their data and the establishment of the most suitable and lucrative way of exposing such data to customers. Current models such as subscription and advertising are suitable for businesses where large amounts of data are purchased as a whole (Novak and Hoffman, 2001). However new models such as pay as you go and micro-payments will need to be developed to keep up with the development of new technologies to enable greater flexibility for users to purchase data on the web. Currently users are subject to using free data with little knowledge over the reliability of the content. For example, Wikipedia content can be edited and created by anyone and thus highlights issues of reliability. This problem also exists for the organisation of control over how its data is used and reused (Mitchell and Wilson, 2012). There is also the problem of the transition between free data and paid data from one organisation. **OS** for example now has free data and data which can be paid for, but the transition between free content and paid content needs to be seamless to ensure continuity. This is also a problem for users who are unsure of the benefit they will get from paid for data. There is also a factor of perceived value of the data that will affect the price that the user is willing to pay for the data. Without a way of showing the difference between free data and paid data it is difficult for a user or consumer to make a decision regarding the transition from free to paid. This research aims to explore the factors which affect the business of **LD** on the web, in particular we look at the factors which affect users in situations where free and paid content are available and how these factors affect purchase decisions.

We mention users throughout the thesis. Users for this purpose are potential users of **LD** on the web or people who purchase information on the web. We note that these ‘users’ may not be aware that they are using data that is linked. They will therefore not be developers and have little or no specific knowledge of the technology or how it works. In this thesis we refer to users as lay people who use the web and who will continue to do so, despite any changes to the underlying architecture. This architecture is the way in which the data is structured on the web behind the webpages which they interact.

The web has dramatically changed the way in which we use information and **LD** will affect the way data is consumed on the web. This may have repercussions for content providers such as how data is made available and sold in different formats. Much effort has gone into the technologies and tools used to create and manipulate data, but currently no suggestions are available to address the problem of the consumption of data from the general public with little technical knowledge (Bizer, 2009). For example, **OS** has built its business around supplying products (maps) as a whole. It has large consumers such as insurance companies or local authorities and they have the funds to be

¹<http://www.theodi.org>

able to make purchases of whole products. To date it has been possible to sell the data in packages. Users may not be able to initially say which parts of a data package they need and may require some time to establish what they need. This is a challenge for data providers to be able to offer a subsection of a product with a pay-as-you-go access option to certain parts of the product. For example, a leisure user with no or very little budget for access to data executes a search for points of interest in a particular city, the search may return a number of options but the user may only require one of those options and would not necessarily want to purchase them all. Criticality of purpose is another factor which we are aware of with different consumers from various sectors. For instance large insurance companies will not only have readily available funding to make purchases but could view the purchase of data more critical than a member of the public looking for points of interest in a local village.

The technical issues surrounding LD lie with being able to search, distribute and secure data in small packages rather than provide users with complete datasets. As we have explained above, users may want to scan a large dataset and establish specifically a small amount of data for their end use. The issue for data providers is how to partition data and sell it or give it away for free, which leads us to the importance of a smooth transition between free and premium products to ensure that if a user has been interacting with a free product, they are able to purchase the premium product and transfer without any issue. We look into this in Chapter 8, where we look at a case study for Linked Geospatial Data. We propose that users may be able to search a whole dataset and make a selection for data, which they are then able to make a micropayment to receive or are able to obtain it free.

The web has evolved into a space where consumers are able to readily find data and information for free with the cost of some effort spent searching. With LD, applications can be built around real data which can significantly reduce costs associated with accessing data. We are concerned with the users interaction with LD in this thesis and focus on their interaction with data via the technology rather than a directly manipulating the data. With this in mind we are concerned with essentially a new product on the web, how we might be able to charge for data which was once potentially not available or available for free. We want to consider different scenarios and how the user may react under different conditions.

1.2 Research Hypotheses

Based on the issues we outlined in the earlier part of this introduction, our experimental hypotheses are as follows:

h_1 – Does criticality of purpose affect a user’s decision whether or not to pay for pre-

mium LD when free alternatives are available?

We are concerned with understanding how the intended use of data will affect a decision the consumer makes in regards to making a purchase. We are aware that there are a number of free alternatives to data on the web, and therefore the consumers will have a variety of data to choose from. We wish to find whether this is a contributory factor in the purchase of data on the web.

h_2 – Is willingness to pay for premium data positively influenced by consumers' perceptions of its value?

The same datasets may be used by a number of different users and we would like to consider whether different types of users rate the same data differently depending upon the end use it will be put to.

1.3 Approach

In order to test the hypotheses outlined above, we propose to find answers for the following questions

1. Which factors affect a user's perception of information quality?
2. Which of these factors are specific to LD?
3. What proportion of users choose free data over data which must be paid for?

The main contributions made in this thesis are:

1. Identification of the different types of consumers of LD
2. Identification of the factors which affect a users willingness to pay.
3. Determine the quality factors of data which affect a users choice of data
4. A technical framework for the consumption of LD

We state that the work carried out in this thesis has limitations which should be considered. The key areas which require further development are outlined in the final chapter. We have begun by introducing the key literature surrounding the topics and find this a valuable resource for introducing the key points and issues surrounding LD. We demonstrate the use of LD through discussions and 2 prototypes of a LD model. However we

do not venture into the specific pricing of the data or implement the architecture which has been defined. We suggest that this is a further development for future work outside of this research.

1.4 Thesis Structure

Chapter 2 covers the technical side of this research. We begin by outlining the architecture of the World Wide Web. We describe the web as we are accustomed to using now and the developments which will be noticed with [LD](#) and [SW](#).

Chapter 3 introduces [PSI](#) and its importance to [LD](#) and more specifically we describe [GI](#). This chapter also describes in more detail [OS](#) which is the case study for thesis.

Chapter 4 provides a background to the business of [OS](#) and demonstrates the potential for the application of [LD](#). This chapter provides a use case for [LD](#) and looks at the different types of products provided by [OS](#). It also looks at the pricing and licensing of these products to enable a view of the possible issues which arise.

Chapter 5 covers the important background material regarding the issues surrounding the business of [LOD](#). We address the potential to add value to linked data and how [LD](#) fits into the [SW](#).

The remaining chapters of this thesis detail our contribution which aims to answer the hypotheses. Chapter 6 details the requirements elicitation for the [LD](#) experiments we carry out in chapter 6. Chapter 5 includes details on the informal engagement and qualitative consultation with the community, which included the TerraFuture workshop and an investigation into users of the [OS](#) OpenSpace platform. This chapter introduces the need for encouragement to use new technologies and describes an experiment designed to bring together [GI](#) and linked data communities to discuss the possibilities. This chapter also includes details of a questionnaire which was carried out in order to discover the criteria which may affect a consumers decision to pay for data.

Chapter 7 details an experiment which was carried out to discover if users are prepared to pay for premium data even if there is a free option and what are the factors which affect their decision to move from free to premium or premium to free.

The empirical research we carry out in chapters 6 and 7 give us the detail required to outline a case study using [OS](#) Linked Geospatial data in chapter 8.

Chapter 9 concludes this thesis by reviewing the work and proposing ideas for further research.

Chapter 2

Architecture of the World Wide Web

In this chapter we begin by setting the background for this research by exploring the architecture of the web as we experience it now, including the technologies and developments noticed over time. We then introduce the concept of [LD](#). We highlight the key technologies and detail the uses and benefits which support the business case for [LD](#). We also detail the [SW](#) and explain its extension from the web which we have become familiar with using from its early stages of development.

In order to understand the technology of [LD](#) in more detail it is important that we look at the history of the web from its introduction through to how we use it today. The web was intended to be a network of information resources and was originally used as a medium for retrieving information.

The W3C outlines three mechanisms used to make resources on the web available. These are [Uniform Resource Identifier \(URI\)](#)'s, [Hyper Text Transfer Protocol \(HTTP\)](#) and [Hyper Text Markup Language \(HTML\)](#).¹

Everything on the web has an address or more specifically an identifier which is encoded by a [URI](#) and given an address at a [Uniform Resource Location \(URL\)](#). We explain [URI](#)'s in more detail in the next section. Pages are written in [HTML](#) and hyperlinks are used to link to other pages ([Chakrabarti et al., 1998](#)). In the past anyone could upload pages about their business and individuals were able to create pages about hobbies and personal projects. These pages were primarily just text and pictures, displayed with little formatting. Later some pages began using [Extensible Markup Language \(XML\)](#) to structure and format large sets of information. [Figure 2.1](#) demonstrates a simple web page from the beginning of the web.² Notice this page shows no text formatting, no

¹<http://www.w3.org/TR/html4/intro/intro.html>

²<http://www.w3.org/History/19921103-hypertext/hypertext/WWW/TheProject.html>

background colour or flash. More importantly when we look at the source for this page, it contains:-

1. **A header** *The World Wide Web Project*
2. **Title** *World Wide Web* and
3. **Body** *The WorldWideWeb is a wide-area...*

Web pages now contain formatting of text, background colours, links, videos and sounds. We also note that early web pages had no social input, i.e. comment pages and uploading of user content.

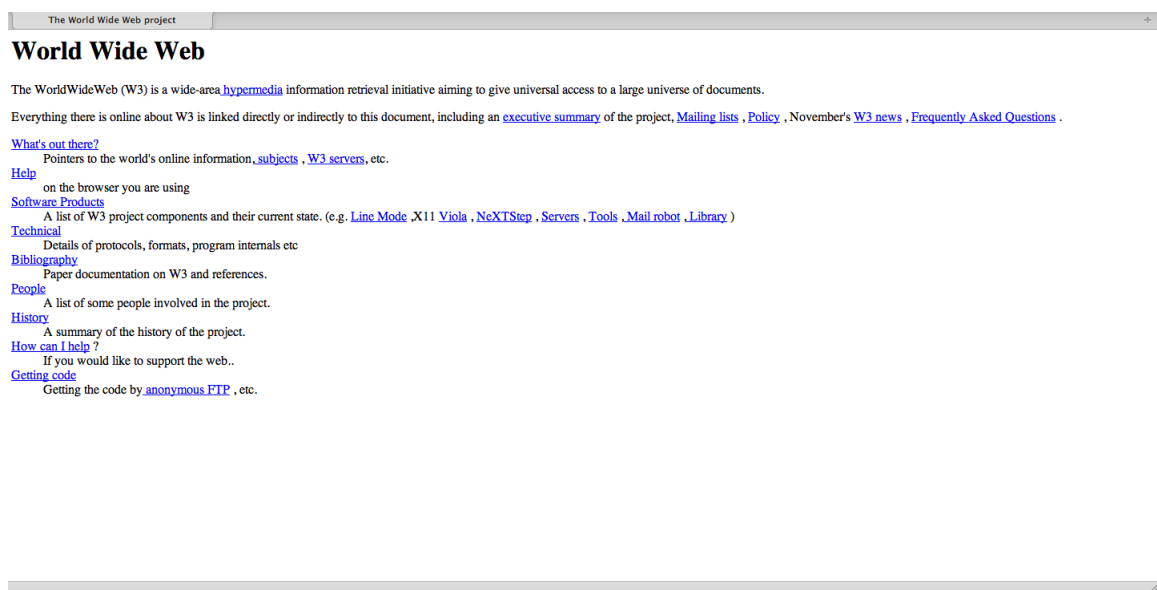


FIGURE 2.1: An example of an early web page taken from the W3C from 1992
<http://www.w3.org/History/19921103-hypertext/hypertext/WWW/TheProject.html>

Web pages now contain several other elements such as Adobe's Flash which is a multi-media platform used to add animation, video and interactivity to web pages. [Cascading Style Sheet \(CSS\)](#)'s³ are also used as a mechanism to add style to pages in the form of fonts, colours and spacing. We also see that users are able to add their own comment, pictures and content and suggest links to similar sites. This trend has become popular for various sites, for example pages displaying the news and social websites such as Facebook[©] and Twitter[©]. Web 2.0 or the social web is the same but we notice that in web 2.0 there are better tools available for people who are not the author of the page to add comment and content to pages. As a result of users being able to contribute more to the web by sharing content and comments, the need to observe the quality of

³<http://www.w3.org/Style/CSS/Overview.en.html>

this rapidly growing information resource is essential. The web as we use it has no real control of what is published and if individuals or organisations are to add more data to this, we need to ensure there is a way of making decisions about the quality and trustworthiness of the additional resources we contribute. Information quality, value and trust is a major consideration which we will explore in more detail in chapter 4.

Despite this evolution into the formatting and interaction with webpages, there still lies an area of the web which has not been utilised - there is no formal way of creating links or reasoning between the content on disparate web pages. We are able to create links to pages but it is the utilisation of the links between the content which we explore using [SW](#) technologies and [LD](#). The [SW](#), seeks to overcome this problem by creating a web of data ([LD](#)) rather than a web of documents. [SW](#) data is stored in pages in a structured format using standards such as [RDF](#) ([Miller and Manola, 2004](#)) and [OWL](#) ([Hitzler et al., 2009](#)) which are set out by the W3Ca.⁴ Content is described using ontologies and knowledge is inferred through reasoners applied to these ontologies. Data can be copied, adapted and re-used. Whereas, in formats such as CSV or XML, the data can be changed and the meaning is lost as it is passed around. [SW](#) standards also enable the publishers of [LD](#) to remain in control of their data.

Web pages on the [SW](#) are structured in [RDF](#), as opposed to [HTML](#), which makes the content of these pages accessible to machines. Typically [HTML](#) only enables machines to read the formatting of the data, but encoding data as [RDF](#) enables further reasoning to be performed. Computers or machines are able to reason with parts of the page specific to search requirements of users and can then be displayed in a web browser in a human readable form ([Halb and Raimond, 2008](#)). This structure enables a computer to make decisions and reason with data without a prior knowledge of the subject, which was previously not possible due to the incompatible formats of different data stored on the web ([Shadbolt et al., 2006](#)).

2.1 Key Technologies

In this section we outline in more detail the key technologies of the web. We detail [URIs](#) which are used to identify names or resources. We describe [HTML](#) the publishing language for the [World Wide Web \(WWW\)](#) and [HTTP](#) which is the protocol used for interaction between webpages. We also explain [XML](#), a set of rules used to encode documents in a machine readable format.⁵

⁴<http://www.w3.org/standards/semanticweb/data>

⁵<http://www.w3.org/TR/REC-xml/>

2.1.1 URI

A [URI](#) is a way of linking documents and resources together. These can include images and objects.⁶ A URI is made up of string of characters and is used to identify resources individually ([Masinter et al., 2006](#)). There has been some confusion between the existence of URIs, URNs and URLs. A URN in general refers to an item's identity (for example a book's ISBN number) whereas a [URL](#) provides a method which can be viewed as the pathname or address.

In order for objects to be retrieved on the web we need a way of identifying and retrieving them and the use of [URIs](#) solves this problem. The act of retrieving this information using a URI is know as *dereferencing* that URI.⁷ When a user clicks on a URI they are directed to the URL which tells the computer where to find the item.

An example of a http URI could be:

```
http://exampleuri.co.uk/example/example.html
```

- The http in the [URI](#) refers to the [HTTP](#) which is the communication protocol used for the request, this tells the server how to send the information. We explain [HTTP](#) in more detail below.
- 'exampleuri' part of the [URI](#) specifies a computer on the internet which is storing the page (the server)
- '/example/' outlines the name of the directory stored on the server or host, this is used where there are many documents stored on the same computer.
- 'example.html' is the name of the file stored on the server or host. The extension can be any type. For example .txt, .jpg or .pdf depending on the document.

The '#' is used in the URI to help eliminate the issue of naming things and representations of things. To solve this the '#' is used at the end of the URI to reference a page about something whereas the '\' extension represents the thing itself. For example the [British Broadcasting Corporation \(BBC\)](#) uses the document [URI](#) with #programme to refer to a recommendation about a programme it is broadcasting. Then '\programme' is used to refer to the programme itself.

For example the URI for the programme is:

```
http://www.bbc.co.uk/programmes/b006m86d#programme
```

The programme itself would be:

```
http://www.bbc.co.uk/programmes/b006m86d\programme
```

⁶<http://www.w3.org/TR/uri-clarification/>

⁷<http://www.w3.org/TR/uri-clarification/>

2.1.2 HTTP

This is a protocol for interaction between webpages. It is used to enable computers to communicate with other computers. The **HTTP** protocol is a ‘request/response’ protocol. A client sends a request to the server containing information such as request method, URI and message. The server then responds with a status line and a code of success or error and the URI is resolved (Fielding et al., 1999).

When a webpage has an error and cannot be loaded, a set of status codes were developed to handle this. These codes help to identify causes of problems when web pages do not load. The HTTP status line includes a code and a reason phrase.

Table 2.1 shows the two types of classes used when the client’s request is successful and when further action is required.

Client Request Successful	Further action Required
200 OK	300 Multiple choices
201 Created	301 Moved permanently
202 Accepted	302 Found
203 Non-authoritative information	303 See other
204 No content	304 Not modified
205 Reset content	305 Method not allowed
206 Partial Content	307 Temporary redirect
207 Proxy authentication required	

TABLE 2.1: Table to show the successful and further action required http status codes

The two types of error code are Client Error codes (400) which in the instance of providing **LD** are on the users side and Server Error codes (500) which are on the data providers side. The client error is used when a request for a webpage contains errors. The server error is used when the request for the webpage is understood but is not capable of fulfilling the request. The most commonly found client and server error codes and phrases are listed below in 2.2 .

The status codes beginning 2XX refer to actions which have been requested by the client and have been processed successfully. Codes which begin with 3XX represent redirection and in order for the request to be completed further action must be taken. Codes which begin with 4XX mean that the request contains bad syntax and cannot be fulfilled. Codes which begin with 5XX are server side errors and suggest that the server failed to fulfil a valid request.

When considering **LD** and the possibilities for the sale of data we must consider ways to restrict access to content on the web. We suggest a number of http status codes suitable for this in more detail in our suggested technical framework for **LD** in Chapter 8.

Client	Server
400 Bad request	500 Internal server error
401 Unauthorised	501 Unauthorised
402 Payment required	502 Bad gateway
403 Forbidden	503 Service Unavailable
404 Not found	504 Gateway timeout
405 Method not allowed	
406 Not acceptable	
407 Proxy authentication required	
408 Request timeout	
409 Conflict	
410 Gone	

TABLE 2.2: Table to show the client and server http codes

2.1.3 HTML

[HTML](#) gives authors of webpages the ability to publish documents online with text, tables and other elements such as photos. It also allows the retrieval of information via hypertext links. Figure 2.1 shows an early webpage written in HTML.

An example of a simple page written in [HTML](#) is shown below.

```
<HEADER>
<TITLE>The World Wide Web project</TITLE>
<NEXTID N="55">
</HEADER>
<BODY>
<H1>World Wide Web</H1>The WorldWideWeb (W3) is a wide-area
<A NAME=0 HREF="WhatIs.html">hypermedia</A> information retrieval initiative
aiming to give universal access to a large universe of documents.<P>
Everything there is online about W3 is linked directly or indirectly
to this document, including an <A NAME=24 HREF="Summary.html">executive
summary</A> of the project
</BODY>
```

2.1.4 XML

[XML](#) is similar to [HTML](#); it is a language used to display content on webpages. It allows users to design their own tags and structure data and therefore enables users to make data transportable. [HTML](#), however, is designed just to display the data.

Although text written in HTML is easier to read by humans, XML is readable by both machines and humans, whereas only humans can understand text written in HTML.

XML is a format which enables computers to reason with other data in XML format. XML is one example of a basis of RDF, which is a standard model for data interchange on the SW.⁸ which we also explain in more detail later

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<CATALOG>
<CD>
<TITLE>Empire Burlesque</TITLE>
<ARTIST>Bob Dylan</ARTIST>
<COUNTRY>USA</COUNTRY>
<COMPANY>Columbia</COMPANY>
<PRICE>10.90</PRICE>
<YEAR>1985</YEAR>
</CD>
<CD>
<CD>
<TITLE>Greatest Hits</TITLE>
<ARTIST>Dolly Parton</ARTIST>
<COUNTRY>USA</COUNTRY>
<COMPANY>RCA</COMPANY>
<PRICE>9.90</PRICE>
<YEAR>1982</YEAR>
</CD>
```

The difference in formatting between XML and HTML is that XML is displayed with no formatting whereas HTML enables colours, fonts and backgrounds to be modified. XML also enables information to be structured which is important especially when looking to reason with the data.

2.1.4.1 OAuth

OAuth is an authorisation protocol which allows a third-party application to obtain access to an HTTP service (Hammer-Lahav, 2010). OAuth is the bridge between users and the Service Provider (SP) or owner, acting as an authorisation layer. The system uses access tokens which are assigned to users, which replaces the need for a username and password. This is a system which could easily be integrated with LD to provide a suitable method to authenticate transactions on the web.

There are a three actors within the OAuth Authentication System:

⁸<http://www.w3.org/RDF/>

- The Service Provider - A web application that allows access via OAuth.
- The User - An individual who has an account with the SP.
- The Consumer - A website or application that uses OAuth to access the SP on behalf of the user.

There is one key in the system which is granted through a Consumer Secret:

- The Consumer Key - A value which is used by the consumer to identify itself to the SP.
- The Consumer Secret - A secret which is used by the consumer to establish ownership of the consumer key.

There are also two tokens in the system:

- The Request Token - A value used by the consumer to obtain authorisation from the user and is exchanged for an Access Token.
- The Access Token - A value which is used by the consumer to gain access to the resource on behalf of the user instead of using the SPs credentials.

Figure 2.2 (taken from <http://oauth.googlecode.com/svn/spec/core/1.0/>) shows the data flow through the OAuth authentication system. There are 7 steps in this system which we describe in more detail below:

- (A) The Consumer requests an unauthorised Request Token. (This includes the consumer key).
- (B) The Service Provider grants a Request Token. (This includes the token and the token secret).
- (C) The Consumer (website or application using OAuth to access SP on behalf of the user) directs the user to a Service provider.
- (D) The Service Provider then directs the User to the Consumer.
- (E) The Consumer Requests the Access Token. (This includes the consumer key and the token).
- (F) The Service Provider Grants the Access Token. (This includes the token and the token secret).
- (G) The Consumer Accesses the Protected Resources. (This includes the consumer key and the token).

OAuth Authentication Flow

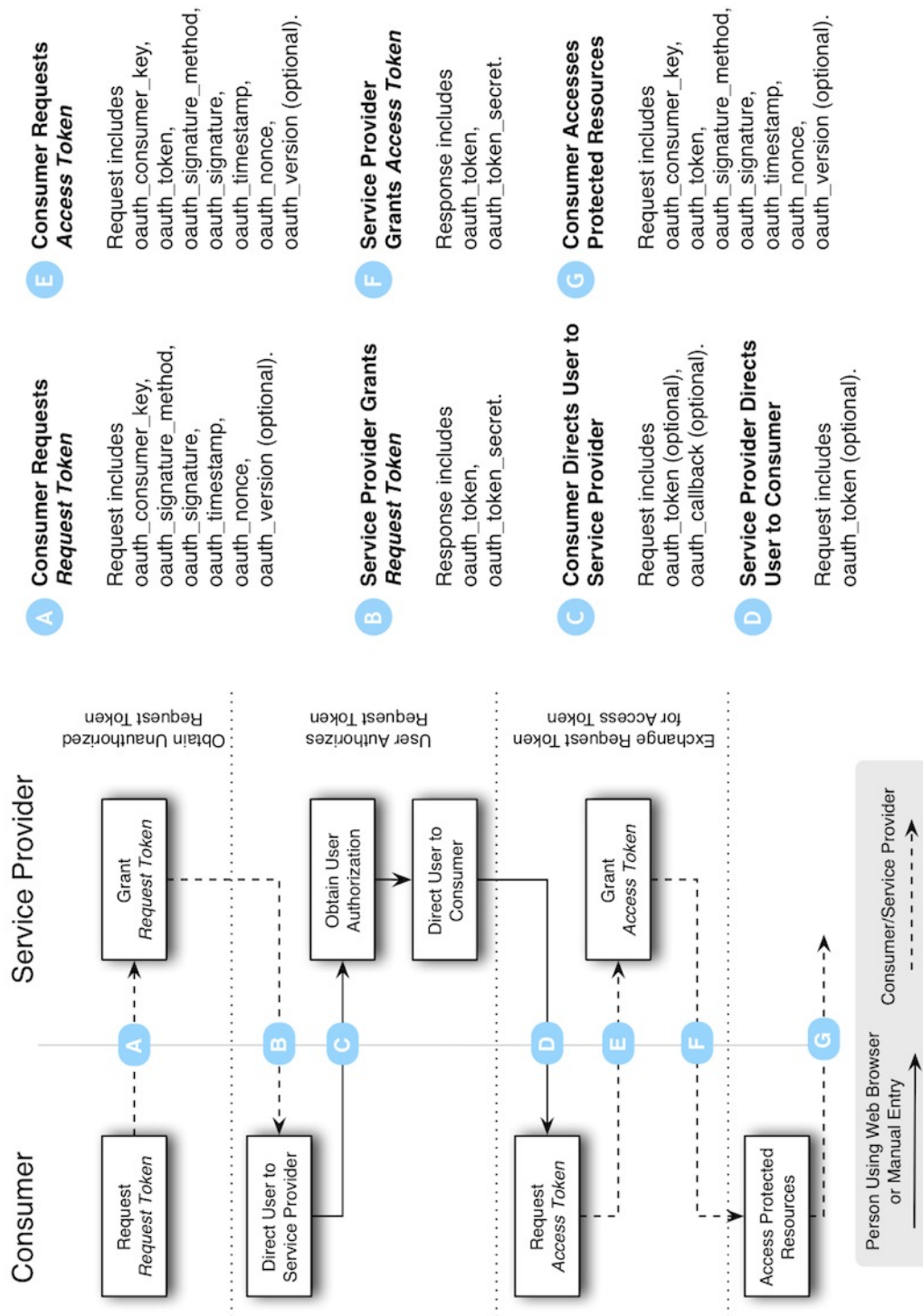


FIGURE 2.2: OAuth Authentication Flow Diagram from <http://oauth.googlecode.com/svn/spec/core/1.0/>

2.1.5 Limitations of data on the Web

Despite the huge number of capabilities of the web as we use it now there are a number of limitations which we will explore in more detail below.

Tim Berners-Lee, the founder of the web, intended it to be for data as well as documents. However, the web as we use it now is a web of documents which although are linked, the content in these documents is not. Many ways of creating colourful and interactive websites have been developed making them accessible and easy to use, but do not readily enable computers to interact in ways which we would like, for example only returning relevant results of queries which we execute. Due to the fact that the web is geared to use by humans, a great deal of additional work and time is spent trying to search for and make decisions about the data and information which is retrieved.

The first issue of time consuming tasks on the web is search. A considerable amount of time spent on the web by users is searching for required information.

Pages stored on the web are indexed and can be retrieved through standard web search engines such as Google⁹. The Google search engine uses a web crawling robot (googlebot) to find and retrieve web pages (Price, 2001). An indexer is then used to sort and store words from every web-page. This results in an index of words which is stored in a database. A query processor compares a user's search query with the database index and then suggests the documents which it finds most relevant.

The search however is a keyword search, where the search engine used will look for occurrences of a word despite its relevance to a users search term. For example, if a user enters a search for 'Lion', they may retrieve pages regarding the operating system 'Lion', the animal 'lion' or book about a 'lion' and so on. This initial search may return all the instances of lion but is not able to reason with the information found in the documents containing that reference. The aim of the SW is to enable users to search for the search term 'lion' and return pages which contains metadata about the subject. For example, the technology will be able to infer that lion is a particular version of the Mac OSX operating system and that Mountain Lion is another instance of the operating system, from this it will be able to offer other instances of OSX for example Leopard and Snow Leopard. This level of inference will enable users to reduce search times for data and information on the web as searches will be carried out more precisely and return more accurately reasoned results. Other benefits of the SW include the ability to integrate different datasets easily with others in structured formats. These datasets can then be easily shared and added to at later stages without the need for repetition of similar datasets.

Search engines are unable to find specific content on the web due to the large amount of information available. Much content is stored in a part of the web which is not

⁹<http://www.google.com>

seen from a users browser (Wright, 2008). This is called the ‘deep web’ (He et al., 2007). The content on the ‘deep web’ is stored in searchable databases. Traditional search engines do not yield results from the deep web and will only produce a result to a direct request (Bergman, 2001) but, how do we search the databases in the deep web? If the content in the ‘deep web’ is stored in standard formats using RDF, it becomes much more manageable, as links can be made to particular resources and more structured responses from searches are achieved. This is possible due to the nature of the technology which aims to give everything a URI. The unique identifier represents physical resources, concepts or information resources on the web (Hyland, 2010).

Organisations who hold a great amount of digital information find that their data is unstructured, and, consumers of this information, although able to search for the information, are unlikely to find specifically what they are looking for. More powerful search engines are only useful if the content is structured in a way that users can find exactly what they are looking for with minimal extra effort (Harris, 2010; Hendler, 2010).

LD will enable the structuring of knowledge to be more refined with the use of XML Schema (Antoniou and van Harmelen, 2008). LD will enable individuals and organisations to structure the data and information which they hold in a way which will help to remove problems such as inconsistencies in the data and the addition and removal of data which maybe distributed across many systems. It is also useful for the retrieval of information as it can be stored in structured forms which rather than searching via keywords, queries can be executed and presented in a human readable form.

Once in a structured form the data is then easy to re-use and share by others and therefore makes the data useful for a wider audience than has previously been experienced. We also note that the data can be integrated with other datasets more easily due to being in standard formats.

2.2 The Semantic Web

2.2.1 What is the Semantic Web?

The SW is not a new concept in computer science. Its origins come from knowledge representation techniques (Davies et al., 2003). Tim Berners-Lee quotes that the SW is the vision he had of the web from the beginning. However it has taken time for technology to improve and more tools become available to realise this vision.

2.2.2 The Technologies

We explain in the next section, there are key technologies used on the web and for LD. We now explain the critical technologies for explicitly specifying semantics.

By using these methods we are able to add context to the information we publish: i.e., where it came from, who created it, what it is about, make links to other data and where semantics of the data is explicitly specified.

In the previous section we outlined [URIs](#). Further to this we note that a [URI](#) identifies a document on the web. This document is data about a thing, be it a place, person or concept. In order to understand the [SW](#) we should clarify that a [URI](#) of a thing is separate to the [URI](#) of a web page which talks about it. The power of the [SW](#) means we can use tools to infer that things are the same. For instance if two individuals provide information about Southampton we are able to infer that these things are the same thing or have the same name but are in fact different ‘things’.¹⁰

2.2.2.1 Ontologies

An ontology is a way of formally representing knowledge as concepts within a domain ([Antoniou and van Harmelen, 2008](#)). The ontology allows relationships to be made between the concepts and also allows reasoning about entities within the domain.

[OWL](#) is a [SW](#) language designed to represent rich and complex knowledge about things, groups of things, and relations between things ([Hitzler et al., 2009](#)).

2.3 Linked Data

2.3.1 What is Linked Data?

The introduction of a new technology (in this instance, [LD](#)) influences the added value of data ([Longhorn and Blakemore, 2007](#)) and models for distributing this [LD](#) have, therefore, become a pivotal area of research ([Lytras and Garcia, 2008](#)). [Feigenbaum and Herman \(2007\)](#) detail organisations who are adopting [Semantic Web Technologies \(SWT\)](#): this includes British Telecom, who have used the technology to build an on-line prototype to help vendors develop new products together. Vodafone are using the technology to enable their consumers to download content to devices faster. Renault have used the technology in car repair and diagnostic documentation as they found that numerous objects are found across many different systems, and that the technology was well suited to linking these together ([Servant, 2008](#)). Alongside the development of the technical aspects of a new system, the environment in which the organisation operates must also be investigated to inform a model the organisation can use to carry out business ([Picard, 2000](#); [Kanliang, 2004](#); [Latif et al., 2009](#); [Allemang, 2010](#); [Chan-Olmsted, 2004](#)). A pivotal area for research for the [LD](#) community would be to discover how the technology will impact the business or revenue model of [LD](#).

¹⁰<http://www.w3.org/DesignIssues/Axioms.html>

LD is structured data from multiple sources, the technology enables users to create links between data from different sources (Bizer et al., 2009). **LD** is currently published on the Web in a way that can be read by machines. This differs from most data which is published on the web as it can only be understood by humans. Data written in **RDF** is able to be linked to other datasets in the same format (Bizer et al., 2009; Becker and Furness, 2010; Yu, 2011).

LD is where hypertext meets data. Hypertext is the link between text documents (Rizk et al., 1990). **RDF** is used in **LD** to link data to other data using triples. We go into more detail about triples below.

We point out at this stage that **LD** is concerned with the data itself, in particular the storage, linking presentation and distribution of data, whereas the **SW** is concerned with the reasoning behind the data, and the relations and vocabularies used for manipulating the data. With this in mind we will explain **LD** in more detail. By structuring data in this way, different data sources are able to interact with each other in a useful form. The linked web of data uses **RDF** to link the documents, whereas traditionally Hyper Text Mark-up Language (**HTML**) pages used hyperlinks to other pages have been used on the Web (Bizer et al., 2008). The concept of a linked Web of data will encompass many different data sources linked by features that previously were not feasible, for example relations to things and similar features. **SWT** enable these data sources to be linked and reasoned with and have the capability to expose previously inaccessible databases to users who would otherwise not be able to access such data sources.

Tim Berners-Lee¹¹ has outlined how links work in **LD**. The example given shows when an individual searches for information about a person on the Web, they will encounter a Web page with a URI beginning `http://`. This Web page will have information about the person, perhaps their date and place of birth. Each Web page will contain links to other web pages which will contain information about the place, events, places to visit etc which, as a whole, are much more valuable than just plain information.

Tim Berners-Lee also outlines four key principles for creating **LD**:

- URIs are used as names for things.
- HTTP URIs are used so that people can look up those names.
- When someone looks up a URI, useful information is provided in **RDF** form.
- **RDF** statements are included that link to other URIs so that they can discover related things.

Figure 2.3.1 shows all the datasets which have been published and are available as **LD**. The larger the circle in the diagram shows an organisation which has published large

¹¹http://www.ted.com/talks/tim_bern timers_lee_on_the_next_web.html

quantities of data. The diagram shows the links between 295 datasets and arrows are used to illustrate the links. The diagram is updated as more datasets are published. The diagram below is the most recent version updated in September 2011. The first [LD](#) diagram published in 2007 (see [2.3.1](#)) showed 12 datasets so we notice that in four years this diagram has grown considerably.

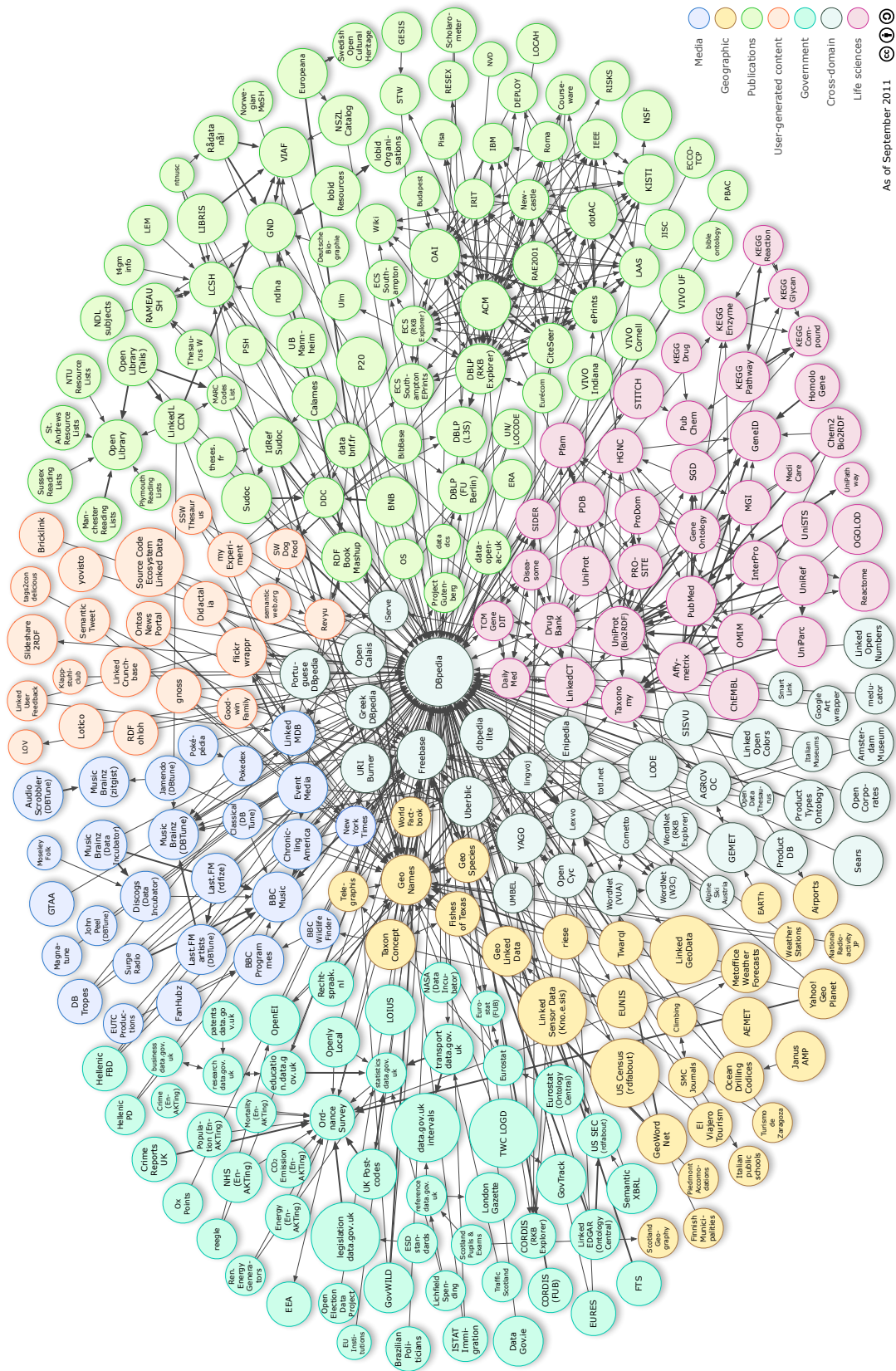


FIGURE 2.3: Linking Open Data cloud diagram, as of September 2011, by Richard Cyganiak and Anja Jentzsch. <http://lod-cloud.net/>

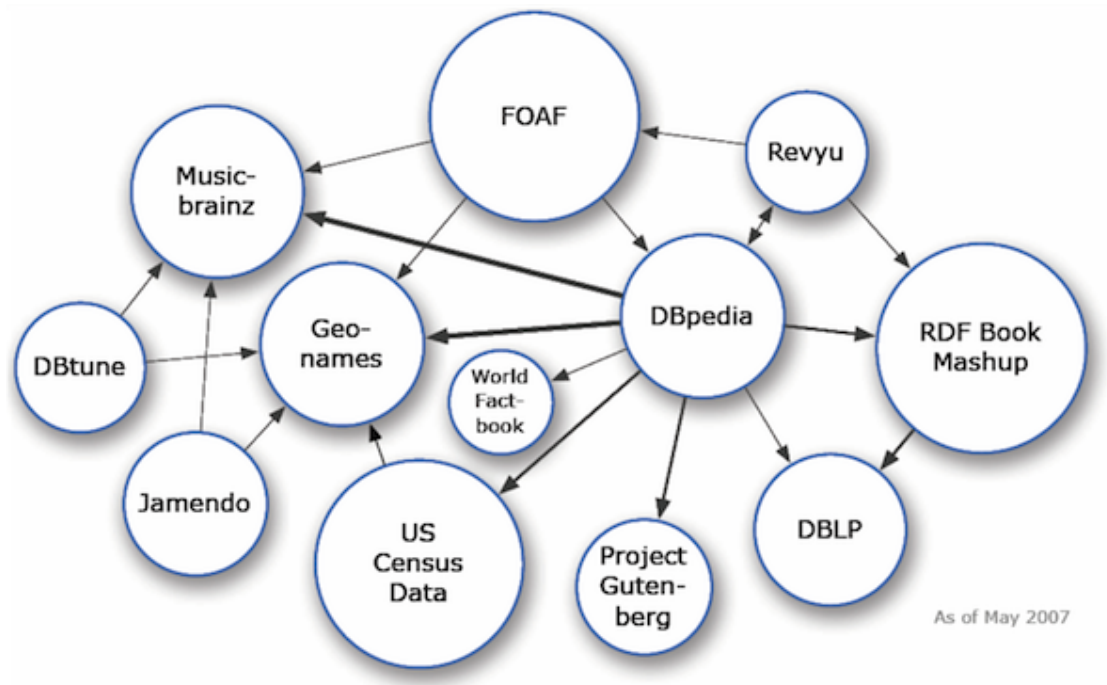


FIGURE 2.4: Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. <http://lod-cloud.net/>

We can see from the cloud and the classifications which have been used that government data includes [OS](#) data but this is not necessarily a good classification as governmental data does not necessarily mean it is geographic and geographic data is not necessarily governmental.

2.3.2 Datasets in the Linked Data Cloud

This next section highlights and examines the key datasets in the Linked Data Cloud illustrated in [Figure 2.2](#). We have sampled Open data with the greatest number of links. [Table 2.3](#) outlines the data and in what format it is downloadable.

Data	What is it	Download Format
MusicBrainz (http://musicbrainz.org)	MusicBrainz is an open music encyclopaedia that collects music metadata and makes it available to the public.	XML
Ordnance Survey (http://data.ordnancesurvey.co.uk)	Geographical data about England, Wales, and Scotland. Provides identifiers for counties, cities, wards, census areas, identifiers for postcodes, Relationships between geographical areas (containment, borders)	RDF/XML, Turtle and JSON
DBpedia (http://dbpedia.org)	DBpedia is a crowd-sourced community effort to extract structured information from Wikipedia.	N-Triples and N-Quads
RKB Explorer (http://www.rkbexplorer.com)	As part of the ReSIST Project, a Resilience Knowledge Base (RKB) has been built; it has gathered data from many bibliographic and other sources, and structure has been added to allow it to be queried by topic. Particular attention has been given to topics concerned with Dependable Computing and Resilient systems	Linked Data, SPARQL endpoint and RDF dumps
The Gene Ontology (http://www.geneontology.org)	The Gene Ontology project is a major bioinformatics initiative with the aim of standardizing the representation of gene and gene product attributes across species and databases.	MySQL, OBO, XML, OWL, RDF, XML, SQL
Freebase (http://www.freebase.com)	An entity graph of people, places and things.	Turtle RDF
Geonames (http://www.geonames.org)	The GeoNames geographical database covers all countries and contains over eight million place names	RDF

TABLE 2.3: Key Linked Datasets from the OS Cloud in Figure 2.2

We notice from the data we have listed in the table above that there is a lack of commercial data. In Chapter 4 we examine the possibilities for commercial data and the potential for charging for data.

The cloud has been sectioned into different classifications. These include media, geographic, government and life sciences. We notice however that the classifications are not necessarily true. We take OS as an example, we can say that the business of OS is truly geographic and yet it has been classified into the government section in the diagram. We would suggest that this classification is not necessary as it tends to blur what it illustrates in the diagram.

2.3.3 W3C and Data Publishing

The W3C produces technical recommendations for publishing of data. Five rules for publishing data were outlined by Tim Berners-Lee to help organisations understand the aim of publishing data and how we can make the most of it. By introducing a ranking system it allows people to have an introduction into publishing data without being overwhelmed with requirements. This makes the publishing of data more accessible with the long term aim to achieve all of the stars.¹²

- * On the Web, open licensed
- ** Machine-readable data
- *** Non-proprietary format
- **** RDF standards
- ***** Linked RDF

We explain what each of these stars represent in more detail below. To achieve one star the data can be available on the web in any format as long as it has an open license. For example a scanned image of some data or a pdf document. To achieve two stars the data must be available in a machine readable format with structure, such as an excel file rather than a scanned image. To reach three stars the data must be in a machine readable format plus a non-proprietary format, that is, rather than using the excel file format, the data would be saved in CSV instead. The four star tier comes closer to the ideal standard required which is as three stars but the data is available in open standards set out by the W3C. This would include [RDF](#) and [SPARQL](#). This enables identification of objects and enables items to be linked to them. Five star data states that it must be all of the four star ranking, but with links to other data in order to provide a context to the data which a user has available.

¹²<http://www.w3.org/DesignIssues/LinkedData.html>

2.3.4 The Application of LD

LD has the potential of application across both the public and private sector. Different organisations have their own reasons for making use of LD and the surrounding technologies. First we notice organisations such as the BBC who have large datasets on say wildlife and music. They are not selling their data to make revenue but have produced their data in RDF which in turn will make the data more structured, more accessible and easier to direct to. The end user may not see the data or realise that it is being stored as LD, but, for the holding organisation it has the benefit of being structured with the ability to link to other datasets if required. There are also other organisations who have data which they wish to sell for profit. We recognise that organisations such as OS have a large amount of data which if sold in parts could be of benefit to its users and still maintain economic benefit to the holding organisation. In order to sell data in portions rather than whole datasets it is important that all aspects of a change in business model are investigated to ensure that the most suitable model is selected in order to maintain profit.

Despite the apparent economic and commercial issues surrounding LD we notice other issues for organisations which include privacy and sensitivity of data. This reaches to both the public and private sectors, as data published must maintain privacy of individuals and organisations and therefore must adhere to certain restrictions to ensure the data published does not cause damage to organisations and individuals either financially or ethically.

The web as we use it now allows social interactions from sites and communities with similar interests but the value is added by using semantics stored in, for example, a travel ontology. This is also known as the network effect which is extended when different sites containing information about different topics can be linked together.

Hyland (2010) discusses the key features of LD which include decentralising and exposing large stores of data enabling users to build new applications and acquire better resource discovery and re-use.

The next section goes on to explain the technologies surrounding LD and gives specific examples of the technology in order to help the reader understand the context in which the technology can be applicable.

2.3.5 The Technologies

As we detailed earlier, there are the four main technologies used on the web. These are URI's, HTTP, HTML and XML. LD however, extends the use of the key technologies and a number of further technologies are used which make LD possible. These technologies

include [RDF](#), [Resource Description Framework Schema \(RDFS\)](#) and [SPARQL](#) which we outline in more detail below.

2.3.5.1 RDF

[RDF](#) is a general purpose language from the W3C which is designed as a way to represent information and model data interchange on the Web ([Miller and Manola, 2004](#)).

[RDF](#) uses [URIs](#) to name relationships between two objects. This is what we call a triple which contains a subject, a predicate and an object. The [RDF](#) vocabulary contains classes and properties.

In order to explain [RDF](#) in more detail we explain how ordinary data about a person would look in RDF. The example we use to demonstrate [RDF](#) is information about a person.

The subject in this instance is the resource <http://www.jbexample.com> which identifies the person Jennifer Black. Jennifer Black has an email address jblack@example.com and has a nickname Jen. She is interested in the [SW](#) and her picture can be viewed at www.jbexample.com. She also knows a person called Bob Farmer.

Information is represented in ‘graphs’ which outline the information in the triple. [RDF/XML](#) is a syntax which expresses an RDF graph as an XML document. [Turtle](#) was introduced as a more accessible alternative to [RDF/XML](#) as it is easier to understand by humans.

The text below shows what the same information would look like in [RDF/XML](#). Note that this data is a description about a thing, in this instance a person. The [RDF](#) enables each property of the person to be identified and therefore gives the opportunity to link to that specific detail.

```
<?xml version="1.0"?>
<rdf:RDF xmlns:foaf="http://xmlns.com/foaf/0.1/"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
<foaf:Person rdf:about="http://www.jbexample.com/me">
<foaf:name>Jennifer Black</foaf:name>
<foaf:mbox rdf:resource="mailto:jblack@example.com" />
<foaf:homepage rdf:resource="http://www.jbexample.com/" />
<foaf:nick>Jen</foaf:nick>
<foaf:depiction rdf:resource="http://www.jbexample.com/img_small.jpg" />
<foaf:interest rdf:resource="http://www.semanticweb.org" />
<foaf:knows>
<foaf:Person>
```

```
<foaf:name>Bob Farmer</foaf:name>
</foaf:Person>
</foaf:knows>
</foaf:Person>
</rdf:RDF>
```

The properties of the resource are identified by the elements `<foaf:person>`, `<foaf:name>` and `<foaf:nice>` etc.

This is what the TURTLE version of the RDF looks like:

```
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
<http://www.jbexample.com/me>
  a foaf:Person ;
  foaf:name "Jennifer Black" ;
  foaf:mbox <mailto:jblack@example.com> ;
  foaf:homepage <http://www.jbexample.com/> ;
  foaf:nick "Jen" ;
  foaf:depiction <http://www.jbexample.com/img_small.jpg> ;
  foaf:interest <http://www.semanticweb.org> ;
  foaf:knows [
    a foaf:Person ;
    foaf:name "Bob Farmer"
  ] .
```

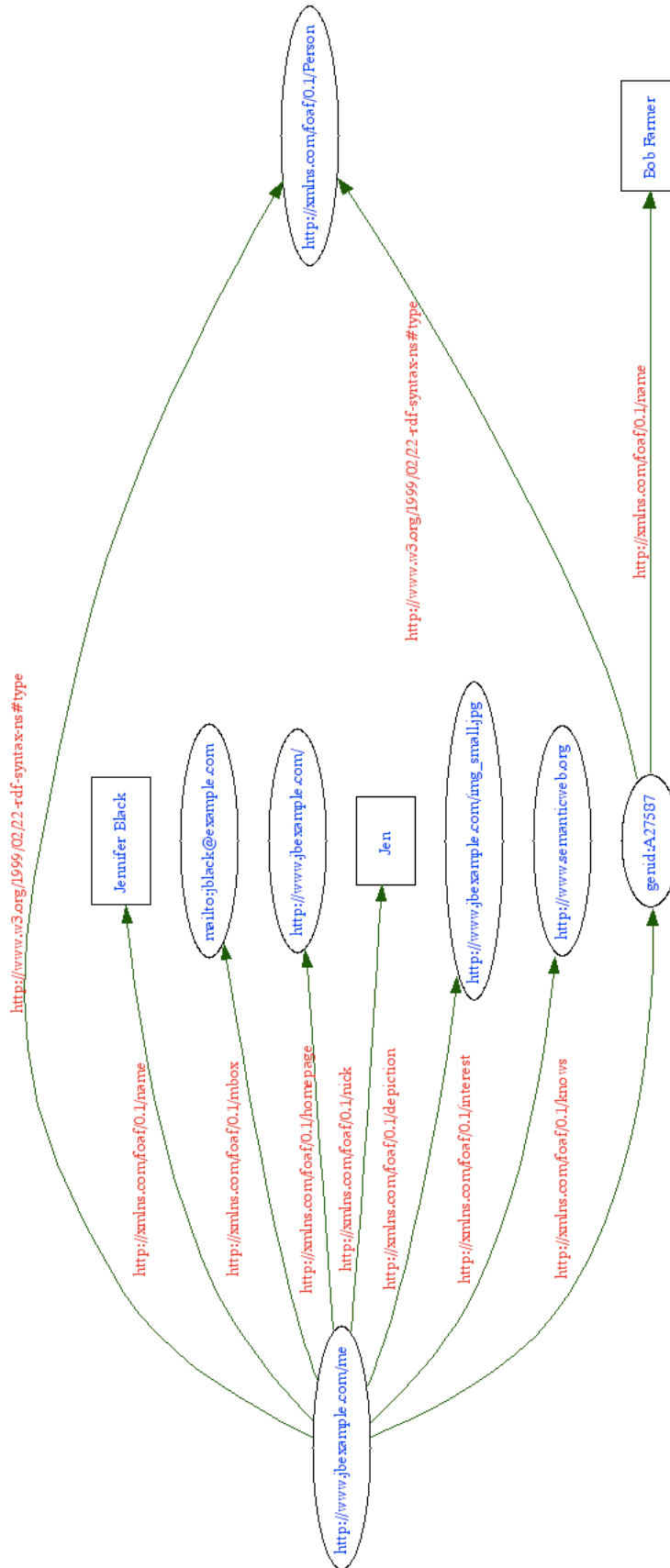
The triples for this RDF example are shown in Table 2.4 below:

Figure 2.5 shows the graph of the RDF so we can see the relations between the data.

Subject	Predicate	Object
http://www.jbexample.com/me	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://xmlns.com/foaf/0.1/Person
http://www.jbexample.com/me	http://xmlns.com/foaf/0.1/name	"Jennifer Black"
http://www.jbexample.com/me	http://xmlns.com/foaf/0.1/mbox	mailto:jblack@example.com
http://www.jbexample.com/me	http://xmlns.com/foaf/0.1/homepage	http://www.jbexample.com/ "Jen"
http://www.jbexample.com/me	http://xmlns.com/foaf/0.1/nick	http://www.jbexample.com/img_small.jpg
http://www.jbexample.com/me	http://xmlns.com/foaf/0.1/depiction	http://www.semanticweb.org
http://www.jbexample.com/me	http://xmlns.com/foaf/0.1/interest	http://xmlns.com/foaf/0.1/Person
http://www.jbexample.com/me	http://www.w3.org/1999/02/22-rdfsyntax-ns#type	genid:A11365
genid:A11365	http://xmlns.com/foaf/0.1/knows	genid:A11365
http://www.jbexample.com/me	http://xmlns.com/foaf/0.1/name	"Bob Farmer"
genid:A11365		

TABLE 2.4: Table to show the triples in the RDF example

FIGURE 2.5: A graph to show the data model of the example RDF



JavaScript Object Notation (JSON) is a text format which is readable by both humans and machines and was designed to be a portable subset of JavaScript (Crockford, 2006). It differs from **XML** in that it does not have closing tags and therefore the resulting data is shorter and easier to read. **RDF JSON** represents a set of **RDF** triples as a series of nested data structures which aims to serialise **RDF** in a structure that is easy for developers to work with (Alexander, 2008).

2.3.5.2 RDFS

RDF is a language used for representing information on the web, specifically metadata, which we outline in more detail later. **RDFS** enables users to express simple statements about resources using properties and values.¹³ Users also need to be able to define the terms which they intend to use in those statements so that they can describe specific classes of resources. Classes are used to represent categories of things.

In order to explain **RDFS** in more detail we use the example outlined by the W3C.¹⁴

An organisation (example.org) wants to provide its consumers with information of the different types of motor vehicles it sells. To do this they use classes.

```
ex:MotorVehicle rdf:type rdfs:Class
```

When we describe things we can include additional classes

```
ex:Van      rdf:type  rdfs:Class .
ex:Truck    rdf:type  rdfs:Class .
```

In order to represent something as a type of something we use `rdf:type`. For example a van or truck is a type of motor vehicle. This links two classes together.

```
ex:Van      rdfs:subClassOf  ex:MotorVehicle .
```

As we detailed earlier, the schema explained above can be represented as triple, subject, predicate, object.

By using **RDFS** we are able to create classes of things and relations to them. In the next section we outline how we retrieve information stored in these formats.

¹³<http://www.w3.org/TR/2004/REC-rdf-primer-20040210/#rdfschema>

¹⁴<http://www.w3.org/TR/2004/REC-rdf-primer-20040210/#rdfschema>

2.3.5.3 SPARQL

SPARQL is a query language for databases. It is used to retrieve and manipulate data stored in the **RDF** format.¹⁵ A query written in **SPARQL** is a very powerful way of retrieving data from a dataset. It allows a query to consist of triple patterns, conjunctions, disjunctions. The key feature of **SPARQL** queries is that the query itself is unambiguous. The user is able to specifically outline the types of data they would like to be returned from the query.

In order to understand **SPARQL** queries we have outlined an example **RDF** dataset below.

```
@prefix foaf: <http://xmlns.com/foaf/0.1/> .

_:a foaf:name "Jennifer Black" .
_:a foaf:mbox <mailto:jblack@example.com> .
_:b foaf:name "Rachel Black " .
_:b foaf:mbox <mailto:rblack@example.org> .
_:c foaf:mbox <mailto:andrew@example.org> .
```

An example of **SPARQL** query to request data from the data set is detailed below:

```
PREFIX foaf: <http://xmlns.com/foaf/0.1>
SELECT ?name ?mbox
WHERE {
?person foaf:name ?name.
?person foaf:mbox ?mbox.
}
```

This query is asking the dataset to select the name and email address of the people in the dataset. It defines how the data is stored and will return only names and email addresses of people stored in that specific dataset. Queries can be extended to answer specific questions such as what are all the counties in the England, so it is capable of selecting only counties found in England?

The resulting query will be

name	mbox
'Jennifer Black'	<mailto:jblack@example.com>
'Rachel Black'	<mailto:rblack@example.org>

¹⁵<http://www.w3.org/TR/rdf-sparql-query/>

Some view the [SW](#) as a collection of databases and that [SPARQL](#) is a way of retrieving data from all of the databases in one query. This has multiple benefits for organisations as new technologies do not need to be developed or implemented to query their data. Also users are able to access more than one database using the same query. Although we are now able to do this using search engines on the web, we are not able to readily access data from databases stored in the deep web.

2.3.5.4 Tools for Linked Data

Since the [LD](#) movement really gained momentum, [LD](#) support and accessibility for users is becoming more comprehensive. Emerging vocabularies and tools allow digital content to be more richly experienced by automating processes, thus, aiding the generation of links ([Luczak-Roesch, 2009](#)). [LD](#) support is becoming more comprehensive over time, making [LD](#) more accessible to all ([Bizer et al., 2009](#)).

Some of the tools being developed include those which will make the process of translating pre-existing datasets into [RDF](#) and writing and executing queries more efficient. Making the transition to [LD](#) for non technical users much simpler.

We list a number of these tools below, but we also note that as more data is becoming available on the web and more users are beginning to use the data, more tools are becoming available to help simplify this transition.

Current tools available include RDFisers. These are groups of tools which have been developed to convert data on the web into the [RDF](#) format, allowing the data to be structured in web pages, searched for and linked to.¹⁶ Creating [RDF](#) datasets can be a time-consuming task, and to make [LD](#) more accessible, tools have been developed which make the task of creating the data on the web easier. Examples of the types of conversions which can be carried out include from email, calendar, GPS, BibTex to [RDF](#). [Davies and Donaher \(2011\)](#) began to explore the development of more specific tools to enable non-technical users to create their own linked data. This is important as it means the utilisation of the technology is for everyone.

Tools such as those which enable users to clean data have also become more readily available. These tools look for any inaccuracies within the data to ensure the data is consistent in labels and tags. The tools are also used to transform data into different formats. An example of such a tool is google refine.¹⁷ (Previously known as Freebase Gridworks)

A number of [SW](#) search engines are being developed to enable users to query specifically documents which are written in [RDF](#).¹⁸ A [SW](#) search engine enhances an ordinary search

¹⁶<http://esw.w3.org/ConverterToRdf>

¹⁷<http://code.google.com/p/google-refine/>

¹⁸<http://esw.w3.org/topic/TaskForces/CommunityProjects/LinkingOpenData/SemanticWebSearchEngines>

engine by indexing [RDF](#) data and providing an interface to search through this data. Whereas traditional search engines index key words which often have no real relevance to the search term the user specifies. Searching using data indexed in [RDF](#) enables a machine to infer relations with the data rather than displaying data and leaving the user to make decisions about the relevance of the data.

Semantic search engines are more powerful when data is expressed with meaning and they are able to produce more definitive results by making decisions. This will provide machines with the power to give more targeted answers to questions and will try to help users with further searches that relate to the original search. [SW](#) search engines are a strong development in web browsing, as they can reduce the time spent by individuals looking for information on the web. The question which we want to answer is how do we quantify this reduction in search time? Do users value their time searching as valuable or are they happy spending time searching for answers to their queries.

Some examples of [SW](#) search engines include [Sindice](#),¹⁹ [Swoogle](#)²⁰ and [Falcons](#).²¹

There are also a number of browsers which display [SW](#) pages more consistently than current web browsers, examples of these include [Tabulator](#),²² [Disco](#),²³ and [Zitgist](#).²⁴

2.3.5.5 APIs

An [Application Programming Interface \(API\)](#) is an interface used for different software components to communicate with each other. APIs can be used in web development and is essentially a web service as defined by the W3C ([Vedamuthu et al., 2007](#)).

A mashup is created by combining different web resources and data to create new web applications [Benslimane et al. \(2008\)](#). With the use of [LD](#) technologies we are able to use [RDF](#) data to create mashups using [RDF](#), for example we can link data on crime figures in a town with data about house prices and display this data on a map or generate information about bands and concert venues and also display this on a map. What we notice when starting to generate web based applications is the need to pinpoint all of this data to something and this is location. To refer to a map which displays crime figures is much richer if we are able to see where the crime is in relation to a location and other locations we find of interest.

[Data.gov.uk](#) for example contains data from many different sources such as health, crime, transport and local government, this data is available under license. [Data.gov.uk](#) has its own API which is used to access the catalogue of data or users are able to download

¹⁹<http://www.sindice.com/>

²⁰<http://swoogle.umbc.edu/>

²¹<http://ws.nju.edu.cn/falcons/>

²²<http://www.w3.org/2005/ajar/tab>

²³<http://www4.wiwiw.fu-berlin.de/bizer/ng4j/disco/>

²⁴<http://dataviewer.zitgist.com/>

the data in CSV or [JSON](#) format. We explore data.gov.uk in more detail in chapter 4 on and illustrate how different mashups can be created using the data.

2.4 Conclusion

This chapter has provided an insight into the current architecture of the web as we use it now. We outlined the key technologies and then go into more details about the relevant technologies for [LD](#). We then introduced the [SW](#) and how the two can be applied to give benefit to consumers and producers of information. We also outlined the key affordances for both [LD](#) and the [SW](#) and look at the potential for a business case for both.

The web has changed the way we view documents online and [LD](#) provides us with the opportunity to apply the same changes to viewing data online. [LD](#) has the potential to open up new markets which have previously not been accessible. The next chapter will address the key area we are investigating which is [PSI](#) and in particular [GI](#) and the following chapter will look at the business case for [LD](#).

Chapter 3

PSI and Geographic Information

3.1 Public Sector Information

In the previous chapter we introduced the technologies specific to this thesis. This included detail about how LD has been created and how the technologies of the SW can be used to reason with this data. In this chapter we examine the key focus for our research which is GI. We note that PSI contains a strong element of GI and in order to define specific characteristics of GI we begin by specifically defining different types of PSI and then go into particular detail about GI. We also detail how this relates to LD and we then investigate GI and how important GI is to LD.

PSI includes all information produced and maintained by the government. There are many different types of PSI. We notice that GI can occur in all of the types of PSI as well as as an entity on its own.

In order for us to establish what types of PSI we are concerned with in terms of LD, we look at the data available from the Data.gov website. ¹<http://data.gov.uk> We have listed the different types of PSI data below.

1. Economic and Business
2. Social
3. Legal
4. Meteorological
5. Scientific
6. Transport

¹.

7. Environmental, agricultural and fisheries
8. Cultural
9. Political

The list above is clearly not exhaustive, as it lists the data from one central resource - Data.gov.uk. This site is a large resource of data but is not the only resource available. We felt that this gives us a general idea about the types of PSI that is available. We also note that the data on the site is not specifically in a LD format and therefore the data is available on this site in varying formats.

The development of the Open Data initiative in the United Kingdom (UK) and of the site data.gov.uk is focused towards the access and reuse of PSI which has significantly improved the reuse of PSI (Sheridan and Tennison, 2010). This site contains datasets for various types of topics including crime, health and public spending. This data is now readily available on the web, where previously it has been hard to find. With this data now more accessible through the data.gov site it will enable people to find and reuse the data and create more applications, which we already notice the increasing growth of applications containing PSI, made possible by the development of LD technologies.

PSI is produced by organisations to inform government, businesses or individuals. The Met Office provides information in the form of forecasts to allow organisations to make informed decisions about impacts of weather. OS collects and distributes mapping information. The Driver and Vehicle Licensing Agency however provides complete, accurate and up to date registers of drivers and vehicles and its majority of income is from registration fees for drivers and vehicles.

The PIRAIInternational (2000) report suggested that there have been barriers to PSI information. These barriers include the format and accessibility of the data. A similar report OFT (2006) also outlined some of the potential difficulties experienced with public sector organisations and the use and re-use of data. This highlighted the requirement to make the information work better for consumers.

Pollock (2008) outlines the Key Features of PSI and a Public Sector Information Holder (PSIH) to be: non-rivalry; high fixed costs; high potential for use and re-use and the two-sided nature of a PSIH.

These features are outlined in more detail below:

Non-rivalry (Zero Marginal Cost) – This means that if one consumer purchases the information, it will not prevent another consumer from purchasing the same piece of information. Unlike non digital products which, once sold, cannot be used at the same time.

High Fixed costs – The collection, processing and storage of data can be high. Despite costing a small amount to reproduce, the cost of producing the first item can be high.

High Potential for use and re-use – Digital data can have many uses across a wide range of markets. Information can be used and re-used in many different ways as it does not lose its quality if it is sold to many users or just one.

Two-Sided Nature of PSIHs – OS for example collects data and changes to maps and then this information is supplied to third parties. Two sided in this instance shows that costs are involved in the collection of the data and then with the dissemination of the data.

Pollock (2008) discusses when considering the supply of information, not only price should be specified, but what can be done with the information required. This shows different terms for different charging policies. For example,

- Profit maximising and cost recovery - maintain a strong control over the re-use and distribution of the data
- Marginal cost pricing - allow the data to be ‘openly’ available. Free to re-use and redistribute the data.

Price elasticity of demand is a term in economics used to describe how a demand for a product can change the price of the product (Flores and Carson, 1997). Elasticity of demand for PSI is illustrated by Pollock (2008) in this report and states that a change in charging policy by a PSIH (or other entity) allows one to elicit the elasticity of demand by comparing prices and demands before and after the change. It is suggested that the changes in demand noticed in this study could be due to a backlog of demand. Consumers have wanted to purchase the information but at the high prices experienced before they chose not to purchase the information. As a price reduction was experienced, these people chose to make a purchase, resulting in a large increase in demand, however, over time this demand would stabilise as more people had the information. Despite this stabilisation of demand it is favourable that once people are aware of the opportunities to make new products from the data, availability of re-used GI products will be higher, thus maintaining a reasonable level of demand.

Aichholzer and Burkert (2004) discusses a study funded by the Dutch Federal Geographic Data Committee which suggests that by lowering prices of GI would in turn lead to higher turnover and growth in employment. This is a relevant area for this research as we discuss the concern of pricing and the possibility of having a free version of data alongside premium versions. We discuss this further in chapter 5.

Pollock (2008) discusses the Australian Bureau of Statistics where information was given away, showing a significant increase in the usage of data once it was free. This evidence shows that there is a definite trend in making information available for less or for free. Further investigation is needed into what data can be given away for free without sacrificing potential profit or custom.

Pollock (2008) illustrates a possible analogy which can be made between information products and the telecommunications sectors and suggests that both are involved with innovation and new technologies. He also suggests that telecommunications is the route through which most information is distributed and thus making telecommunications fundamental to the distribution of information. Therefore in this research we emphasise the importance to publish data in a format which is easy to distribute but also to consider the other factors which affect the distribution of information.

Other factors such as pricing and how to charge for data are key to the introduction of LD in a potential commercial environment. The conclusions regarding charging regimes made by Pollock (2008) show that the pricing at *marginal cost or below* is most suitable for PSI. This is due to a number of reasons: the high costs of average cost pricing; the high demand for digital data and the benefits to be gained from encouraging users to innovate and make new products from the data available de Vries et al. (2011). Therefore investigation into the potential costs and revenue to be generated from the data is important to ensure the sustainability of a LD market.

A study commissioned by the European Commission and carried out by de Vries et al. (2011) investigated the impact of different models of supply and charging for PSI. The study investigated the different charging models which were outlined in the study by Pollock (2008) and proved that the suggestions and recommendations made in this study have in fact deemed to be true.

This study carried out a number of case studies and found that there was a clear trend towards the lowering of charges and the facilitating of reuse. They found that some organisations were only charging for commercial use and allowing non-commercial use to be free or for a reduced fee.

Of the organisations where there was a reduction in cost recovery, the number of re-users was increased by 1,000% to 10,000%. It also showed that where charges for PSI were reduced it attracted new types of re-users.

3.2 PSI Data Providers

In this section we consider publishers of PSI and the types of data which they provide, what they have been doing with it, how they are publishing and releasing it (including

licensing of the data) and the affordances we recognise they have received as a result of creating their data in LD formats.

A large amount of PSI held in databases has recently been exposed on data.gov.uk and has a huge potential for interlinking with other databases.

We note that although organisations may publish their data, it may not necessarily be in a LD format. We refer back to the LD star ranking system in chapter 2 which ranks data. Data which is merely in a machine readable format is not necessarily in a LD format and may still require some formatting to make it suitable for LD. The main publishers of data on the data.gov.uk website in the United Kingdom include:-

1. Cabinet Office
2. Department for Business, Innovation and Skills
3. Department for Communities and Local Government
4. Department for Environment, Food and Rural Affairs
5. Department for Transport
6. Department of Health
7. Office for National Statistics

Access to PSI differs in the UK to the United States of America (USA) for example. Access to PSI in the USA is unrestricted and the data is considered ‘open’, whereas until recently Europe maintained strict pricing and licensing policies (Aichholzer, 2004; Weiss, 2004). In the USA, where PSI is regarded as open and unrestricted, it is felt that this has contributed to the rapid growth of industries such as the geographic information and environmental service sectors.

Alongside the issue of releasing data for public use, there is also the matter of the licensing of data. Different datasets and organisations will require differing levels of use and recognition for their datasets and therefore licenses which are able to cover this are important to ensure that data can be reused to its full potential. We explore the licensing of PSI later in this chapter.

3.3 PSI and LD

There is great potential for value added products to be created using information and in particular PSI, but one of the key issues with this information is that it is held in many different databases, often in incompatible formats. For example some datasets may be

scanned images of a document stored in .pdf format and others maybe stored in excel databases and although the data may be available, it is more useful to potential users of the information if it is held in standard formats which enable interoperability.

There is a large amount of [PSI](#) held by government agencies and the use of [LD](#) to expose this data is proving to be increasingly beneficial. Research by [Alani et al. \(2007\)](#) demonstrates the suitability of [SWT](#) to unlock various sets of [PSI](#). The Enakting project² has created a number of demonstrations of the power of linked [PSI](#). [Pollock \(2008\)](#) gives an account of the Economics of ‘Public Sector Information’ and illustrates the issues surrounding [PSI](#). Policies regarding access, maintenance and re-use of [PSI](#) have a significant impact on the economy. With the emergence of new technologies such as [SWT](#) and the incorporation of [LD](#), new policies must be adopted to facilitate sharing and re-use of such data on the Web. (See Appendix D)

To enable the sharing and re-use of [Public Sector Geographic Information \(PSGI\)](#), it is necessary to investigate not just the technical issues but also the socio-economic issues. For example the pricing, copyright and licensing agreements held by [PSGI](#) producing organisations ([Giff et al., 2008](#)). Although we state here that this applies to [GI](#), this also applies to non [GI](#) data.

Until the release of the [OS Open Data](#) in April 2010, building an application which used a post code breached copyright laws ([Heath and Goodwin, 2011](#)). This is an issue when trying to link datasets from many different sources as each may have a different licence and even price which can cause difficulties for users who want to access the data. We consider the problem of [LOD](#) and [Linked Closed Data \(LCD\)](#) in [Cobden et al. \(2010\)](#) where we address the issue of how we consume LCD.

Due to the vast array of different datasets becoming available such as crime figures and school performance results, the first action we want to take is to find out where an event took place and by using [LD](#), we are able to illustrate these events on maps. We notice how important it is to be able to pinpoint these events and then link these events to other similar events in the locality.

A lot of data tends to reference location be it full addresses, post codes or grid references. Because of this we notice that location is a useful central point for linking to other datasets. Everything we do has a location and in terms of creating links on the web if we can pinpoint something we do to a location we are able to make further links to other related items on the web ([Hendler and Golbeck, 2008](#)).

We have begun to notice a trend in [LD](#) applications being developed. People are keen to use information released by government and the easiest way to visualise this data is by placing it onto a map background.

²<http://www.enakting.org/>

This opens up an opportunity for revenue to be created from applications developed by individuals who are able to create tools which are useful for others.

3.4 Geographic Information

In the previous section we detailed the different types of [PSI](#) available. In this section we go into more detail regarding [GI](#) in particular. We explore [GI](#) specifically as it is a key component in [LD](#) as it enables us to pinpoint other data to a location on a map.

This section will begin by exploring [GI](#), what it is and the -key technologies and terms used when discussing it. We then go on to explore traditional mapping agencies in the [UK](#) and the [USA](#). We then look at the business of Britain's National Mapping Agency - [OS](#), its key products and how the introduction of [LD](#) will affect it business and look at its current [LD](#) products. We also investigate user generated [GI](#) and how this differs from national mapping agencies and the strengths and weaknesses of both types of data providers.

We notice that [GI](#) is an excludable good and is not a public good until it is in the public domain ([Coote, A Smart, 2010](#)), that is, until the data has been published it remains a private good. The issue which we investigate further here is that organisations such as [OS](#) can publish their data as open [LD](#) under license but there remains the issue of maintaining control of derived data

3.4.1 What is GI?

[GI](#) relates to geography, location, addresses, or a place on the earth's surface. Typical examples of [GI](#) include crime scenes, event locations and property. Gazetteers are used to define indexes consisting of geospatial features ([KK Breitman and M A Casanova and W Tuskowski, 2007](#)).

3.4.2 Why is it expensive/costly?

[GI](#) is a valuable entity; it has various different forms and varying levels of value depending upon its specific user ([Longhorn and Blakemore, 2007](#)). When distributed with added value, it becomes more precious to the holding organisation. Allowing access to this data at the appropriate level is vital. Yet, what is the value of the data held by each individual and how does this value affect their decision to purchase the information? Investigation into the values held by individuals will enable suitable levels of pricing for products to be established that are more suited to users and specifically tailored to the use of [SWT](#) ([Longley et al., 2005](#)).

GI is an expensive commodity from the perspective of the holding organisation (Peuquet, 2002), it has inherent costs associated with its collection and maintenance. GI is used widely across various sectors, for instance private sector organisations require the information for construction, whereas local councils may require information for planning purposes, and the general public may use the information for leisure. Expensive GI is limited to use by only companies who can afford to purchase it or whose business relies on GI to function and therefore has accounted for its cost.

GI is encoded in different formats such as vector (encoded as points and lines) and raster (rendered maps encoded as bitmaps), and its value to a user varies depending on the nature of the user (Longhorn and Blakemore, 2007). The value of GI is subjective; a company relying on GI to do business will value it highly, whereas an individual that could get by without it will value it less. If an organisation or individual cannot afford the asking price of GI, it is effectively without value to them.

If prices of GI remain high, there exists a threat to the organisation generating the GI in the form of competitors offering comparable datasets for free. In the case of the UK national mapping agency the OS,³ the Open Street Map (OSM)⁴ project is already starting to provide a suitable substitute for many consumers of GI. OSM provides user contributed mapping data which is free and updated regularly. For many users this level of mapping is adequate, but for those who require more detailed mapping for building purposes or accurate planning this is not the case.

3.4.3 Key Technologies

A Geographic Information System (GIS) is used to display digital map data and query and analyse the data provided.

A gazetteer is a geographical catalogue that provides an index of geographical features within its scope and coverage. It includes basic information such as shape, location and classification of landscapes.

3.4.3.1 Vector and Raster

In order to view GI using a GIS the data needs to be encoded in such a way that it can be retrieved at a later stage.

Vector and raster are two different methods which are used in order to code geographic data into a computer database. Longley et al. (2005). These two different methods are explained in more detail below.

³<http://www.ordnancesurvey.co.uk/oswebsite/>

⁴<http://www.openstreetmap.org/>

1. Vector

In a *vector* representation, all the lines are captured as points and are connected by precise straight lines. An area is captured as a series of points or vertices connected by straight lines. The vector representation is often called a polygon due to the straight edges between the vertices.

Figure 3.1 shows how the area is captured by a series of points or vertices connected by straight lines.

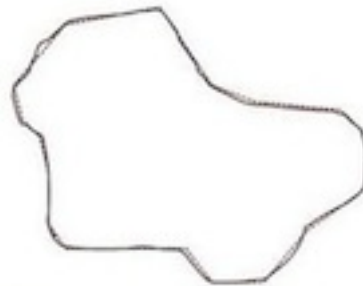


FIGURE 3.1: An area and its approximation by a polygon taken from Longley et al. (2005)

2. Raster

In a *raster* representation space is divided into an array of rectangular (usually square) cells. Any geographic variation is then expressed by assigning properties or attributes to these cells. The cells are sometimes called pixels.

Figure 3.2 shows how each colour represents a different value of a nominal - scale variable denoting land cover class.

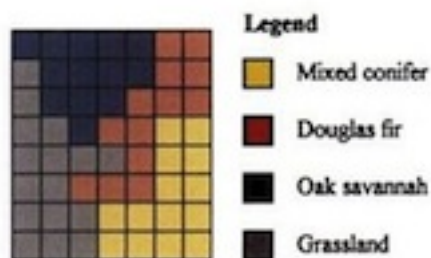


FIGURE 3.2: Raster Representation taken from Longley et al. (2005)

The formats which GI is held bring a new complexity to the application of GI in LD, as modelling data which has different forms is a challenging task and is something which

the LD world has begun to address (Scharrenbach et al., 2012). Some of these issues have been detailed by Utery and Varanka (2011) but raster data in particular is problematic as it is structured as pixel values or digital numbers.

3.5 GI Data Providers

3.5.1 Ordnance Survey

The OS is Britain's national mapping agency (Goodchild, 2012). They produce the most 'accurate and up-to-date geographic data, relied on by government, business and individuals'.⁵ OS holds a monopoly on the market (Gillespie, 2007). A monopoly is where a firm dominates the market and determines the price of its products rather than the price being determined by the market.

OS is a Trading Fund under the Government Trading Funds Act 1973. Despite being government based, OS has the responsibility to earn its own revenue through the distribution and sale of its products. As part of a government agency, however OS must adhere to the specific guidelines laid down by the government.

Before the Open Data movement in the UK, organisations such as OS needed to understand the factors which would affect the organisation in the transition from a high cost low volume market to a low cost high volume market, where large quantities of data can be sold to many consumers for low costs rather than very expensive datasets sold only to a minimal number of consumers. This is due to the way in which markets have changed over time, we illustrate this in chapter 5, looking at the music industry where consumers have changed their spending habits from buying complete albums to individual tracks online, rather than purchasing full albums from physical shops. This trend has been reflected across other areas including the news industry and has come about due to the rise in technology which will enable such transactions to take place. Therefore if this is the direction in which commerce is taking place, LD needs to follow this in order to create products demand.

As the key topic for this research into LD is the GI aspect, we use OS, a Trading Fund to illustrate the potential issues which are involved with the introduction of LD. In order to understand the criteria which OS work under as a Trading Fund we will briefly explain what a Trading Fund is.

Under the Government Trading Funds Act 1973, a Trading Fund is required to recover their costs through income derived from operations within a Trading Fund (Newbery et al., 2008). OS does this by protecting the intellectual property rights over its products and services

⁵www.ordnancesurvey.co.uk

A Trading Fund is part of a government department and its employees are considered civil servants (Bailey, 2006). It receives income from the services and goods it provides and has the advantage that the Trading Fund status allows the organisation to have a more commercial approach to its business.

The OS operates as a Trading Fund under both the Trading Funds Act 1973 and the Ordnance Survey Trading Fund Order 1999. Trading Funds in the UK provide data about a wide variety of subjects, including GI, weather, registered companies and vehicles. The six largest Trading Funds in the UK include Ordnance Survey, the Met Office, The UK Hydrographic Office, HM Land Registry, the Driver and Vehicle Licensing Agency and Companies House (Weiss, 2004).

In order to ensure that OS operates fairly the total income they charge for its data may not exceed the sum of cost of collection, production, reproduction and dissemination and a reasonable return on investment.

OS derived data is any data which has been created using OS base data. For example, if a point on a map was captured and then used as a background to the point on a new image this is considered to be derived data.⁶ OS define derived data as data created by the Licensee that has used Ordnance Survey Digital Mapping Products in its creation.⁷

3.5.1.1 Ordnance Survey Data Products

The digital products at OS contribute towards 90% of its business and the remaining 10% is paper products. This shows a considerable change in types of sales where previously we would notice more sales of paper products.

OS data products are sold in layers, each layer has its own unique common reference a Topographic Identifier (TOID) which allows the layers to be used together, note that the TOID is only used in OS MasterMap. There are over 450 million geographic features in the real world including individual buildings and roads.

In order to establish a means of identifying geospatial features on OS Maps, OS developed a reference called a TOID. Every OS MasterMap feature has a unique identifier which is used to refer to the feature. Key characteristics of the products are those such as complete up-to-date coverage, the seamless data, orthorectified aerial imagery, topographic areas and a topologically structured transport network (Longley et al., 2005). The TOID can be used in for identification in LD products by including it in the URI which we detail in the section on OS LD products.

⁶<http://www.freeourdata.org.uk/blog/?p=256>

⁷<http://www.ordnancesurvey.co.uk/oswebsite/aboutus/foi/questions/docs/PanGovtAg.pdf>

3.5.2 User Generated GI

As well as Great Britain's national mapping agency, there are also other user generated efforts of GI.

[Goodchild \(2007\)](#) outlines the various user-generated efforts towards geospatial information including those such as [OSM](#) and Wikimapia.

The issue of derived data is a key problem with user-generated GI as users may find that the data they use is originally derived from OS and has been inadvertently copied. Therefore, the licensing terms on such user-generated efforts must clearly state from where the data came. The significant issue here is most data reverts back to OS which has strict licensing terms, that is however only for the products which have not been released as Free data. These products do not hold the same strict licensing terms.

A number of issues have been highlighted with availability of user generated content ([Flanagin and Metzger, 2008](#)). One of the key issues we notice which will transfer to the LD world is source credibility or trust as we shall refer to it. When a user uploads content, be it contributions of GI to OpenStreetMap or contributions to Wikipedia, it is the judgement of the user to decide whether they wish to trust the source. We find that this is especially complex when there are many different contribution from many different authors. In the next chapters of this research we will aim to highlight a number of the factors which firstly affect their decision to choose data and then further which will affect their decision to pay for data.

3.5.2.1 Open Street Map

[OSM](#) is a free source of map data that is produced through volunteer efforts ([Auer et al., 2009](#)). Despite being a comprehensible map form, it is not always accurate enough for a users requirements ([Goodchild, 2007](#)). This inaccuracy can be considerably detrimental to organisations requiring precision from mapping products. Examples of end users who would not benefit from the use of [OSM](#) include, the Land Registry and Utility companies. [OSM](#) only can only guarantee accuracy to 10 metres, whereas [OS](#) provides much finer granularity. Therefore for organisations who require the information for plotting boundaries and pipelines for instance the lesser accurate option would not be suitable. We also note that [OS](#) maps display more consistency in their mapping and regardless of the area of the country you look at on an [OS](#) map, you will see the same features as they have created a standard for their maps. [OSM](#) maps however are generated from many different sources and therefore have no official controlled standard to work to, therefore discrepancies in different mapping areas can be noticed ([Mooney et al., 2010](#)).

The key value of [OSM](#) information at this stage is the ability for local people to be able to link the data that they share with local information and knowledge. Individuals,

(such as those who contribute to [OSM](#)) enjoy adding value to such user generate efforts. They notice that their additions are also valuable to others which makes it popular. With community contributions, users are able to notice more value than they would to maps which have not had any contributions ([Goodchild, 2007](#)).

3.5.3 User generated GI vs Traditional Mapping

We outlined earlier the specific details about user generated [GI](#) and [GI](#) produced by national mapping agencies. In this section we explain in more detail the differences between the two. User generated content is produced free of charge by community members. National mapping agencies in this instance, [OS](#) produce high quality mapping but at a cost. There is a cost to organisations producing these maps and there either cover these costs through direct funding from government, or in the case of Ordnance Survey through revenue generate from map sales. In fact there can be an overlap between the two whereby professional organisations such as the [United States Geological Survey \(USGS\)](#) and NASA manage and motivate volunteers to provide contributions. In the case of [USGS](#) this is achieved via the National Map Corp who contribute towards the creation of the National Map (<http://nationalmap.gov/TheNationalMapCorps/>).

On issues of quality we notice that user generated GI has no formal method of checking for quality and accuracy([Flanagin and Metzger, 2008](#))⁸ . OS maps are regularly monitored and formal methods of data collection are used to ensure precision. The data is also collected using standard methods by all surveyors who conform to a precise specification. Whereas the user generated maps are collections of data from various different surveyors, possibly using different standards. This means that there is the chance of differing levels quality of the records made.

With respect to positional accuracy a number of studies have shown that at least with respect to positional accuracy at medium scales (1:10k 1:50k) crowdsourced data can be as positionally accurate as equivalent professionally sourced data [Haklay \(2010\)](#). However, below these scales where positional accuracy really matters there are no crowdsourced equivalents to data such as Ordnance Surveys MasterMap with positional accuracies at the sub-metre level. Here it is therefore generally acknowledge that the professionally sourced data is the most accurate data available. In addition it should be noted that even at the scales where crowdsourced data is collected the comparisons have been made against professional data that has been deliberately generalised and simplified from more accurate and detailed data: for example comparing [OSM](#) with [OS Meridian](#), where the latter dataset has been derived and significantly simplified from the much more detailed and accurate [OS MasterMap](#).

⁸In the case of [USGS](#) and other professional organisations that actively engage the public this is not the case as they contributors have to conform to well-defined specifications and quality controls are applied and well defined. However, for the purposes of this work we will not discuss these further and references to User Generated GI will be specific to purely voluntary bodies.

We do also consider that there is a benefit to maps created by many users. This gives the possibility that the maps may be more up-to date. We use an illustration of the Encyclopaedia Britannica⁹. Until March 2012 the Encyclopaedia Britannica was available as a printed version (Kahin and Varian, 2000). This meant that if certain elements of an entry changed over time, it would not be updated in the encyclopaedia until it was reprinted. This meant that users may end up using out of date information. The information may have been subject to a series of quality checks before publication, but after a certain time it will become out of date. Wikipedia, an online, user generated encyclopaedia, however, has the benefit of being online and can be edited by anyone who has registered to become a contributor. Therefore if someone with domain specific knowledge notices a gap in the content, they are able to add to it or make any relevant changes (Giles, 2005). However, with user generated content, which has no formal policing or checks before it is published and therefore may be left incorrect until someone who knows the area or topic in question notices it.

We also note that in the case of OS, as a national mapping agency they are required to provide mapping of the whole of the Great Britain and as a result may be allowed to access areas which are not public and therefore would not be covered by a user generated mapping agency such as OSM.

Figures 3.3 and 3.4 illustrate a section of a map showing the University of Southampton. The left image is displayed using data collated from OSM and the image on the right shows the same area but with data collated from OS.

We can see from the maps that there are a number of small differences. The first one we notice is that the OSM map show car parks. The OS one however, does not. The OSM image shows the blue P symbol for the car parks but what it doesn't show is that the whole area is a university and therefore the car parks are not public car parks and require a parking permit. We also see from the images that the OSM map does not include all of the buildings above Burgess road at the top of the map. This is an inconsistency which we notice in OSM and which we turn to the OS version of the map for a more precise images of the local area.

The OSM map however shows Southampton common and the various routes through it, but, the OS map gives more detail about the tracks and specific areas on the common. We also notice that the OS map outlines each specific building in the area but the OSM map just shades in full areas which contain buildings. This is where we notice that the OS maps contain much finer detail about the structures which exist and which could only possibly be generated by a national mapping agency. The area shown on the map above Burgess road does not show any buildings and to a new user or someone who is not familiar with the area this may be misleading.

The data found in OSM is not complete or consistent across the whole of the UK

⁹<http://www.britannica.co.uk>

and there are no official and thorough quality assurance processes as part of the data collection process (Mooney et al., 2010; Haklay, 2010; Haklay et al., 2010)

For a leisure user the OSM map gives a suitable map to use for a rough guide to an area and does have the advantage for example of displaying car parks but for a user who may require a higher level of detail, i.e. for the commercial environment, this may not be reliable enough and they may prefer to refer to well trusted source such as a national mapping agency.

The car parks for instance are shown on the OSM map on the left but they are not shown or highlighted on the OS map on the right. The OSM image fails to provide information that the car parks shown are in fact permit holder only car parks and not public car parks which some users may find misleading.



FIGURE 3.3: Comparison of OSM vs OS - This image show a snapshot of an OS map

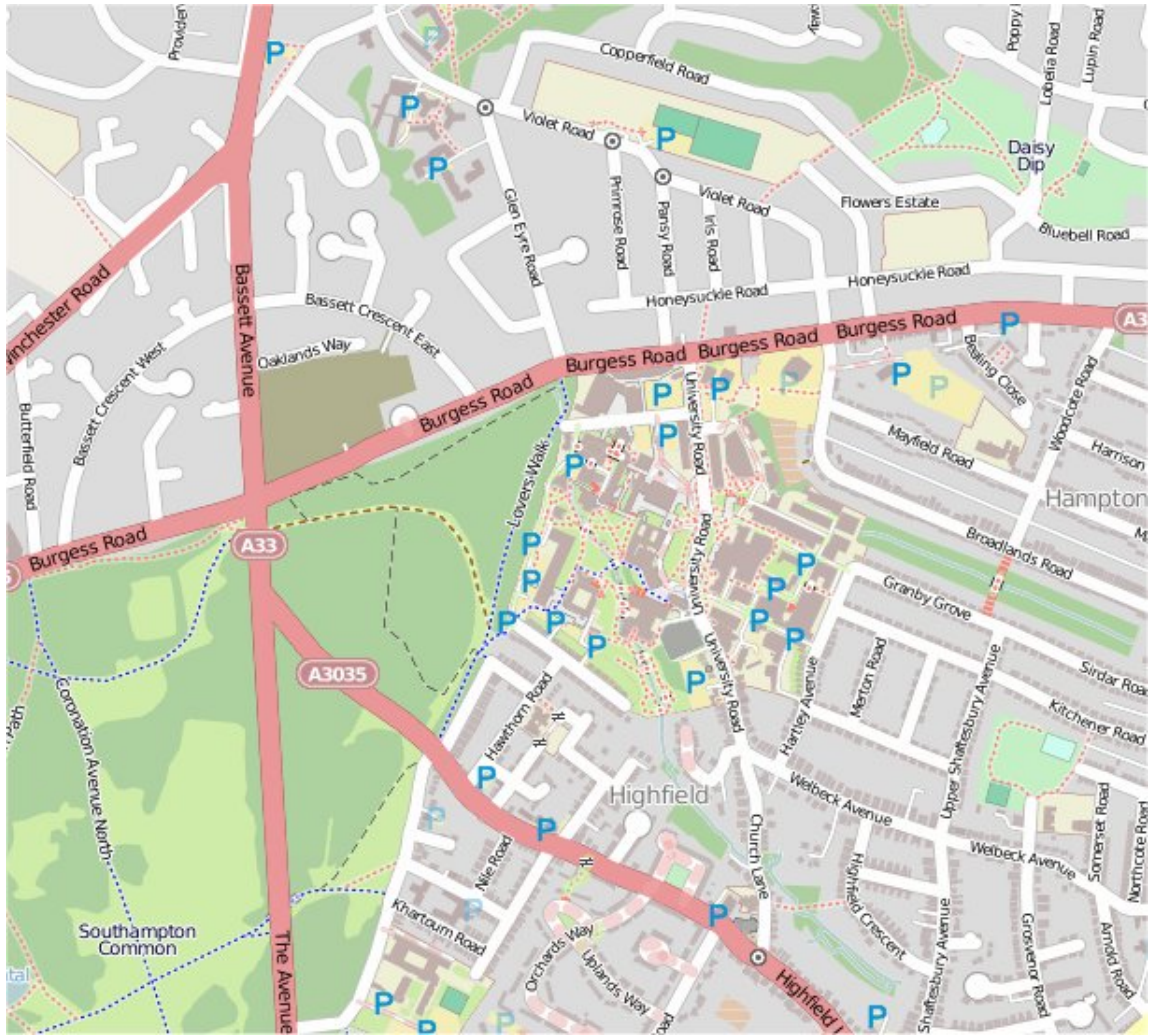


FIGURE 3.4: Comparison of OSM vs OS - This image shows a snapshot of an OSM map and illustrates the same area as shown on the OS map above

We note work carried out by [Haklay et al. \(2010\)](#) has suggested that the more people who contribute to the same area on [OSM](#) are more likely to increase the validity of the map due to more people marking coordinates for specified landmarks in the area. Once a certain number of contributions per area have been recorded the quality of this areas often decreases, thus reducing the overall quality of the map. There still lies the issue of consistency as there may be areas which have had no coverage at all where some areas may have received more coverage than others ([Heipke, 2010](#)). This leaves it uncertain that the point on the map you wish to view is still left with accuracy or completeness issues.

However [Haklay \(2010\)](#) does conclude that [OSM](#) has better positional accuracy than [OS Meridian 2](#) but is less accurate than [Mastermap \(Chilton, 2009\)](#). Therefore we suggest that for the purpose of this thesis where we will use [MasterMap](#) as the premium data source, [OS](#) data or the ‘premium’ version of data, which has had quality assurance checks and has full coverage of all geographic areas.

3.6 Conclusion

In this chapter we have investigated the different types of [PSI](#) and the organisation which produce it. We then go on to explore [GI](#) as an important type of [PSI](#). [GI](#) is a strong topic itself but is also considered a sub topic of [PSI](#). We note that [GI](#) has a central focus for the linking of data as it provides a location for other data to be attached or linked to. With more government produced [PSI](#) available, it becomes easier to create ‘mashups’ of the data. As data is available in similar formats it makes it easier for correlations in the data to be found and gives the potential for more value to be generated from applications which display not just one dataset but potentially more. Although it would require more effort to get it into the [LD](#) format initially (by transformation into [LD](#) formats) the overall benefit (value) to the organisation following this will be more valuable products.

The previous chapter has detailed the key topic technologies ([LD](#)) for this research and the importance of [GI](#) with this technology. With these topics in mind, the next chapter outlines the business of linked [GI](#) and details the specific elements required for analysis into generating value from linked [GI](#), we look specifically at value, revenue and willingness to pay for products online.

The next chapter examines the [UK](#) national mapping agency [OS](#) and outlines its contribution to [LD](#)

Chapter 4

Ordnance Survey Use Case

As the National Mapping Agency for Great Britain, much of its data is relied upon by government bodies and agencies. For instance there the emergency services use OS for route planning (Beaumont et al., 2005) and it is also used by local councils for planning and certain types of private sector areas such as insurance. It is widely used by many other groups including hobbyists and enthusiasts, therefore we consider it to have a wide spectrum of uses and has plenty of depth to be able to explore different areas of its business. If OS is to operate in an environment where data is linked, it is important that we are able to understand how the different products within the company are structured and the formats in which they are made available to its consumers. From this we are then able to draw some conclusions about how the data is to be used in a LD environment.

4.1 Ordnance Survey Vector Data Products

The two types of data raster and vector data are the basis for the different types of data available from OS and these products are outlined in more detail below. OS MasterMap is the main and most detailed product range produced by OS. It is built up in layers which can be added and which contain different resources such as post codes and imagery. Table 4.1 illustrates the OS MasterMap® products and the layers of MasterMap which are available. The OS MasterMap® is a digital product range that contains information structured into different products which are called layers. These layers can be purchased separately or with other layers and used together. Further Vector data products which are available from the OS and are outlined in table 4.2.

Product Name	Product Description
Topography ¹	A detailed, geographic database containing almost half a billion features. Surveyed to a high degree of accuracy. The Topography Layer forms the foundation of the OS MasterMap ®
Address 2 ²	Precise coordinates for over 27 million residential and commercial properties in Great Britain (GB) . Originates from Royal Mail's postcode address file. Coordinates for each address are determined through the use of on-the-ground Global Positioning System (GPS) survey and aerial imagery. Postal and topographic geography is joined up which creates a fixed link between the property and its address.
Prebuild Address ³	A dataset which provides consistent and comprehensive address information for England, Scotland and Wales. Future builds and approximate spatial location across GB are identified.
Imagery ⁴	A maintained dataset of high quality aerial photography of GB . This layer has seamless coverage of GB in high-resolution and accurately reflects the position of features at ground level.
Integrated Traffic Network™ ⁵	A detailed and up-to-date digital dataset consisting of a Roads Network and a Road Routing Information (RRI) theme for GB . Contains all types of road categories from unnamed minor roads to motorways. RRI also contains many features such as the height, weight and width restrictions; traffic calming and one-way roads.

TABLE 4.1: OS MasterMap Products

Product Name	Product Description
1:50 000 Scale Gazetteer	1:50 000 Scale Gazetteer provides a reference tool or location finder, allowing location of areas of interest. The gazetteer can also be used to navigate around the map, geocode data or create lists of places within a specified area.
ADDRESS-POINT® ⁶	A dataset that uniquely defines and locates residential, business and public postal addresses in GB. Each address has a unique Ordnance Survey Address Point Reference (OS-APR). This product is slowly being phased out and is being replaced by Address Layer
Boundary-Line™ ⁷	A specialist 1:10 000 scale boundaries dataset. It contains all levels of electoral and administrative boundaries, from district, wards and civil parishes up to parliamentary constituencies.
Code-Point® and Code-Point® ^{8 9}	Code-Point provides a precise geographical location for each postcode unit in the United Kingdom. Code-Point® with polygons is produced by tessellating individual address records from ADDRESS-POINT® then nested within the sector boundaries prescribed by Royal Mail®. These polygons enclose every fully matched address in the correct boundary and are more accurate than previous products.
Land-Form PROFILE® Plus and PANORAMA® ^{10 11 12}	Land-Form PROFILE provides detailed height data defining the physical shape of the landscape of GB. It provides a consistent foundation for 3-D modelling applications, to maximise the potential of information.
Meridian2™ ¹³	Meridian™ 2 is a mid-scale digital representation of GB that allows customisation of its transport network and topographic themes, allowing the user to create geographic solutions for their business needs.
Points of Interest ¹⁴	Points of Interest is a dataset of around 3.9 million geographic and commercial features across GB. These highlight location and function information, with a postal address for all postally addressable Points.
OS VectorMap™ Local ¹⁵	OS VectorMap Local is a flexible product that helps users to visualise information on a map. It enables users to customise the look and feel of their map, incorporating their own information.
OS Sitemap® ¹⁶	This product provides customers with extracts of Ordnance Survey mapping in a number of different formats and to different scales. Developed to suit the requirements of a broad range of customers - from private individuals requiring paper map copies for planning applications to architects and engineering businesses wanting electronic map data to be used for a development project.
Strategi® ¹⁷	Strategi is detailed digital map data, used for applications requiring an overview of geographical information. Geographical features within Strategi are represented as vector data, enabling users to link business information to relevant features on the map for planning purposes, analyse trends or create simplified routing information.

TABLE 4.2: OS Vector Data Products

4.2 Ordnance Survey Raster Data Products

The Raster data products ¹⁸ which are available from the OS and are outlined in more detail in table 4.3.

Product Name	Product Description
1:10 000 Scale Raster	The 1:10 000 Scale Raster map data is the most detailed product in the raster portfolio, providing large-scale background mapping upon which information can be added or overlaid
1:25 000 Scale Colour Raster	One of a range of backdrop mapping products, Ordnance Survey's 1:25 000 Scale Colour Raster is backdrop map data of the OS Explorer Map series for outdoor activities.
1:50 000 Scale Colour Raster	Provides a comprehensive map base for detailed work where street names are not required, such as demographic analysis.
1:250 000 Scale Colour Raster	1:250 000 Scale Colour Raster map base combines roads, railways and other key features to make a cartographic backdrop for overlaying business information.
Historical Map Data	An archive was scanned to create Historical Map Data. National cover available, dating back to the 19th century and derived from 1:10 560, 1:10 000 mapping & 1:25 000 scale County Series, post-War National Grid, and superseded mapping that includes 1:1250 & 1:500 scale Town maps.
MiniScale® (1:1 million nominal scale)	MiniScale is a small-scale product, nominally at 1:1 million scale, designed for use within desktop graphic applications to provide simple backdrop mapping covering the whole of Great Britain.
OS Landplan® Data	OS Landplan Data is the largest scale of Ordnance Survey raster data to show contours, providing an overview of the lie of the land. Fences, field boundaries, road names and buildings are also included.
OS Street View®	OS Street View is street-level, backdrop map data that is designed for online applications. It provides a scanned image of street-level mapping that can be combined with other data in a geographical information system (GIS), enabling visualisation of a wide range of information.
OS Locator™	OS Locator is a fully searchable national gazetteer for use with Ordnance Survey's range of mid-scales raster map data products.

TABLE 4.3: OS Raster Data Products

¹⁸<http://www.ordnancesurvey.co.uk/oswebsite/products/>

4.3 Ordnance Survey OpenSpace

OS OpenSpace gives free access to the same detailed data available in the paid for versions of data. It enables non-commercial users to embed the maps into public websites and use the data for leisure use or commercial users to experiment with the data before making a purchase.

Users are required to register for their own [API](#) key which asks them to accept the terms of the OS OpenSpace Developer Agreement. It is also a requirement that the [URL](#) of where the map is to be used is provided when the registration takes place. The OpenSpace maps are then available to users via the Web Map Builder which enables non technical users to select the maps they require and embed them into their website.

There are daily limits on the use of the data which are aimed to prevent over use of the free product and give users an idea of where the boundary lies between free use and use which requires payment.

- 65 000 tiles of mapping data in a 24-hour period.
- 1 000 Place name look-ups (Gazetteer service) in a 24-hour period
- 1 000 Postcode look-ups in a 24-hour period
- 1 000 Boundary look-Ups in a 24 hour period

4.4 OS OpenSpace Pro

Further to OS OpenSpace, which was designed to promote experimentation with Ordnance Survey datasets and available for use by anyone including commercial organisations, the Pro version provides businesses and developers access to detailed map data for Great Britain. This service is available Free of Charge for 90 days and then after this the following charges apply. Table 4.4 4.2 displays the pricing model for OpenSpace Pro. The model is made up of the data royalties which are determined by the relevant Partner Contract plus a service charge for the volume of data supplied.¹⁹

¹⁹<http://www.ordnancesurvey.co.uk/oswebsite/web-services/os-openspace/pro/pricing.html>

Charge	Relevant Contract/Service	Amount	Minimum
Access Fee	Annual fee for access	£1,900	N/A
Royalties	Payable as standard under the relevant Data Contract	Dependent on Data Contract	N/A for OS OpenData™, As set out in Framework and Contract, N/A for OS OpenData™
Session charge	Consumer Applications and Websites Contract (standard on-demand), Viewing, Tracking and Scheduling Contract (on-demand view/track-/schedule)	£0.005 per session	Minimum Service Charge of £500 each Contract Year
Subscription charge	Consumer Applications and Websites Contract (standard on-demand)	£0.60/month per subscription	Minimum Service Charge of £500 each Contract Year
Data Volume charge	Consumer Applications and Websites Contract (premium on-demand and software packages), Navigation Contract (on-demand and software packages), Viewing, Tracking and Scheduling Contract (on-demand and software packages, tracking and scheduling, but not viewing), OS OpenData™ (where no contract is in place)	£0.30/Gigabyte downloaded	Minimum Service Charge of £500 each Contract Year

TABLE 4.4: OpenSpace Pro Pricing and Terms

4.5 Ordnance Survey Linked Data Products

Since the **LD** concept has really gained momentum **OS** has begun to develop a number of its existing products into **LD** formats. The products are divided into vector and raster products and point data products. All of these products are available from the **OS** website via download for free or via a medium such as a compact disc which carries a small fee.

Of the products which **OS** offers which we have outlined earlier, the ones listed below have been released as part of **OS** OpenDataTM. These products are available under the **OS** OpenDataTMLicense which is outlined in Section 4.5.²⁰

- OS VectorMap District
- 1:50 000 Scale Gazetteer
- 1:250 000 Scale Colour Raster
- Boundary-Line
- Code-Point Open
- Land-Form PANORAMA
- Meridian 2
- Miniscale
- OS Locator
- Strategi
- OS Street View

We notice that some of the products available from **OS** are more relevant to **LD**. The vector products are more suitable to being specifically defined using **LD** as the polygons, lines and coordinates can be encoded into **LD**. References to raster products may be possible but not necessarily suitable for referencing the finer details.

We note that the raster products available from **OS** are not suitable for **LD** due to the format of the data being hard to encode as **LD**. We note that the MasterMap products of which are vector are considered to be the most suitable for the purpose of **LD**.

The three key **LD** products available are:

- 1: 50 000 Gazetteer

²⁰<http://www.ordnancesurvey.co.uk/oswebsite/docs/licences/os-opensdata-licence.pdf>

- Code Point Open
- Administrative Geography Gazetteer for Great Britain

OS began its portfolio of LD products by producing a gazetteer of the administrative regions of Great Britain. This gazetteer is the LD version of the Boundary Line Open-Data product. A unique identifier in the form of a URI is given to each region and this is described by its name and relation to other regions. We use the example of Hampshire. If a user chooses to explore Hampshire as a county, they may decide to enter a search term into a web browser and this search may return a list of the counties which are adjacent to it for example, Surrey and Berkshire. It may also places which are contain in Hampshire such as Winchester.

Following the production of the 1:50 000 gazetteer, OS produced further data which contain URIs for every postcode in the country. This dataset is identified as Code-Point. This data was then linked to the URIs produced for administrative regions. We use the university as an example of how the post code is used in the URI:

<http://data.ordnancesurvey.co.uk/id/postcodeunit/S0171BJ>

The release of these products enable users to begin to experiment with LD and create applications and make use of data which is linked.²¹ In order to clearly illustrate the potential of linking GI, a demonstration of the possibilities should be created, in order to prove this to non-technical users. There are already some efforts which have been produced where companies are beginning to create revenue from public data, in particular Local Authorities spending data. For example Agresso is producing LD on local authority spending.²²

The introduction of the data.gov site has enabled users to see the type of datasets available and gives them the opportunity to explore the potential of linking these datasets together. We note here that although there is a large amount of data which has been published on this site, not all of it is available as specifically LD. Data which is already in a machine readable format can easily be translated into a LD format.

Without publishing the data and enabling people to see what data is available meant that people were not able to visualise or begin thinking about the possible benefits. Now the data is available it has enabled people to begin seeing and experimenting with the data and creating more useful datasets which others can use or adapt. With the release of OS data people are able to create mashups of data which uses GI as its focal point. For example people can create mashups of data which contain crime figures for local

²¹<https://www.ordnancesurvey.co.uk/opendatadownload/products.html>

²²<http://www.unit4software.co.uk/about/news/art?aid=3746>

ares and pinpoint the crimes to post code areas. This can then be extended to health and education figures which can be added from different datasets.

Table 4.5 shows the datasets which have linked to OS data. These datasets are accurate up-to September 2011. We created this table by following the links from the OS in the Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch.²³ From the links we looked at the data which was attached to OS data and took the number of links to the OS LD from the website <http://datahub.io/dataset/>.

²³<http://lod-cloud.net/>

Name	Description	Source	Number of Links to OS LD
Street level crime reports for England and Wales	The data presented here is a linked data representation of the street-level crime reports first released for England and Wales in 2011	http://crime.rkbexplorer.com	684394
EnAKTing Population Dataset	Data extracted from the census data provided by UK's Office for National Statistics grouped by parliamentary constituencies.	http://population.psi.enaktng.org	533
EnAKTing Mortality Dataset	UK's Mortality data per region for the year 2008/09, delivered by the UK Home Office.	http://mortality.psi.enaktng.org	399
EnAKTing Crime Dataset	UK Crime statistics per region for the year 2008/09, provided by the UK Home Office.	http://crime.psi.enaktng.org	133
Openly Local: Making Local Government More Transparent	The project provides a unified way of accessing UK Local Government information.	http://openlylocal.com/	13500
statistics.data.gov.uk	Linked data about administrative areas used within UK government official statistics.	http://statistics.data.gov.uk/	32284
Metoffice Weather Forecasts	Weather forecast data scraped and converted to Linked Data	http://metoffice.dataincubator.org	413
Yahoo Geoplanet RDF	This is a Linked Data version of the publicly available data dumps from the Yahoo! GeoPlanet database.	http://kasabi.com/dataset/yahoo-geoplanet	9591825
UK Postcodes	Linked data for every UK Postcode.	http://www.uk-postcodes.com/	3392176
UK Legislation	API access to UK primary and secondary legislation.	http://www.legislation.gov.uk/	16000000
EnAKTing Energy Dataset	UK's Road Transportation consumption data provided by the UK Department for Business, Enterprise and Regulatory Reform (BERR).	http://energy.psi.enaktng.org/	368
Renewable Energy Generators	The Renewable Energy Federation maintains an online database of renewable energy generators located in the UK.	http://kasabi.com/dataset/renewable-991-energy-generators	991

TABLE 4.5: Linked Datasets which use OS Data

There are a number of other datasets which have linked to OS which incorporate OS data. These include:

- The Stationary Office (TSO), have used OS boundary data and London Gazette corporate insolvency data for a mash up showing information on firms entering insolvency mapped on to council ward, local authority.²⁴
- In the LOCAH Linked Archives Hub dataset,²⁵ links are made using the OS vocabularies from archival repositories (as places) to the OS Postcode Units within which they are located, e.g.<http://data.archiveshub.ac.uk/id/place/repository/gb96> is linked to <http://data.ordnancesurvey.co.uk/id/postcodeunit/WC1E7HU>
- OS data is used in the LUCERO project,²⁶ to get Postcode information of the university buildings. For example: <http://data.open.ac.uk/page/location/building/r05notb>. This enables the extractions of Latitude and Longitude details, enabling buildings to be plotted onto a map.
- OS data is also key to the Data Enrichment Service which automatically adds linked data to text.^{27 28}
- <http://data.southampton.ac.uk/bus.html> uses the postcode URIs and RDF provided by OS to easily resolve locations of postcodes to let people find nearby bus stops.

As we can see from the examples above there are a number of cases of use of the data, but at this stage we do not see any commercial usage of the data. We emphasise here the difference between the value of raw data and the value of data which can be made into an application and the value of the application per say and not the raw data itself. When data is released for free, it is hard to determine how much of the data is being used for different purposes. For instance, it is very hard for OS to discover what the data it released for free is being used for as some users may not have disclosed the reasons for their download, therefore we are unable to gain a true understanding of the actual usage at this stage.

4.6 Licenses

The OS Open datasets available free of charge are released under the OS OpenData™ License which has been created specifically for OS OpenData™.²⁹ The data sets include raster

²⁴<http://openup.tso.co.uk/developer/demos/insolvency>

²⁵<http://archiveshub.ac.uk/locah/>

²⁶<http://lucero-project.info/lb/>

²⁷<http://openup.tso.co.uk/content/images/7112%20OPENUP%20Info%20Sheet%231.pdf>

²⁸<http://openup.tso.co.uk/des>

²⁹<http://www.ordnancesurvey.co.uk/oswebsite/docs/licences/os-opendata-licence.pdf>

and vector mapping, height, boundary and gazetteer products. The license is outlined in Chapter 5.6.

OS also offers trial versions of its products for potential commercial use under 3 different license types which enable users to try the data for free.

- **Discover Data Licence** - A free sample of data is distributed under the terms of the Discover Data Licence to give an indication of what the data will be like. Free samples of all OS business products can be downloaded under this licence.
- **Evaluation Licence** This is for new or existing customers who would like to take a larger area of OS data to evaluate, test or demonstrate internally for a period up to 3 months.
- **Developer Licence** If developers have or are developing a new product or service that will use digital mapping this licence enables them to develop, test and demonstrate OS data to potential customers.

We can see from the licenses available here that OS is offering free trials of its products in order to give users a chance to work with the products before they buy them. This demonstrates a free-trial pricing model which we outline in more detail in the next chapter.

4.7 OS MasterMap Pricing

The price of OS data products are determined by the areas selected, users are able to select a predefined area of interest and there is a minimum charge for each order. The number of terminals required for the data to be used on and duration of the contract is taken into account in order to calculate the final total price. We outline how each layer is calculated in Table 4.6 as of September 2012. There is a discount available if the product is to be used on more than 101 terminals.³⁰

³⁰<http://www.ordnancesurvey.co.uk/oswebsite/docs/ordnance-survey-business-portfolio-price-list.pdf>

Product Name	Pricing Structure
OS MasterMap Topography Layer	Based on a 1 km by 1 km classification of geography type that covers the extent of Great Britain. Each square kilometre is allocated to be one of three geography types and priced accordingly for a one-year contract for use on 101 or more terminals. Orders for less than 1 km ² are priced according to the underlying geography.
OS MasterMap Imagery Layer	Based on a single flat km ² price, calculated individually. The first km is £54.40 per km ² . The next 24km ² is £12.00, the next 9975 km ² is £5.44 and then each subsequent km ² is £0.76.
OS MasterMap Integrated Transport Network (ITN) Layer	These products are priced using a km ² density model created for the Roads Network theme. This theme can be ordered and used as a single theme. The Road Routing Information theme is only available in conjunction with the Roads Network theme and cannot be used independently of it.
OS MasterMap Address Layer 2	OS MasterMap Address Layer 2 links any property address to its location on the map. It provides precise coordinates for over 27 million residential and commercial properties in Great Britain. OS MasterMap Address Layer 2 is overlaid on OS MasterMap Topography Layer. In order for customers to use this layer, they have to complete a form in order to comply with the license. This is the Royal Mail® Multiple Residence Data customer registration form. The price is calculated by establishing the number of addresses found in the required dataset. The first five million addresses are £0.0148, the next ten million addresses are then £0.0074 and then any additional addresses are £0.0038.
OS MasterMap Address Layer	Prices are calculated using the number of addresses in the dataset. The number of addresses in the area of interest are added up and priced as follows for a one-year contract for use on 101 or more terminals: First five million £0.0102, Next ten million £0.0051, Additional addresses £0.0026. For a one year contract covering the whole of Great Britain on over 101 terminals is £130600.
AddressBase	Similar to Address Layer the whole of Great Britain on over 101 terminals can be purchased for £129 950 or addresses can be purchased individually. Prices are calculated using the number of addresses in the dataset. The first five million £0.0080, the next ten million £0.0051, any additional addresses are £0.0030
AddressBase Plus	The total price for a one-year contract covering Great Britain for use on 101 or more terminals is £175 000. Or the addresses can be purchased individually. The first five million addresses £0.0108, the next ten million addresses £0.0068 and any additional addresses £0.0031.
AddressBase Premium	The total price for a one-year contract covering Great Britain for use on 101 or more terminals is £189 370. Or the addresses can be purchased individually. The first five million addresses £0.0116, the next ten million £0.0074 and any additional addresses £0.0074.

TABLE 4.6: Product Pricing

There are additional terms with the use of the AddressBase products. If the user displays any AddressBase product on a publicly-available website. I.e. they do not just use the addresses internally for commercial purposes then there is an additional annual fee of £4000. Terms for central government departments who license the Great Britain coverage on 900 or more terminals will also be charge an additional annual fee. More detail on the terms and prices for the OS products is available from the OS website and are available for the public to see. The prices we have listed here are correct as from September 2012 until September 2013.

The Web Map services called OS OnDemand have separate prices and are shown in Table 4.7 below. There is a minimum term contract here which is one year

Band	Terminals	WMS and WMTS	Internal and external serving	Includes OS MasterMap Topography Layer	Price per annum
A	<100	WMS only	Internal only	Yes	£1 500
B	1011 000	WMS only	Internal only	Yes	£5 000
C	1 0012 000	WMS only	Internal only	Yes	£9 000
D	Unlimited	Yes	Yes	Yes (WMS only)	£20 000
E	Unlimited	WMTS only	Yes	No	£6 000

TABLE 4.7: OnDemand Pricing

The next product, Address-Point also has its own pricing structure. This is outlined in Table 4.8

No of terminals	Licence fee Great Britain	Licence fee Government Office Regions
101+	£132 500.00	£13 250.00
51 to 100	£119 250.00	£11 925.00
21 to 50	£106 000.00	£10 600.00
11 to 20	£79 500.00	£7 950.00
6 to 10	£59 625.00	£5 962.50
3 to 5	£39 750.00	£3 975.00
2	£26 500.00	£2 650.00
1	£16 562.50	£1 656.25

TABLE 4.8: AddressPoint Pricing

Table 4.9 displays the pricing structure for the CodePoint product.

We can see that a discounted rate is applied the more of the product is purchased however there are prices for singular purchases of certain products. Due to the nature

No of terminals	Licence fee
101+	£5 852.75
51 to 100	£5 267.48
21to50	£4 682.20
11to20	£3 511.65
6 to 10	£2 633.74
3 to 5	£1 755.83
2	£1 170.55
1	£731.59

TABLE 4.9: Codepoint Pricing

of this pricing we can see that it may be in a suitable structure to form the basis for a pricing structure for LD. This can be applied in the architecture which we outline in detail in Chapter 8 where the links can be used to restrict access to certain features and the prices for the products here could be translated into the data to produce restrictions to important or valuable datasets. The prices shown in these tables have been carefully calculated by the Pricing and Licensing department within OS to ensure that all aspects of a potential product purchase have been covered.

4.8 Pricing changes since 2008

In order for us to observe any price change on OS products over time, we use firstly Address Layer 2 as an example for comparison and then Address-Point.

In 2008 a minimum fee of £500 for an annual contract was applicable for the use of each individual layer of OS MasterMap, this has remained the same in the 2013 Business Portfolio Price list.

In 2008 the first 5 million TOIDs were priced at £0.0148 and in 2013 this was also priced at £0.0148. The same applies for Address-Point, in 2008 the License fee for the product on 1 terminal for the whole of Great Britain was £16,562.50 and in 2013 it was also £16,562.50. Therefore we see that there is no difference in price over five years, so what has changed?

Although the core pricing for products has not changed, we notice that there are alternative pricing mechanisms in place for smaller, non commercial users. For instance Ordnance Survey Getamap is available with a number of different levels of subscription.³¹ These options include a 12 month, 3 month and 1 month subscription option. Getamap³² allows users to select an area of the map they require and print it under their

³¹<http://www.shop.ordnancesurveyleisure.co.uk/products/digital-maps/digital-maps-for-get-a-map/digital-maps-for-get-a-map-getamap-subscription>

³²<http://www.shop.ordnancesurveyleisure.co.uk/products/osdigital-maps/digital-maps-for-get-a-map>

subscription allowing them access to OS data at the level they require without high cost.

We can see from the data outlined above that the pricing of the OS product ranges are highly complex. Each product has different features and potential uses, so the prices have been calculated to ensure that OS adheres to the strict terms of being a Trading Fund, which is not to make a significant profit from the sale of its products.

As with any commercial organisation, the ways in which OS calculates the prices for its products is held in strict confidence within the organisation and is not publicly available. We have however demonstrated in this section the levels of pricing which have been adopted and the products to which it applies. This gives us an understanding of the different pricing structures which apply for each product. We also note here that the usage of the data products, for example before and after new releases, is also confidential and not available to the public.

4.9 Conclusion

In this chapter we have provided a background to the business of OS and demonstrated the potential for the application of LD. This chapter also provides a use case for LD and looks at the different types of products provided by OS. We have then outlined the pricing and licensing of these products to enable a view of the possible issues which may arise. These issues include quality of user-generated content against a national mapping agency such as OS. We also note that pricing is a complex area of discussion and that OS in particular has very complex pricing regimes. The need to explore pricing and its counterparts including value is addressed in more detail in the next chapter.

Chapter 5

The Business of Linked Data

In the previous chapter we explored the potential for [LD](#) and [GI](#) using [OS](#) as a potential use case. We explored the power of [GI](#) and outlined the business carried out by [OS](#) and why investigation into the [SW](#) is important for organisations in a wider context.

In this following chapter we will introduce the business and economic side of the technology. Specifically, we investigate the issues which arise when investigating new models for generating value from [LD](#) on the web and how this is relevant to [GI](#). We investigate new revenue models as the current models of generating revenue are constantly changing. With a new concept such as [LD](#) we must ensure we are aware of the factors which will contribute to the success of the technology in the future.

We outline the characteristics of information goods which differ considerably from tangible goods and then investigate the economic factors surrounding [LD](#). Having detailed the economic factors of such products we investigate the concept of willingness to pay for information goods. Following this we, discuss similar industries to the [LD](#) industry and draw together the review in a summary of our findings. This chapter sets out the framework for our further empirical research and enables us to ask questions regarding the business case for [LD](#).

It is important that we outline all of the concerns regarding the business case for [LD](#) now so that the empirical investigations we have carried out in the latter chapters are informed to form the technical framework for the consumption of [LD](#) as the final aspect of this research.

The common assumption is that [LD](#) is a free commodity and that all organisations and individuals will publish their data free of charge. We do however, need to address the fact that not all data can be made free and there are organisations who will publish their data but will need to expose this data for a fee in order to recoup its costs. [OS](#) is a clear example of an organisation which holds a vast amount of data which when published as

LD will be of great interest to its consumers. Furthermore, OS is a Trading Fund and therefore needs to make a revenue from the products it creates and maintains.

Latif et al. (2009) suggests a value chain for LD. They suggest three different types of LD providers. There are raw data providers such as the BBC and Wikipedia. There are specific LD providers such as MusicBrainz and DBPedia and then further to this there are LD application providers such as the BBC. The difference between these are DBPedia for example just provides the data where as an application provider such as the BBC provides a potential ‘mashup’ of many datasets in a format which is visually attractive to the end consumer.

We consider other organisations who are publishing their data for free, including the BBC which we discuss later in this chapter. The Met Office¹ has released some of its forecasts as LD and MusicBrainz² which is an online music encyclopaedia, have contributed data which can be linked to from other data sources.

We are yet to see the true value organisations are getting from LD, as the initial costs of producing data in LD formats is expensive and there are issues which we address later, including trust, provenance and value to consumers, that will affect the value of LD to its producers.

This chapter will address this issue in more detail as we are aware there will become a situation where there is linked open data (data which has no cost or licensing restriction, note that it must still have a license even if it is an open license) available alongside linked closed data (data which may be charged for or has restrictive licensing).

5.1 Excludable vs Non Excludable, Rivalrous and Non Rivalrous

In this section we discuss the concept of excludable and non excludable goods in reference to digitally available goods on the web. We also detail what is meant by rivalrous and non rivalrous. As we will explore in the next section, there are public, private, club and common goods. Each of these in their own right is either excludable or non excludable. Firstly we define for the purpose of this research, what we mean by a ‘good’. A good is merchandise or a product which a consumer can purchase.

In traditional economics, most goods are rival and excludable, that is one person’s consumption of a good prevents another person’s consumption of the same good and therefore reduces the amount of the product available for others (Kahin and Varian, 2000). Information and digital products however, do not have these same features and

¹<http://thedatahub.org/dataset/data-incubator-metoffice>

²<http://musicbrainz.org/>

large numbers of consumers can download the same set of data with no impact on other consumers. [Kahin and Varian \(2000\)](#) also suggests that as a result of this traditional pricing, models which have existed in traditional markets with non digital products are no longer suitable to sustain profitability for the digital market and therefore new models must be explored. We outline these models in more detail in this chapter where we discuss the business of Linked Data and introduce the idea that we may not only have [LOD](#) but also there is the potential and the need for [LCD](#).

A rivalrous good is a tangible good where one person's use of that good prevents others from using that same good at the same time. We give the example of a domain name. If one user is using that domain name at a certain time, another cannot. However, if the user gives up that domain name, it can be used by someone else at a later time.

Contrasted to rivalrous goods are non-rival goods; where the goods can be consumed by many people simultaneously and it will not prevent another from using it. We give the example here of a website, many users can view the same website at the same time.

5.2 Public, Private, Club and Common Goods

To understand the type of products we are looking at we explore different types of goods in economic terms. There are four types of goods available which are public, private, club and common. We will outline each individually but will explain in greater detail, public goods which is the basis to this research. A public good is one which is also non-rivalrous, that is the consumption of the product by one consumer does not affect the consumption of it by another consumer ([Fraser, 1996](#)). It also has the characteristic that individuals cannot be excluded from its consumption. For example street lighting, or clean water or air. A public good however can be subject to restrictions which then make it a club or private good via copyright or paywalls. This is where we notice that [PSI](#) may encounter restrictions. By the use of [LD](#) standard however, we feel that it has become more accessible and if additional features are incorporated into the product making it more valuable then here lies the opportunity to create restrictions to its use via paywalls etc. This is where we begin to notice that there is a possibility for [LOD](#) alongside [LCD](#).

A private good is excludable and its consumption is rivalrous ([Meyer, 2010](#)). That is, a consumer is able to purchase the product and once they have made the purchase, the consumption of this product by others is prevented. Examples of such goods include food and clothing.

Club goods are those which are excludable, i.e. they can be charged for but are non-rivalrous ([McNutt, 1999](#)). For example a cinema or a service such as television; one persons consumption of the product or service does not affect another consumers con-

sumption of the product. Finally we have common goods which are rivalrous in that once it has been consumed it can no longer be consumed by another and is also non-excludable; that the use of it cannot be restricted. We give the example of fish in international waters; there is no way of excluding people from fishing but those who do fish there affect the stocks of fish for fishermen later who wish to fish there.

The OXERA (1999) study was commissioned by OS to estimate the contribution which OS makes to the Great Britain Economy. GI information as outlined in the OXERA (1999) study states that it cannot be either a pure private good or a pure public good. It has characteristics of both public and private goods. GI is non rival in consumption but charges can be used to limit access. Love (1995) suggests that GI is in fact a quasi-public good. The OXERA (1999) study also states that OS itself cannot be seen as purely a public or private good provider. This is due to the two sided nature of its services. Where it is a Trading Fund it adheres to that of private-goods where the National Interest Mapping Services Agreement (NIMSA) is official recognition that the goods OS provides are public.

The types of goods outlined above are summarised in Table 5.1 below.

	rivalrous	non rivalrous
excludable	Private	Club
non excludable	Common	Public

TABLE 5.1: Summary of Types of Goods

5.3 Information Goods

Varian (2000) details the three main properties of information goods. Firstly, information is an *experience good*; the user must experience the goods before they know what it is and decide to buy it (Clay et al., 2003). Secondly, information goods have high production costs and low reproduction costs. This means that the initial cost of collecting and producing the information is costly, this may be from data entry or from the methods which are used which are costly. For example GI is costly to produce due to the scale and methods used to collect it. Once collected however, it is cheap to reproduce. Thirdly, information goods are typically non-rival and non-excludable, making them *public goods*.

As we detailed in the previous section *public goods* are *non-rival* and *non-excludable* (McNutt, 1999). This means that ‘one persons’ consumption does not diminish the amount available to other people, while non-excludable means that one person cannot exclude another person from consuming the good in question.’ (Varian, 2000). It is noted that the non-rival aspect of the good is a property of the good itself whereas the excludable part of the good is a ‘social choice’. That is, it is up to the holding

organisation how the good is available and to whom. Therefore, issues such as how to limit access to LD are key areas for investigation.

Information is an *experience good*, that is, in order for a consumer to know the benefit of the product they need to experience it, they cannot make the decision before they experience it. Consumption of such goods in a LD community highlight some research challenges. These challenges include: the evaluation of information quality with regards to a certain task, selection of a suitable dataset given a number of options and the integration of information from different sources. We aim to address some of these challenges through this research.

In order to sell information goods and maintain revenue, organisations could consider price discrimination of their goods. Price discrimination is where different prices are charged to different users. It is considered that this could enable organisations to recover revenue from the low demand sector (in the case of OS its leisure users) without destroying the revenue for the high demand sector (business users) (Linde, 2009).

Price discrimination can be in the form of varying prices of goods or by varying various aspects of the goods such as quality and timeliness. In terms of information or digital goods this is possible due to the nature of the goods. Time delays in loading of data and quality of data can be managed in order to differentiate between products.

How does this work in a world where we are trying to make data ‘open’ and more readily available? If links are consistently created to other datasets then the decision lies with the user whether or not to purchase the data. Web 3.0 makes data or information readily available, whereas previously in Web 2.0 the data was not linked.

In the next sections we investigate the characteristics of the business of information goods and specifically LD. The first issue we consider is information value.

5.3.1 Information Value and the Value added by Linked Data

We consider the factors which affect a consumers choice of products or information. The key issues highlighted by Zeithaml (1988) are price, quality and value.

We investigate value as the first of these three factors. The value of data on the web is a critical issue, especially for organisations such as OS which have a large amount of GI which they exploit to make profit. In order to maximise profit, it is essential that all digital GI products are easily accessible, readily available, and distributed to their utmost potential. We consider value important from both the consumer and the holding organisation. As we stated earlier there is the possibility of free and open data being available together. In the determination of prices for products, Zeithaml (1988) suggests that the value of data to the customer is a key determinant in the pricing level. We

also consider that in order to determine the qualities of the products available we must understand the value of this information to consumers.

Value theory is associated with decision theory that tries to explain why people place a positive or negative view on products. An investigation into value will be beneficial to this study to determine reasons for the parameters that affect a person's decision based on value.

Value added by the introduction of new technology can impact the new actual value of data (Longhorn and Blakemore, 2007). Ways in which this data can be distributed then become a foremost subject for exploration.

The *network effect* is discussed by Shuen (2008) and a positive network effect is suggested to increase the value of a good or service. Therefore the more people who adopt the use of a good or service increase its value. We use the example of Facebook to illustrate the network effect. If only two people use the Facebook networking site, it holds little value, as only two people are able to communicate using it. However, if more people adopt the use of the Facebook, the more valuable it becomes as more people are able to communicate with a wider audience. We foresee that this will be a positive concept for LD. As more people create links to data sources and the data becomes more used, it will increase the value of this data source. Therefore we must consider that encouraging people to link to data will in turn create more value as not only will it be easier to find, it will contain other links to more sources related to it.

Chesbrough and Rosenbloom (2002) indicate that value can be observed by a customer according to the ability to reduce the cost of a solution to an existing problem, or its ability to create new possibilities and solutions. In terms of OS and SWT it can be noticed that through the use of SWT, value will be added to the customer through the ability to purchase and link data sets exactly to the needs required, thus enabling the information to be made accessible to a much wider market than previously as a result of advances in technology.

We give the example of, where two datasets have been combined to give more value. A 'mash-up' of data about deprivation in certain geographical locations combined with a dataset regarding crime figures would be a valuable asset to organisations such as insurance companies as they would be able to use this information to determine the highest crime locations. This would enable them to make decisions for their business regarding how much they charge for their insurance premiums depending on high crime areas. Another example may include health figures with deprivation which may help organisations such as the National Health Service make decisions regarding provision of more healthcare resources in certain geographical locations. These companies already have access to this knowledge but with LD such efforts may become faster and therefore cheaper to produce useful models to make decisions as less effort needs to be put in to make datasets which quickly provide answers to questions.

According to [Chesbrough and Rosenbloom \(2002\)](#), on identification of the market, it is possible to determine the price and how a customer will pay for a product. As a Trading Fund, OS is required to return a reasonable profit and build funds for investment. In order for organisations to understand where the majority of their business comes from the identification of the market is required which will help them to distribute its efforts into areas which require more profit generation. We suggest that for the purpose of this research it would be beneficial for us to understand where the majority of its consumer base lies and we carry out a preliminary investigation into the type of consumers of OS data in detail in chapter 5.

The literature surrounding value of data suggests that the greater part of research into the value of GI is in the *content* value of the information ([Meeks and Dasgupta, 2004](#)), and suggests that a future suitable approach to research into this area would be the estimation of the value of the GI *relative* to the needs of new geospatial users.

5.3.2 Affordances

SWT will not only allow computers and people to work together but will also allow an organisation to supply its customers more easily with the level of information which they require ([Heath and Bizer, 2011](#)). This information will be obtainable via links authorised by access mechanisms which will enable differing levels of access to datasets, thus reducing information overload for users. Currently, users who have been given access to datasets may have to download the whole dataset, or parts of it, which have been returned by a general search engine and may or may not contain the required data. The power of SWT will enable organisations to grant access to its data to users at various levels. In the instance of OS a SW search may enable a user to search for Southampton and return only information regarding this specific area. It will however enable users to follow links to other information regarding Southampton if they so wish but does not overload them with information which is not relevant to their initial search term.

SWT will allow users of the data to search more accurately and precisely and have the ability to extract data ([Bizer et al., 2009](#)). SWT enables data stored in ‘the deep web’ to be structured in a way that enables reasoning on the data. The implementation of SWT will enable this information to be more widely distributed and linked to further resources across the web and in turn be more useful to consumers ([Latif et al., 2009](#)).

With the linking of datasets, there will be a suitable space for the creation of applications built around data which will quickly solve answers to queries in the form of single applications rather than browsing through multiple web pages.

By storing data in RDF on web pages we are making the link between other pages easier by having standards to work to. If for example we have a webpage created in

[RDF](#) containing product information, a second page contains reviews stored in [RDF](#) about these products and third page showing where we can purchase this item and if it is in stock we have created a much better environment for users to make purchases by reducing time spent and make assertions which can easily be inferred by machines if they are stored in a standard format.

The key benefit for consumers using [LD](#) is that whilst consuming data consumers are able to discover more related data than was previously possible. Data which contains links to other data is more useful to a consumer as it eliminates the need to actively search for related resources. Links exist on the web now, but with the power of machines inferring links to other datasets we are able to reduce time spent searching. Benefits to consumers of data in this format include the ability to create links to data and to reuse this data by linking to other sources.

For the data publisher there are a number of costs involved which include the time spent in publishing the data into machine readable non-proprietary formats. Time will be spent organising how the data will be represented and assigning [URIs](#) to the data.

The benefits to the data publisher include making the data more discoverable and thus will increase the value of the data due to its usability [Kobilarov et al. \(2009\)](#). It will also enable the holding organisation to maintain control over the data and with this high level of control there is the opportunity to restrict or allow access to the data.

If data sources are stored and published in a format to which it is readily able to be linked it makes the process of creating links easier for other users. Once in a machine readable format there is the opportunity for machines to infer relations to data which makes the process of linking data much easier. It also means that for consumers if they have readily available datasets in the same format it is easier to create links to the data as they can create links without having to parse the data into another format.

We investigate the economics of [LD](#) in more detail in the next chapter but some of the economic benefits of [LD](#) include the ability not only to sell the data itself but to build services on the data which is available ([Auer et al., 2007](#)). [LD](#) applications are much easier to build and can be seen to contain more data than previously possible with mashups built with Web 2.0 technologies. The power of [LD](#) means that where previously fixed data sets were used, applications can be built on a wide range of datasets and more questions can be asked from data sources ([Bizer et al., 2009](#)).

This can be done not only by the holding organisation but by other users if the data holds a license which enables them to do this. There is also opportunity for revenue to be made from the data used in these apps which may be recurring, as sales may be made by high volumes of consumers.

We believe that there will be another divide in the type of consumer and the data. There will be consumers who will want to download the data onto their devices and reuse this

data again for other purposes. There will be types of users who will want to use the data for a one off reference. There will be data which is useful on repeat usage. For example, **GI** can be reused to show how a place has changed over time, or to refer back to as a direction. Other data (for example news) soon becomes outdated and therefore may not have the same value as data which is not time limited.

In this section we have investigated the potential of the technology of **LD** and the advantages it has for organisations. We investigate the possibilities to generate value from **LD**. There are two primary areas we address. The first is in organisations building applications using their own **LD** the second in generating value from providing data which enables individuals to create applications. This in turn can help to draw people to an organisation's data and thus may generate value. We have also described potential of **SW** and clarify the separate roles which **LD** and **SWT** have and that **LD** does not need a **SW** as such in order for it to exist (Hausenblas, 2009).

Given that **LD** is a new concept and the technologies for the **SW** are in development, the commercial options for data available in this format are at an early stage. Seen as the data can be made available as a commodity itself, organisations could sell the data as 'raw data', which can be used by consumers to create their own applications. Alternatively organisations could create applications using their own data and sell these applications. There is also the option to carry out both and supply the data alongside the applications. We would like to find out if there is more appetite for the data or for the applications. We investigate the appetite for data further in the empirical research detailed in the remaining chapters of this research.

5.3.3 Information Quality

As the web of **LD** grows and more data is published from many different sources, an understanding of the quality of this data is important both for data providers themselves as their data may not be trusted or recognised if others exist and for the consumer, as they are not able to determine which datasets to use or trust.

One of the problems of data quality is overlapping (Mendes and Mühleisen, 2012). This is where two datasets may have two separate identifiers (**URIs**) for the same object. For example, a pub in Southampton may be identified by one dataset where someone else may identify Southampton the place in their dataset but not make the distinction that the Southampton that they are referring to are in fact the same place. This illustrates where we will have issues with quality and trust. Users may navigate from one dataset to another but not necessarily be aware that the link between data is the same. This leads to them not being able to trust the dataset they use, unless there is a measure in place to enable them to make a decision about the quality.

Another issue of quality with **LD** is the open nature of the web. Anyone is able to

Accuracy of Data	Believability Accuracy Completeness Reputation
Relevancy of Data	Value Added Councils Relevancy Timeliness Flexibility
Representation of data	Interpretability Consistency
Accessibility of Data	Security Cost

TABLE 5.2: Categories and Dimensions of Data Quality - Adapted from [Strong \(1996\)](#)

publish data on the web without any checks or policing to ensure the accuracy of the data ([Bizer et al., 2012](#)). Consumers need to be sure they are accessing data which is correct and from a source which they are able to trust.

[Sansone et al. \(2012\)](#) suggests that rather than assessing whether a dataset is of good quality, it is easier to identify the areas in which it is bad. They also suggest ways to conform to the LD ranking system as outlined earlier. Although it is an alternative way of investigating how one dataset differs from another, by looking at what is missing from one dataset, it will not give the consumer all of the data they may require. This may lengthen the process as users may be aware of the features they want from data rather than what they do not want. We recognise the importance of understanding the need for information quality standards and how this is going to be achieved.

[Strong \(1996\)](#) outlines some of the the key criteria they have found to be most important to data consumers. They summarise four categories in data quality - accuracy, relevancy, representation and accessibility. Within these categories lie the criteria we can identify our data with. Table 5.2 illustrates the categories of the criteria and examples of each. From this we have been able to identify key points for further empirical investigation later in this research.

5.3.4 The Economics of Linked Data

As we established above, digital products available on the web are experience goods with low reproduction costs and are excludable in nature ([Shapiro and Varian, 1999](#)). In order to understand the economics of LD explicitly, we need to investigate the economics of

information products to understand how they may be applicable to [LD](#).

We anticipate that there may be more revenue to be generated from applications and services which are created to benefit potential consumers rather than directly from the data. We notice that there is great potential for applications to be developed using various different datasets such as crime figures, accident rates and school league tables. This provides us with more beneficial tools which have not previously been experienced. Revenue may still be created, though it may be for small portions of data.

Although we are looking for the potential economic benefits from LD we are also looking at the second order benefits - adding value in other ways which are harder to assess. There is a great quantity of literature in this area but currently there are more questions than there are answers. How do you account for external value? Externalities of value include saving human cycle time, reducing information friction (effort required to process data into useful formats and standards for all to understand), cost of republishing in multiple formats etc.

We also note that there are difficulties where a number of datasets have been linked together from various different sources. How do we attach advertising to multiple datasets, each holding a similar level of relevance to the data?

New ways of doing business on the web have emerged over time, and, although they may have been deemed infeasible at the start, have achieved significant success ([Lassila and Hendler, 2007](#)). For example, Amazon is a key example of the online marketplace, which began the trend of individuals trading online. In order to determine if these models are suitable for LD business on the web it is important to investigate possible replacements for unsuitable models.

5.4 Revenue Models

Through this research, we find that current revenue models, such as the advertising model, are no longer as profitable as has been experienced in the past ([Lopes and Galletta, 2006](#); [Picard, 2000](#)). These models are unsuitable where large amounts of data are linked together from a number of different sources: we aim to contribute to finding suitable ways to exploit both [LOD](#) and [LCD](#).

[Lassila and Hendler \(2007\)](#) discuss how revenue models on the web have evolved over time, but are models, such as the subscription model, suitable for [LD](#) on the web? It is becoming increasingly apparent that these existing models will not allow organisations to exploit the full capabilities of the technology ([Alani et al., 2007](#)). Revenue models used in Web 2.0 can be implemented easily, but are they going to be particularly suitable for [LD](#)? Current [LD](#) practice is based around ‘open data’: however, some organisations

may not be in a position to give away all their data for free (for example, OS,) therefore, there is a need to investigate models which can allow access to LCD.

A revenue model is a component of the wider business model. The business model defines the way in which the company will do business, including the customer, the product and long term planning (Teece, 2010). The revenue model however determines how it will monetise this business.

Sliwinski (2004) states that the ideal for a profit maximising business is to charge the maximum a customer is willing to pay (i.e. the monetary equivalent of the perceived value) for what is offered. Current models such as advertising are not suitable for industries such as the newspaper industry and there needs to be investigation into more suitable ways of generating revenue. Beuscart and Mellet (2008) suggest that the current models on the web are weak and suggest that the advertising model is no longer as profitable. We notice this in the newspaper industry (which we describe in more detail later). Newspapers such as the Sunday Times have recently introduced a pay-wall to the online version of the paper (Thurman and Herbert, 2007a). This is due to the decreasing revenues received from advertising. This may be due to a number of reasons such as people are intolerant to advertising on the web where it can be deemed as a nuisance whilst browsing. It is now easier to remove advertising or block pop-up windows whilst browsing the web and therefore companies could be more unwilling to pay high costs for advertising. Sliwinski (2004) also suggests that traditional pricing models (quantity, area, feature and zone based) fail to mirror the value of the product to the individual user. This is due to the fact that users value products differently and what is valuable to one user may not be to another, and therefore we note that consideration into the factors which affect users value of products is important to ensure that revenue can be generated from potential LD products.

Longhorn and Blakemore (2007) suggest that current-pricing strategies can be improved if customer value of information is taken into account and suggests that there is no need for cost based value; the value should reflect the customer value instead. This research aims to investigate the reasons why people chose whether to make a digital product purchase, and, in order to do this, a thorough examination of the data required and used by OS consumers and their reactions to new products should be gauged, in order to determine the pricing model required for such a service (Lowe, 2005).

There are several different classifications of revenue models. Organisations need to ensure they generate the maximum profit and therefore need to ensure they have selected the most suitable model for their business. Picard (2000) suggests that a number of current models for online services have become outdated and with developments in technology, audiences for online content have changed and therefore the demand for online services has also changed, requiring models to suit both the providers of online content and its consumers.

There are three types of models found on the web, which include; paid, free and advertising supported models (Picard, 2000; Novak and Hoffman, 2001; Chehade, 2011). Further to this the models are broken down more specifically, we outline the six key models for business on the web below which have been defined by Shuen (2008)

- Subscription
- Advertising
- Transaction fee
- Volume (unit-based)
- Sponsorship and co-marketing

5.4.1 Revenue Models for Digital Goods

This next section outlines the potential revenue models specifically for digital goods. There are a number of different possibilities which we outline but then we narrow them down to the most suitable ones for a LD environment.

Organisations may not use just one single revenue model and may use multiple revenue models in order to generate the required level of profit. Flickr for example uses the subscription feed, sponsorship and advertising models to ensure that each level of consumer is supported. We may consider that a combination of models is suitable for LD to ensure each aspect of LD is covered from free, to premium data.

5.4.1.1 The Advertising Model

The advertising model is an extension of traditional media modelling as seen on television and in print newspapers (Rappa, 2004). The advertising model is most suited to websites or media which attract high volumes of traffic so the advertising is seen by large numbers of users.

Web or online advertising is often displayed as a banner across a page or spread across one side of a page which can contain flash and images to attract consumer attention. There is also advertising which can appear as a pop up which consumers must actively close or minimise in order to continue with their task. Revenue is generated from advertising in a number of ways. The most common ways are explained in more detail. Revenue can be made per click, so each time a consumer clicks on a link in the advert the company is charged a small transaction fee for referral to their site. Alternatively it can just be from individual views, so each time an advert is displayed, it is charged. It can also be

charged in a fixed fee which can be time limited; so an advert is placed on a site for set period of time and no extra royalties are charged following the initial fee.

The advantage of advertising online is that it can be directed to its target audience. For example the Spotify application targets advertising depending upon the type of music which the user listens to. This means that the consumer tends not to hear advertising for items which are demographically unlikely to be of interest to them. Facebook also uses the details of a users profile to ensure the adverts seen on their pages are relevant to their interests. We see the potential for advertising to be incorporated into LD by creating advertising content modelled as LD and links created to and from ads from various items. For example a musician may have a page of LD created about them and an advert for a concert they are performing could also be created as LD and linked to the page about the artist.

5.4.1.2 Sponsorship

The sponsorship model works as follows; for example the money section of an online newspaper may be sponsored by a bank such as Lloyds TSB. The bank will provide content which will be displayed on that site and will help the newspaper cover the costs therefore enabling users to view the site for free. We notice that this model does however come under the advertising model but differs in that the advert will only come from one organisation rather than many and may not be well targeted to the entire audience of the newspaper as not every reader will use the same bank.

The disadvantage of this model for LD will be finding and maintaining sponsors for data. There will be large sets of data available which may benefit from such sponsorship but it is not a model which would be sustainable for all datasets as some may be created by individuals which may not be able to find sponsorship or require long term sponsorship deals to maintain revenue. Therefore for the purpose of this research we will disregard this model as a potential revenue model for LD

5.4.1.3 The Transaction Fee Model - Micro-payments

Where a large number of transactions are made on a daily basis on a website, the most commonly found revenue model is the transaction fee model, where sellers of products using sites such as eBay or Amazon are charged a small percentage of the final selling value of the item. Therefore the seller is charged a percentage of the cost of the lists. Micro-payments as outlined by Chi (1996); Dai et al. (2001) are purchases which can be made without making a new account for each seller and are used commonly on sites such as Amazon and eBay. For content on the web which has a date limited profile, such as newspapers which are outdated overnight, the faster they are obtainable the

more valuable they are to a consumer. We consider sites such as Amazon and iTunes for illustration of micro payments. Amazon, does not just concentrate on small, individual payments for single items which we may consider more suited to the term micropayment. iTunes, however supports payments for many small purchases of between 69 pence to 99 pence for individual tracks. It also supports purchases of whole albums and the consumer is charged immediately having already pre-registered their card details for payment. Amazon also requires a user to register an account and provide payment details making the shopping process simpler but their payments are not necessarily under £10. The purchases however are allowed from many different sources under the Amazon umbrella but are managed by Amazon allowing the consumer to just make one payment. We consider micro payments to be a suitable method for LD due to the nature of individual items of data holding a minimal cost and a simple and fast way of managing payments of such data may ensure a smooth consumer purchasing experience.

5.4.1.4 Volume

This model is used mainly for offline products and services and the revenue is generated from charging per unit of item sold. Therefore the consumers pay the same per unit. This model may work for LD but would require investigation into other models which could be applied with this model. We also note that this model tends to be used in offline sales and is not commonly used online. Therefore we will disregard this model from further investigation as there are more suitable models which can be easily applied to LD.

5.4.1.5 The Subscription Model

The subscription model as outlined by Rappa (2004), is a model where a user pays a daily, weekly or money fee to a product or service. The subscription model is often used with a free option or alongside advertising. In a LD scenario we suggest that subscribers to a LD service could receive more resolvable URIs as a benefit to their subscription to a service. We outline how this may be possible in our technical framework in chapter 8. Alongside the subscription model we also note the free element which can be included in this to create a freemium model. We outline free in more detail in the next section.

5.4.1.6 Free

Whilst there are models which require payment or revenue, we also consider the free models. Anderson (2009) illustrates a number of 'free' models which give away some, or all, of the product in order to generate income. The benefit of using such models to

generate custom means that the amount of information which is being re-used is high, and so demand is stimulated through loss of re-users engaging in the market.

Although there are many organisations offering their services and products for free, Google has introduced a cap to its usage which restricts the usage for commercial consumers. Table 5.3 shows a comparison of the numbers of requests available for free users and for business users of the API.

Features	Maps API	Maps API for Business Street View
Geocoding Web Service	2500 requests per day	100000 requests per day
Directions Web Service	2500 requests per day with 10 waypoints per request	100000 requests per day with 23 waypoints per request
Distance Matrix Web Service	100 elements per query 100 elements per 10 seconds 2500 elements per day	625 elements per query 1000 elements per 10 seconds 100000 elements per day
Elevation Web Service	2500 requests per day with 25000 samples per day	100000 requests per day with 1000000 samples per day
Static Maps API maximum resolution	640 x 640	2048 x 2048
Static Maps API maximum scale	2X	4X
Street View Image API maximum resolution	640 x 640	2048 x 2048

TABLE 5.3: Comparison of Google Maps API vs Google Maps API for Business taken from <https://developers.google.com/maps/licensing>

Table 5.4 illustrates the four types of free models as outlined by Anderson (2009). The table shows the four free models and explains what is given away for free and who can receive the ‘free’ version.

Free Model	What is Free?	To Whom?
Direct Cross Subsidies	Any product that entices users to pay for something else	Everyone who is willing to pay eventually, one way or another
The Three Party Market	Content, Services, Software	Everyone
Freemium	Anything matched with a premium paid version	Basic users
Non Monetary Markets	Anything people choose to give away with no expectation of payment	Everyone

TABLE 5.4: Comparison of The Different ‘Free’ Models

The problem with the direct cross subsidies model is that some people will never be prepared to pay for the data, and therefore will continue to seek free data or information

for as long as they need to use it. This model is therefore not an appropriate model on which to base a new pricing regime as it may not make any profit.

The three-party market model is referred to as a two sided market where two user groups are supporting each other, i.e. the advertisers and the consumers. The products are given away for what is seen as free to the consumers, whereas it is the advertisers who are paying the price for the product, in order to reach its targeted advertising groups. The advantage with this model is the win-win situation noticed within the market. Consumers are receiving a product which they see as ‘free’, whilst the advertisers are achieving publicity from the exposure of their campaigns. We have outlined the advertising model earlier in models for revenue generation, but we also consider it here in the free models as the advertising here supports free use for consumers. The advertising model for revenue enables content providers to supply information without requiring a payment from the consumer. This does however suggest that without the revenue from advertising the provider may not be able to continue its operation and may need to seek further assistance from other means.

The last model outlined in the table illustrates how some people are willing to give away things for free because of other gains. Some musicians have realised that they cannot overcome piracy and thus embrace releasing some music for free to stimulate interest in concerts and themselves as artists. Google receives information free from anyone who creates a website, and inadvertently people who search are helping to improve ad-targeting algorithms. All of these examples demonstrate users contributing a small amount of labour, and in return getting back something which is useful to others, which in turn creates a non monetary market.

We have established that current models for content online are somewhat unsuitable due to changes in the economy and consumer spending behaviour (Donker, 2009). Therefore, to continue producing online content we must suggest a suitable alternative to current models. We suggest that a suitable model for further investigation would be the freemium model. This is because we notice the potential for free data to be used to point to paid data. Good Relations³ has enabled some businesses to start providing metadata about products, prices and specifications which google searches and points users to. This Metadata can be LD which is free but which may lead to revenue in an indirect way. We also note that by providing a free version of a product or in this instance data, may bring more traffic to a website and therefore encourage a further purchase of a paid product.

The freemium model aims for a percentage of users to support the rest and does not solely rely on revenue from one stream. For every user who pays the full, premium version of the product, the other users get the basic version for free. This model works because digital products have very low or zero costs for reproduction once the product

³<http://www.heppnetz.de/projects/goodrelations/>

has been collected. Therefore the cost of actually providing this product to the remaining free users is almost zero.

5.4.1.7 The Freemium Model

As we outlined in the previous section, there are 4 different types of ‘free’ models. For the purpose of LD where we notice there is data which cannot be given away for free (LCD) we realise that we need to facilitate a situation where there is a free version matched with a premium version. Therefore we investigate the ‘Freemium’ model as a potential solution to this.

Anderson (2009) describes the freemium model as the opposite of the ‘traditional free sample’. The traditional model gives away 5% to generate income from the other 95%. The freemium model gives away 95% of the product and sells 5%. This model is an interesting aspect of pricing to explore as it holds different opportunities for the holding organisation.

There are four different models within the freemium model outlined by Anderson (2009); Time Limited, Seat Limited, Customer Type Limited and Feature Limited.

The time limited model gives away, for example, thirty days of use for free and then after this time a subscription must be paid. This model gives the customers a real opportunity to test the product and once used the free version for the trial may be more inclined to pay for a full version if it is suitable for their requirements. The OS uses this method for OpenSpace Pro. It can be implemented using an API enabling users access for a restricted time per day or month. Companies such as Spotify also use this method to enable users to listen for a restricted number of hours per month and then if they wish to pay for more are asked to subscribe or wait until the next month.

The seat limited model allows use by a number of users for free and then after that it must be paid for. It is easy to implement but often can take up the low end of the market, whereas the customer type limited model gives away its product to smaller companies and makes charges to larger companies.

The feature limited model is where two versions of a product are supplied, one where the basic version is given away for free and the second version which has more features and must be paid for. This model allows a wide range of customers to be reached and will generate a loyal customer base who are unlikely to be phased by the price. The problem with this type of model is that two versions of a product have to be produced which need to be carefully planned. If too many features are given away for free, then custom may be lost from users who do not need to pay for the full version. However, by giving away for free the organisation is enabling users to try their product or service

without any pressure to subscribe, and, over time, users may find that they naturally progress into the premium version as they require extra space or features.⁴

Kanliang (2004) states that the largest part of added value is through information associated with other products. As the number of people in the network grows, the connectivity increases, and, if users can link to each others content, the value grows at an enormous rate. This is known as the network effect (Hendler and Golbeck, 2008; Shapiro and Varian, 1999). One person may use a service and although they may not subscribe to the premium service, they may refer someone who will use the premium service, which is especially relevant to LD. Someone may create some RDF about a certain topic and publish it on the web, but not necessarily make any specific links. Others who have a dataset may publish more data and link to it and so forth. By encouraging the use of LD, datasets web wide will become more valuable due to the richness experienced through connections between data.

Some examples of feature limited freemium model include Flickr®⁵ where users can sign up for a free account with 300MB of uploads and 2 videos a month or subscribe to the premium service which offers unlimited images and videos per month. Spotify⁶ gives the option of a free account which is advertisement supported or paid subscriptions which remove the adverts, give a higher bit-rate stream and offline access. These models have proven successful to both organisations and others which include Skype⁷ and LinkedIn.⁸

When considering LD we ask the question of ‘How do we account for external value?’ These are externalities of value, such as saving human cycle time, reducing information friction, cost of republishing in multiple formats etc.

Some datasets such as DBPedia (which is crowd-sourced resource where structured information is extracted from Wikipedia) are valuable on their own as the information they provide can be considered useful without being linked to anything else but we note that the value of some LD will be the extent to which it is being linked to by other datasets. i.e. the more heavily-linked it is, the more valuable is the dataset as it becomes more accessible due to more links being made to it from other datasets therefore there is more chance it will be found. We mentioned in chapter 2, the five star ranking system suggested by Tim Berners-Lee. To achieve five stars in this ranking the data must be linked to other data to provide context. Some data published on the web may have no context as it may just be raw figures, but when linked to a geographically location could become more valuable. For example, the bathing water example with data provided by the Environment Agency.⁹ This data demonstrates the areas where clean bathing water is available on the coastline. If the data is provided with the name of the location it

⁴<http://spencerfry.com/freemium-model>

⁵<http://www.flickr.com/>

⁶<http://www.spotify.com>

⁷<http://www.skype.com/>

⁸<http://www.linkedin.com/>

⁹<http://www.epimorphics.com/web/projects/bathing-water-quality>

is useful, but, if it is linked to the coordinates given from OS and also linked to data about the local amenities to that location it becomes more valuable as a resource for consumers to use.

In order to realise this value it is important that all data is made linkable. It must be in a format from which links can be made. If a dataset is not available in a format which is easily linkable then its value is relatively limited as it cannot be linked any further.

We have seen a decrease in the amount of revenue being generated from advertising campaigns alone, and therefore to solve this, we also investigate free models, in particular the ‘freemium’ model as this presents the opportunity for organisations to continue to offer a premium version of their data products whilst enabling users to access a free version as well, which can be supported via advertising (Chehade, 2011).

5.4.2 Willingness to Pay for Information Goods (Online)

We have established in the previous sections the potential models for the consumption of LD online. We now look at the issues which may affect a consumers willingness to pay for goods online.

Research carried out by Ye et al. (2004) found that the willingness to pay for online content was influenced by their perceived value of convenience that the services provide. The study focuses on charging for services which were previously free and outlines the reasons that help to explain why consumers are willing to pay for online services. Investigation into LCD should be considered to enable organisations to ascertain the benefits of a linked web of data and consideration into the factors which will affect a users decision whether to pay for LD.

The cost or price of content online is a key issue which we raise in regards to the willingness to pay online. We consider that this could extend our investigation further, following the clarification of quality features.

When there are free options available, the pricing of content online is even more imperative to ensure it meets the needs of both the supplier and the consumer. It must generate enough revenue for the supplier, and if not then it is subsidised in other ways and that it is at price which the consumer finds reasonable to pay.

There has been a sharp decrease in the profitability of advertising on the web, and, as a result, the freemium model has become increasingly popular Lopes and Galletta (2006). However, it appears that users are unwilling to change their spending habits(Dou, 2004). This is due to users of the internet becoming increasingly aware that free alternatives are available (Dou, 2004). How do organisations gauge the content and value of the premium version of their products? In order to understand this, we feel it is important to investigate the factors which affect willingness to pay for online content, especially

linked online content, as old revenue models become obsolete. Firstly we notice that users' online purchasing behaviour could be shaped by a number of criteria: specifically, demographics, net value and cost benefit and by past online habits [Ye et al. \(2004\)](#).

Consumers' willingness to pay is related to their perception of convenience, added-value and service quality ([Wang et al., 2005](#)). Therefore, when considering LD, we should examine the types of applications for development, the value they give as an application as opposed to a single dataset, and finally, the quality of an up-to date and accurate service. LD will make access to data faster and decrease time spent searching. Therefore, if the audience is targeted accurately the LD movement will transform the way users interact with data on the web, thereby making the data more useful.

We also look at consumers' willingness to pay when there are free options available. We notice that some consumers are willing to pay for content online and some are not ([Guel and Rochelandet, 2006](#)). The ones who are prepared to pay, do so for a number of reasons, such as improved quality, income, and usability. A proportion of users state 'friends' as a reason for paying and this suggests again the power the network effect has on willingness to pay.

In a situation where one user in a social group is a paid user, this can influence the other free users in the group to also become paying users ([Wang and Chin, 2011](#)). This suggests that again reputation can influence a brands name or consumption of data.

Willingness to pay for premium services is strongly associated with the level of social activity of the user ([Oestreicher-Singer and Zalmanson, 2009](#)). This suggests that the LD movement could enable people to participate in generating and sharing content online and that there is potential for applications to be built with both a free and a premium option for both types of users.

It has been suggested that people use free review websites for products more often than those which require payment [Kowatsch and Maass \(2009\)](#). We would like to understand if this is affected by the product, or the payment type or trustworthiness of the resource.

From the research described above, we can see that there are a number of factors which affect consumers willingness to pay. We note that many of the concerns are regarding attitudes towards paying for data. From this we explore the willingness users may show to pay for potential LD in the following later chapters.

5.4.3 Trust

We have outlined the concept of willingness to pay for data and one of the concerns we recognise is trust in the source of a product. The mechanisms used to verify that a source is who they claim to be, contributes to trust ([Artz and Gil, 2007](#)). That is, when

looking at a link to data or the data itself, a user will want to know whether the link is correct and that the data it directs them to is itself correct.

McCole et al. (2010) outlines three further factors for the study of trust: vendor (the person or company who is selling the product), internet (in this case the medium by which a purchase is made, that is, online and not ‘in store’) and third parties. When contemplating a transaction, a consumer will take these three factors into consideration. They go on to explain the two types of uncertainty; firstly uncertainty with the technology and secondly uncertainty with the product.

When we consider trust on the internet, we consider that there are concerns with the ‘trust’, more specifically in the technologies surrounding it. Specifically the authentication (how a transaction is carried out), confidentiality (what happens to personal information about a transaction) and the transaction itself (will it be carried out smoothly, will the correct product be ordered etc). We note that one of the key factors of this trust is understanding of the technology. This is of particular concern with the introduction of a new technology, as new users may not fully understand the way in which the technology works and as a result do not trust it. Therefore education about the technology should be available.

Trust in the vendor in this instance is with the data provider or the product. It is well researched that brand awareness and recognition is a big factor in a consumer’s trust and willingness to pay for a product (Oh, 2000; Macdonald and Sharp, 2000). In this instance we are looking at trust of LD. We have two issues here, not just trust of the source of data, but also trust of the link which is created towards this data. A mechanism to enable consumers to establish grounds for trust is important in a LD situation, as we want the consumers to know that the data they are using or purchasing is recognised and trusted.

The second issue is the problem with the accuracy of the links. The links created between data may predominantly be created by users across many different sectors. We consider here that a way of checking or validating these checks may be beneficial to gain consumers trust in user generated content.

The voiD vocabulary is an RDFS which describes linked datasets.¹⁰ Each dataset is created and maintained by a single provider. The voiD vocabulary aims to create a bridge between the publisher and the user. Omitola et al. (2011) have developed an extension of void to voiDp, which provides classes and properties that publishers can use to describe the provenance information of the data. They describe provenance for the data as when and how it was derived, what data had been used to derive it and who carried out the transformations which achieved the data. This is a mechanism which can be used to help users deduce where the data they are using or linking came from,

¹⁰<http://semanticweb.org/wiki/VoID>

which will help with the issue of trust. We outline in our technical framework in chapter 8 how this will be utilised.

We extend our exploration into the trust of products and information online, in chapters 5, 6 and 7 which help us to understand to what extent trust affects a consumers willingness to pay for content online.

5.5 Digital Content Industries

Having investigated [LD](#), its technology and capabilities we have been able to draw a number of similarities between [GI](#) and other industries such as the news industry, the music industry and the software industry. We explore these similarities in the different industries in further detail below. However, we note that although there may be similarities in the industries, they are still different industries with different business models, but the revenue model, which is what we are discussing at this stage, is similar in that it requires attention due to changes in the market and behaviours of consumers and also changes in technology. Handheld devices have meant that demand for traditional media has changed. We appreciate that the content in these different industry hold no similarities but we do appreciate that changes in the business environment affect the ways in which business is carried out and different revenue models can be applied across many industries. We also acknowledge that there are two types of [LD](#). [LD](#) as a commodity in itself, that is as a product which can be sold and services which have been built around [LD](#).

5.5.1 Newspaper Industry

We consider two factors in this section, firstly the profitability of the news industry and then following this we talk about the use of [LD](#) in the news. Over the years, the internet has become a popular medium for reading news ([Turnor, 2007](#); [Gunaratne, 2010](#); [Zwemer et al., 2010](#)). A study carried out by [Li \(2006\)](#) shows the advantages that online news has over paper print. This includes the ability to be updated more frequently, the addition of audio and video content and the ability to have more interactivity i.e. through comment pages and via blogs. However, despite this increase in online activity this content has widely been available for free and it is clear that the advertising model is not generating the required revenue to sustain the news industry on the web ([Chyi, 2005](#)).

In the past, news organisations have offered news for free, in the hope that it will increase audiences ([Ihlström and Palmer, 2001](#)). With high audience figures, it is hoped that it will in turn attract advertisers in order to generate revenue. Despite this, there has been a decline in revenue from advertising in online newspapers which is in part, due

to increased costs of online advertising a result of this, the companies are reducing the amount spent on advertising all together (Kirchhoff, 2009; Kind and Sorgard, 2009).

We also note that new devices such as wireless mobile phones and devices such as the Apple I-Pad and Amazon Kindle have contributed to the increase in online viewing of news and decrease in paper print news. All the time that people are viewing the news online for free, the sales of the paper based news is dwindling, which is causing news organisations to struggle.

News organisations have been working with different alternatives to the online advertising model such as subscription, micro payments and revenue sharing with search engines like Google.

We note the Guardian as an example below:

- EVERYDAY+ save 41% – 7 day Guardian and Observer papers, plus full iPad and iPhone access - £8.00 per week
- SIXDAY+ save 36% – 6 day Guardian papers, plus full iPad and iPhone access - £7.00 per week
- WEEKEND+ save 29% – Saturday Guardian and Observer papers, plus full iPad and iPhone access - £5.00 per week
- SUNDAY+ save 26% – Observer paper, plus full iPad and iPhone access - £4.00 per week

The Guardian offers 4 different packages and produces vouchers which customers can take to the a physical shop and purchase a paper copy, alternatively they can purchase view the paper online or on hand held tablets devices and mobile phones. This prices shown display a reduction in the cost from a subscription to purchasing the paper over the counter at a shop without the subscription.

The Times offers a different subscription package with just 2 options:

The Classic Pack

- The Times Monday to Saturday
- The Sunday Times
- The Smartphone app

- Access to The Times and Sunday Times websites
- Times+ membership worth over £100 a month
- £6 a week - £26 a month

The Ultimate Pack

- The Times Monday to Saturday
- The Sunday Times
- The Smartphone app
- The Times tablet app, The Sunday Times tablet app
- Access to The Times and Sunday Times websites
- Times+ membership worth over £100 a month
- £8 a week - £34.66 a month

We note that when the pay-wall was introduced to the Times a sharp decline in online viewing was noticed ([Thurman and Herbert, 2007b](#)). This may be analogous to introducing a pay wall to data on the web. Therefore, if this is so, will value be generated from other places, such as applications using the data and other services, rather than the data itself? ([Gallaughier et al., 2001](#); [Morales-Arroyo and Sharma, 2009](#)). We note here that news organisations publish information and not necessarily raw data which is where we consider that perhaps applications which have transformed raw data into useful information rather than the sale of just raw information may be more useful to consumers.

If news organisations need to start charging for their data and yet still remain competitive over competitors, they will need to keep enabling features which attract users ([Zwemer et al., 2010](#); [Sylvie, 2008](#)). LD is especially beneficial to the news industry. LD provides a wealth of relevant data concerning specific locations, therefore, when a news-worthy event takes place, reporters are able to uncover much more detail about a location than previously possible due to the links connecting different datasets together ([Belam, 2010](#); [Troncy, 2010](#)). For example if a serious crime is committed in a location, this crime can be pin pointed on a map and then the news coverage surrounding the event can be attached to this location. Further to this insurance companies can use this data for guidance on charges for insurance premiums, or potential home owners can look to see if the area they wish to move to is safe. This will enable users or reporters to find more specific detail relating to a place or event enabling them to generate more detailed discussion and knowledge regarding the news. LD also enables consumers of linked news to create reports and statistics regarding reoccurring news from specific

areas, more detailed crime reports, health problems and numerous other events which add to consumers' knowledge.

Revenue models for newspapers are in a state of instability (Sylvie, 2008). There is increasing evidence to suggest that people are unwilling to pay for news when there are such vast choices for free news online. It is also noticed that users are unwilling to pay for content online which has only a short term value (Thurman and Herbert, 2007b). Therefore it is important that these news industries offer a premium service alongside its free service. For example, better comment pages, easier and faster navigation and clearer links will give people better choices and use a better tailored service.

However, we notice that if content is valuable, for example music, and is not freely available elsewhere, users can be encouraged to spend money and we notice this in particular with iTunes (Thurman and Herbert, 2007b). People want to build a library full of music and the cost of replacing such items would be expensive. We go on to discuss the music industry in more detail next, but for now we ask, is news valuable enough to users to encourage them to pay?

News organisations offer bundled online articles, like that of a printed newspaper. Research by Stahl et al. (2004) suggests that revenues can be higher for bundled packages. For a set fee, a customer receives a paper full of articles, rather than a pay per article service. This bundling could enable packages to be tailored to a specific user's needs, such as articles by a certain author or subject, thus making them more valuable for reference at a later date (Veglis, 2004). This however, is only valuable to a small proportion of users and most readers of news may not want to refer back to an article again once it has been read. Unique content, such as reader comment, is possible to attract some users. Stahl et al. (2004) states that further investigation into this area needs to be carried out to clarify this, but we suggest that this is a possible model to investigate for the sale of not just news online but for the sale of other LD.

The research carried out by Thurman and Herbert (2007b) shows that most UK online newspapers are charging for something, be it mobile services, games or email alerts. However, which combination of charging is the most lucrative? The evidence above suggests that a freemium model as detailed earlier, is possibly the most suitable revenue model for news online and thus leads us to suggest that a freemium model for LD is most suitable, as pure subscription models do not appear to attract strong revenue streams when free alternatives are available.

5.5.2 Music Industry

Pre internet, the music industry concentrated on physical media, from vinyl records, to cassette tapes and compact discs. Now with mp3s, the internet era not only enables

the sharing and copying of music but has left behind the demand for the physical media we have previously used. Consumers are now able to purchase individual tracks whereas previously whole albums or cd's were purchased, with tracks which were often not required.

Before the rise in digital music, the music industry was partially protected by copyright laws. The copying and distribution of CDs remained illegal, and offenders could be prosecuted. However, the internet poses a real threat to the music industry, where copying and distribution of digital music files is fast, and prosecution is harder, as offenders are harder to locate.

The emergence of peer-to-peer file sharing websites such as Napster has forced companies in the music industry to re-think their business models, as these networks have made music a non-excludable good (Hougaard and Tvede, 2010; Gaustard, 2002; Dolata, 2011; Lin, 2005). This is partly due to the nature of the product: costly to produce and very easy to reproduce (Teece, 2010) and partly due to the changing habits of consumers and the availability of new media applications such as Spotify which allows users to 'rent' music where previously they may have had to purchase.

The concept of access versus ownership is predominant in the digital world where consumers are choosing to 'rent' items such as music, whereas previously they would have purchased a physical product (Heimer, 2011; Wiercinski and Mason, 2010). Spotify is a key example of music 'rental'. Spotify was established by Daniel Ek in 2006, and offers consumers a unique streaming music experience via a downloadable platform similar to iTunes (Kreitz and Niemela, 2010). Spotify uses peer-to-peer technology but adds DRM to the music which prevents it from being played in any other platform. Consumers are able to create and edit playlists of music they like. The free version of Spotify allows consumers 10 hours free listening per month. Following the time restricted version they can choose to pay for access for limited access £4.99 or unlimited access for £9.99 per month. There are links available within the platform for users to download specific tracks they like. Initially Spotify allowed unlimited access but found this was not viable and so introduced a 10 hour cap to the service. The problem Spotify has now is, the more people who pay to use the service to remove the ads, reduces the number of consumers hearing adverts, which in turn puts off advertisers as they are not receiving value for money.

Apple has a wide variety of products from mobile devices to televisions. By offering music for sale via their online music store iTunes and providing a platform for the downloading and listening of music from their store, they are creating a type of services which is totally dependant on their software. With the recent introduction of the ability to download movies or rent them they are also opening up the opportunity for users to develop a need to purchase the physical products they produce in order to create a seamless experience using their products. We see this following the trend of the Gillette

razor. Gillette charges a high price for its replacement blades but offers a whole new set of razor and 'free' blades.

We consider in this section the renting or downloading of films. Traditionally people went to the cinema to see films or waited until they were released on tape or dvd and either rented them from traditional film rental stores such as Blockbuster or more recently LoveFilm and Netflix. However Apple has introduced a new feature into the market which is the downloading of films using iTunes onto available devices. This means that there is now a proportion of consumers who will still prefer to purchase the film, but not necessarily own the physical item and are happy with a digital version of the film.

There is also still a portion of users who do not wish to own the film in any format and will be prepared to rent the film to watch as a one off. We believe that the type of user depends upon the product in question, which leads us back to LD. Different LD products will have differing values and therefore, we must consider the long term use of the data product in question when we consider the type of revenue model suited.

We see the music industry as analogous to the linked GI industry where organisations are holding quantities of data stored in databases which have the potential to generate a large amount of revenue. Once released, music can easily be copied and distributed, and the same applies to data. In order to operate in this world, advances in the revenue models should be investigated to keep up with the advances in technology.

We also note the issues of willingness to pay for music, in a world where free music is readily available (mainly through illegal file sharing)(Guel and Rochelandet, 2006). The model used by the Apple's iTunes is a viable mode of selling music and the micropayment system incorporated here, enables consumers to purchase as much or as little of the music they wish; whereas previously a consumer would have needed to purchase a full album in order to listen to certain tracks.

As outlined earlier, the Long Tail as detailed by (Anderson, 2006) can also be noticed here due to the variety of music available. This suggests that low volume sales of music from unknown or smaller artists may make up the market share from the bigger artists (Dubosson-Torbay et al., 2005). We pose the question as to whether this would be a suitable format for the LD. Would consumers be happy to pay to view the data; or the links to the data or would they prefer to download the data for future use? We suspect that there will be proportions of users who want to have a copy of the data for future reference, but there will also be a proportion of users who just want an instant answer to a query and have no need to refer back to the data or use it for another purpose in the future. We also predict that more and more applications will become apparent, as people build 'mash ups' of data which otherwise would not have been created in the past.

Alongside the music industry we also note the film market. A market industry report

carried out by Key Note Ltd in 2009,¹¹ showed that physical film rental figures had decreased but the existence of firms such as LoveFilm and Netflix have helped to support the film industry by allowing the streaming of films online (Ltd, 2009). They report that the decline in rental sales is potentially due to prices of individual dvds being reduced to a more affordable price for consumers in supermarkets and on auction sites such as eBay. They also note that where strict broadcasting restrictions for television channels have been reduced, more films are viewable via television which has become popular. It is also interesting to note the finding that the ability to temporarily own films via mediums such as iTunes is becoming more popular. We notice that the ‘renting’ of more expensive products such as films is appealing to consumers. Whereas consumers are still keen to ‘own’ cheaper products such as music. In terms of LD, we are interested to find if people are more concerned with the data itself rather than owning it to refer to it again in the future. In order to do this we will consider consumers willingness to pay for data in chapter 5.

We consider this a viable prospect for LD as it would enable users to purchase parts of datasets, rather than the whole dataset which may be deemed useless to the consumer. With this however, we need to consider the pricing of such data items, and maintain the balance between what is economically viable to the data provider but also acceptable to the consumer.

Music, like information goods, is an experience good. Therefore, it is considered that until a user experiences the good, they are unsure of its value. Therefore is the potential for revenue in the data, or is there more potential in the links and added value which the links create rather than in the data itself?

5.5.3 Software

Traditional software licenses are granted by the software publisher under an end-user license agreement but the ownership of the copies of the software still remains that of the owner. This is known as proprietary software.

The first type, which is often the most commonly found, is named user licensing. This is where the license for the software is purchased for a specific user of the software or machine. The second type is a server based license, determined by the number of machines the software is installed on. Alternatively software is available per module, that is a portion of a package of applications. Finally, there is a concurrent user license which is where a number of licenses are available to an organisation and only that number of users can run the software at the same time.

Following on from the traditional software industry, we go on to discuss the Open Source Software, where software is written and when sold, a copy of the source code for the

¹¹<https://www.keynote.co.uk/>

software is released with it. This is comparable with the LOD community, as they give away amounts of their Intellectual Property (IP) but try to retain some control by attaching various licenses to it. We see that this is analogous to LD firstly due to the ability of users to take portions of data, link it to other data and create their own applications and ‘mashups’. We can expect to see this data with licenses attached to it, which we will explore later in this report.

Stallman et al. (2002) outlines the difference between ‘free software’ and ‘Open Source Software’. Free software has no price and is given for free, whereas open source gives away the code for this. In the Open Source world, if you give your product, or in this case software code, away for free e.g. Red Hat,¹² it is the additional features and services attached to this software, which generate the highest revenue and not necessarily the original software. We predict that this is the way in which the future of LD will develop (Rajala et al., 2007). People will give away their raw data for free and organisations will be able to develop tools and applications with this data, which will be the key element for generation of revenue.

Rajala et al. (2007) suggests that revenue can be made from making developments to software and extending it, thus making a collaborative world where software is continually improved by others. This shows an opportunity for users of LD to be creative with the data they wish to link and explore the different possibilities for LD, which can lead to the development of LCD applications; but we note that it is the application which is then closed and not the data itself.

This is not necessarily true for all organisations. Google for example, would not be suited to giving away its code/algorithm for free, as this may inadvertently destroy their business. This may be comparable with the LD industry where organisations may give away their data for free but this may be detrimental to their business. We want to consider this when looking at organisations who are just entering LD market, who are unsure whether it would be suitable to give away all or some of their data.

Raghu et al. (2009) discusses the willingness to pay in an Open Source software environment. Traditional software producers are beginning to find difficulty in operating in an environment where Open Source software and free software is available. Raghu et al. (2009) finds a number of factors which affect the willingness to pay for software. The main factors which affected a users decision to change from paid to free were learning of a new system, format changes from paid software to free and reliability.

We can see again there is proportion of users who will pay regardless of the situation and those who will pay if certain factors affect their decision. When looking at this in terms of LD we need to take this into account. If some data providers start distributing their data for free, where other publishers are charging, there will be a need for companies to ensure they provide a strong competitive advantage.

¹²<http://www.redhat.com/>

Our investigation into three similar industries using linked geospatial data has highlighted a number of areas we believe need further investigation.

Firstly we have highlighted the issues of willingness to pay in an environment when free alternatives are available, how can this be overcome where there are free alternatives? Secondly we have highlighted copyright issues with illegal sharing, what are the options to protect data once it has been published. Finally we address the possibility of the Long Tail effect which means it may not necessarily be the data or large quantities of data which holds the value, but more that applications and possibilities which stem from it which adds the value.

5.6 Experiences With Linked Data

We acknowledge that the business of the [BBC](#) is very dissimilar to that of [OS](#). We visualise [OS](#) to be more of a central point to link data such as that of the [BBC](#) and [DBpedia](#). However we note that the experiences of introducing the [LD](#) technology are similar when it comes to development and integration costs for content management systems which are expensive. These systems require staff to organise and look after the data and require a level of expertise to integrate them with other datasets.

A question which an organisation may have regarding [LD](#) is, how do they start making their business use the technology? An interview with the Technical and development lead for semantic publishing at [BBC News and Sport Online](#) during the 2010 World Cup answers a number of queries ([Milhollin, 2012](#)).

The [BBC](#) has created many sites about its programmes and content, written in [HTML](#). These sites are useful for its followers to get information from but these sites are not linked together in the way we experience with [LD](#).

The [BBC](#) began investigating the use of [LD](#) to better present and share its data. Before [LD](#) the publishing of news aggregation pages specific to a person, athlete, topic or sporting discipline was time consuming and generating links was not possible if done manually.

[Raimond et al. \(2010\)](#) outlines how the use of Semantic Web technologies on the [BBC](#) Web Sites has impacted its business. The [BBC](#) uses [SWT](#) across its Web sites: [BBC Programmes](#) which provides information about its television programmes, [BBC Music](#) about music and artists played on its radio stations and [BBC Wildlife Finder](#) which provides a web identifier for every species, habitat and adaptation the [BBC](#) has an interest in.

Richard Hammond is a television presenter from the [BBC](#), he is well known for presenting [Top Gear](#), but with [LD](#) we are able to present all the categories in which he appears,

for example he also hosts Planet Earth Live. Until the introduction of [LD](#) at the [BBC](#) users were not able to navigate from a page about a programme to a page about an artist played in that programme.

There is however a large amount of community generated data available which can be used to give structure to data for example MusicBrainz is used by the [BBC](#) be for structuring entries of music played on the radio stations. Other sites such as DBPedia can be used to enhance pages with relevant information about topics such as wildlife etc.

The value which the [BBC](#) finds from the use of [LD](#) is in making its content more discoverable and making the user experience in finding the data more accessible. By using and linking to other online resources the cost to the organisation is less as they only need to maintain the data they hold and then link to other sources which have been generated and maintained by someone else, therefore reducing integration and maintenance costs.

The benefits gained from organisations such as the [BBC](#), is not just from having well structure and organised data from within the organisation, but form the potential to create other more beneficial applications from this structured data. [BBC Sport](#) has gained efficiencies by enabling journalists to carry out their main role as a journalist and author content, and letting the automated semantic technologies organise the content on the pages ([Milhollin, 2012](#)).

[OS](#) has a different model to the Met Office. The Met Office has provided weather and climate forecasts fro 150 years. They are again a Trading Fund and operate under the same Trading Fund regulations as [OS](#). However we appreciate that there are similarities and differences with their models. Firstly the technology required for both geographical and weather forecasting is of very high cost and therefore the information gathered by both organisations is high. However, with location we note that forecasting is more valuable when it is attached to a precise location. We note that this is important when connecting data as data connected via a location will enable the user to make other decisions. Accurate weather forecasting has strong implications for everyday life and up-to date and true data can influence important decisions such as events and product placements. Therefore when implemented with data which is linked to other useful data will help increase its use.

We also interviewed Glen Hart the Research Manager at [OS](#) to find out what value [OS](#) are expecting to get from the introduction of [LD](#). In April 2010 [OS](#) released a wide number of products as Open Data. The majority of these products were either raster products or vector products intended to enable cartographic representation. However, three products: Boundary-Line, Code-Point and the 50k Gazetteer were suitable for representation as [LOD](#) and they have been published in both conventional formats and as [LOD](#). The decision to release these data as [LOD](#) was driven by a desire to maximise

the use of Ordnance Surveys Open Data by making it available to not just traditional GIS user but also by new user groups as well. There was also a desire to publish LOD so that this sector could receive a boost. In a similar manner OS has also worked with the University of Southampton and the company SEME4 in a part TSB funded project called RAGLD (Rapid Assembly of GeoCentred Linked Data Applications) to develop a toolset to aid the development of LD applications with a geographic component (<http://www.ragld.com/>). Ordnance Surveys ambitions for the publication of LD go beyond the current released products and LD is now recognised as an important format for products intended for more than just cartographic use. Hence it will be one of the publication formats used by default for future products such as the replacement gazetteer product (currently under development). Work has also been conducted in converting the address holdings to Linked Data, although due to the provenance of these products, releasing this data as LD will also require the permission of GeoPlace, the company that constructs the products.

One benefit to the publication of Open Data is that it may encourage some Open Data users to begin using premium products as well. At present there is no direct link between open and premium products and users of Open Data need to discover the premium products. LOD is seen as a vehicle to enable a freemium model to be more directly implemented with LOD products having implicit links to premium products.

Ordnance Survey Research are also leading a project to investigate the use of LD to underpin future database models in order to maximise the ability to ingest and inter-link new datasets into Ordnance Surveys overall content holding. Here the aim is to significantly reduce the cost and time to acquire new data, something which is extremely difficult with Ordnance Surveys existing databases. These systems have been constructed with the traditional aim of providing efficient data retrieval but have done so at the cost of schema evolution.

For OS, LD is seen as a means to expand the use of Open Data, to help promote the LD Market and to be a potential mechanism to enable a more explicit freemium model. LD is also seen as a possible means to implement internal database solutions where the emphasis is placed on flexibility to easily ingest new data over simple database performance.

Later on in the next chapters we outline how different contexts influence a users decision to pay for data. We reiterate here the difference between raw data as a commodity and data which has been manipulated into an application. Data which is consumed through an application will have differing values to its raw counterpart which introduces an extra dimension to the investigation into the business case for LD.

5.7 Licensing

The purpose of a license is to permit users to use someone else's work, whilst enabling the owner of that work to maintain rights to the work. A license can be restrictive and prevent users from copying or editing the work themselves and passing it as their own.

When looking at online content we must also consider the legal implications of trading online. Goods which although expensive to produce are cheap to reproduce and piracy of these products can be detrimental to the holding organisation. Therefore it is essential that effective licences are available which whilst allowing freedom for users to create their applications also maintain the royalties owed to the holding organisations.

In order for us to understand the legal issues surrounding the [LD](#) we outline in more detail the licenses and their application with [LD](#) below.

Licensing terms for [GI](#) are often seen to be too restrictive. Some licenses allow the re-use of the information freely. Others are limited to just personal re-use and do not allow reproduction of the content for commercial purpose, thus hindering the potential for innovative new products and uses for [GI](#), to become available.

We also note that the Open Licenses, which we outline in more detail later on, are mainly suitable for [LOD](#) and are not specifically suitable to a situation where there is [LCD](#). Therefore, we need to address this issue in chapter 7 where we introduce the architecture for geospatial [LD](#).

Under recommendation 8 of the Power of Information Taskforce Report ([Allan, 2009](#)) two more recommendations were given which are relevant to [OS](#), which will aid the reform of their business.

- Government should ensure that there is a uniform system of release and licensing applied across all public bodies; individual public bodies should not develop or vary the standard terms for their sector.
- The system should create a 'Crown Commons' style approach, using a highly permissive licensing scheme modelled on the Click-Use license that is transparent, easy to understand and easy to use, .

In order to envisage these recommendations, it is important that a clear analysis of the potential licensing policies is detailed in a comparative form, in order to draw conclusions on potential directions for [OS](#) to adopt and promote in the future.

The study of more simple licenses is important due to the complexity and length of current licenses ([Barker et al., 2005](#)). If licenses can be made more accessible to users and the re-use of data encouraged, then organisations will benefit from enabling their

information to be used in ways that would otherwise not happen under current licensing terms.

The report carried out by (Barker et al., 2005) aimed to identify the needs of an organisation, the needs of the potential users of their digital content and to examine the Creative Commons license as a possible licensing solution. This study was commissioned by and carried out for members of the [Common Information Environment \(CIE\)](#) and does not include organisations such as [OS](#) but does however examine [Creative Commons \(CC\)](#) licenses for the use of licensing public sector digital information.

We notice that licensing is particularly problematic for LD as, many datasets, each with their own different license which can be linked together, causing issues for re-use.

5.7.1 Content licenses

- Creative Commons Attribution
- Creative Commons Attribution Share-Alike
- Creative Commons CCZero
- GNU Free Documentation License
- UK PSI Public Sector Information
- MirOS License and
- Free Art License

The data licenses include

- Open Data Commons Public Domain Dedication and Licence (PDDL)
- Open Data Commons Attribution License Data
- Open Data Commons Open Database License (ODbL)
- Creative Commons CCO

From the licenses listed above there are just two licenses which we consider suitable for a [LD](#) framework. These licenses are the Creative Commons CC Zero and the UK PSI Public Sector Information License. These licenses are suitable for [LOD](#) but more restrictive licenses would be required for [LCD](#) which may contain sensitive or confidential information to prevent it being misused.

They are both able to support content and data, whereas the others support only content or data separately. Secondly we choose the UK PSI Public Sector Information License

as there are some bodies who need to make available their data under new government legislation.

We also note that the Creative Commons license is machine readable. Therefore it is suitable for use with [LD](#) and enables the license to be attached easily to the data. The Creative Commons license allows data be licensed in a simple way and can be standardised for particular datasets. For example restrictions can be added for copying, editing and distribution. This will allow the personalisation of licenses for different purposes such as free and commercial use. Creative commons licenses are detailed on the Creative Commons website found here.¹³

5.7.2 Creative Commons Licensing

The Creative Commons license originated in the US in 2001 and enables individuals and companies to grant copyright permissions to their work. The licenses enable the copyright terms to be changed easily from one to another of six types of license.¹⁴

A report by [Barker et al. \(2005\)](#) outlines the [CC](#) licenses and their applicability to public sector organisations in the United Kingdom.

The advantages given in this report by [Barker et al. \(2005\)](#) regarding the use of [CC](#) licenses include

- Ease of use,
- Widespread adoption leading to familiarity,
- Human-readable,
- Machine-readable and symbolic representation of the licences,
- Sharing a common licence with many others,
- A direct link between the resource and its licence.

The most important advantage of the [CC](#) license is that it is available in different forms; a machine readable form and a human readable form. The machine readable form is in [RDF](#) which is perfect for use with [SWT](#) as it enables machine searches to be carried out to discover web pages which are licensed by [CC](#).

The license is encoded in [RDF/XML](#) and describes the license. It gives the name and translations, the description and the properties of the license. This can then be incorporated or linked to the data and will be shown in the [RDF](#) of the data in question.

¹³<http://creativecommons.org/licenses/>

¹⁴<http://opendefinition.org/licenses/>

If a consumer is concerned by the license which may restrict its use, the consumer is able to find this license in the description of the data, which can be returned by carrying out a simple [SPARQL](#) query to search for data which contains a license, or even a specific license.

According to the recommendations given in the report, [CC](#) licenses should be used wherever possible and where it is not possible the first choice after [CC](#) should be Creative Archive or Click-Use licenses. The aim of using only a small number of licenses is to make it clearer for users to understand the terms of these licenses, rather than having a wide variety of licences, each with different terms. If a customised license must be used the directive suggests that this license should still be based as much as possible on the [CC](#) license, again keeping the licenses accessible and understandable by its users.

The report by [Barker et al. \(2005\)](#) states the baseline features of the [CC](#) licences as:

- Licensees are granted the right to copy, distribute, display, digitally perform and make verbatim copies of the work into another format.
- Licensees may incorporate the work into collective works (that is when the work, in its entirety in unmodified form, along with a number of other separate and independent works, is assembled into a collective whole).
- The licences have worldwide application that lasts for the entire duration of copyright and are irrevocable.
- Licensees cannot use technological protection measure to restrict access to the work.
- Copyright notices should not be removed from copies of the work.
- Every copy of the work should maintain a link to the licence
- The rights holder must be attributed.
- The work must not be subjected to any derogatory treatment as defined in the Copyright, Designs and Patents Act 1988.

The six licenses are shown in decreasing restrictiveness and details of each license are outlined below.

- 1. Attribution Non commercial No derivatives (by-nc-nd)** The most restrictive license and allows redistribution. As long as the users of the restricted work mention the owner they are able to download it, however they are not able to change it in any way or use it commercially.

2. **Attribution Non-commercial Share Alike (by-nc-sa)** Unlike the previous license, this license allow others to build upon the restricted work by changing or remixing it, all derivatives will however remain non-commercial as with the previous license.
3. **Attribution Non-commercial (by-nc)** This license enables users to change and build upon restricted work and need to acknowledge the owner; derivative work does not need to be licensed in the same terms.
4. **Attribution No Derivatives (by-nd)** Redistribution of commercial and non-commercial nature is allowed as long as it is not changed and credit is given to the owner.
5. **Attribution Share Alike (by-sa)** Commercial modification of the work is allowed as long as credit is given to the owner. It is similar to open source software. All new work will carry the same license and will therefore allow commercial use. Derivative works can be made.
6. **Attribution (by)** Others can distribute and change the work as long as credit is given for the original work. It is the most accommodating of the licenses in its ability to let others use restricted works.

5.7.3 Click-Use Licenses

The Click-Use License was an online license for the reuse of a variety of Crown Copyright material. In 2001 the [Office of Public Sector Information \(OPSI\)](#) introduced this license in two types: the Public Sector Information License and the value added license. The PSI License covers the information that is central to the government process. There are no charges made for the re-use of this information. The value added license covers the value added material produced by the government and charges are made for this information.

The Click-Use license was not specific to one set of information and once a user applied to the OPSI for the license, they were able to use a wide range of information under the same license. The license lasts for up to five years and permits users to reproduce and publish the material but does not allow it to be modified ([Barker et al., 2005](#)).

According to [OPSI \(2009\)](#) there were 17,934 Click-Use licenses in use as of the 30th June 2009. The Click-Use license enabled the opening up of PSI to a wider global audience than has previously been possible.

The Click-Use License has now been replaced with the Open Government License which we outline in more detail below.

5.7.4 OS OpenData License

This license is specific to OS OpenData™ and also incorporates the Open Government Licence for public sector information which is outlined in the next section. The licence governs access to and use of OS OpenData™ made available at <https://www.ordnancesurvey.co.uk/open> and at <http://data.ordnancesurvey.co.uk>.

The license allows users to make use of the data in any way but with the following restrictions:

- Acknowledge the copyright and the source of the data by including the following attribution statement, Contains Ordnance Survey data Crown copyright and database right 2011.
- Include the same acknowledgement requirement in any sub-licences of the data that you grant, and a requirement that any further sub-licences do the same.
- Ensure that you do not use the data in a way that suggests Ordnance Survey endorses you or your use of the data.
- Ensure that you do not misrepresent the data or its source.

This license was introduced in January 2011 to ensure that there is just a single set of terms for people to use freely available government information. The license also means that the developers do not need to apply for a license to create applications, which can be translated to its users as they will be able to enjoy full benefits from the applications being created such as unlimited access to applications without restrictions on use.

5.7.5 Open Government License

The Open Government License for public sector information which was produced by The National Archives¹⁵ enables and encourages people to reuse information under a small number of conditions. The license enables people to copy, publish, distribute and transmit information, adapt and exploit information commercially; for example combining it with other information or using it in a product or application. This license is particularly suitable for LD applications where users may wish to combine datasets from a number of different sources. In order for people to use this license they need to ensure that a suitable link is created to the license to ensure that people are aware that the data has been reproduced or used under a license.

¹⁵<http://www.nationalarchives.gov.uk/doc/open-government-licence/>

5.7.6 Public Service Mapping Agreement

The [Public Sector Mapping Agreement \(PSMA\)](#) is a contract which enables the provision of core mapping data to the public sector. The agreement became available in July 2011 and has meant that various products are now freely available which were previously available but with strict conditions on reuse. This agreement means that collaborations between public sectors bodies can be created. This will enable the creation of [LD](#) mash ups which previously would not have been possible. It does, however, still leave the issue of use of data for non public sector bodies. Therefore we still see the requirement for investigation into the provision of data for commercial use.

5.7.7 Comparison of the old Click-Use licenses versus Creative Commons Licenses

Table 5.5 illustrates the basic differences between the two licences.

Click Use	Creative Commons
Licence available for an array of information	Licence only available to selected information
Valid for 5 years	Valid for entire duration of copyright
Does not permit modification	Certain licences permit modification

TABLE 5.5: Comparison of Click use licenses versus Creative Commons Licenses

After an online survey was completed by online users, work began by [OPSI](#) to develop a new licensing model which will enable greater interoperability of licenses ([OPSI, 2009](#)). The licensing model will enable other license users such as Creative Commons and [General Public License \(GPL\)](#) to re-use government information more easily.

5.7.8 Open Data Commons Licenses

It has been noted that the [CC](#) licenses are not particularly suitable for use with data [Miller et al. \(2008\)](#). Rufus Pollock from Cambridge University ¹⁶ suggests that the licenses such as the non-commercial [CC](#) license makes it unfeasible to create ‘derivative’ works. He suggests that other licenses such as the [Open Data Commons \(ODC\)](#) licenses are more suitable. The open data commons provides two licenses which were created for data and databases. ¹⁷

¹⁶<http://www.rufuspollock.org/>

¹⁷<http://www.opendatacommons.org/>

1. [Public Domain Dedication and License \(PDDL\)](#)

This license is intended to allow users to freely share and modify work for any purpose without any restrictions. ¹⁸

2. [Open Database License \(ODbL\)](#)

This license allows users to share, create and adapt as long as the user attributes any public use of the database or any works produced from the database. If the user publicly uses any adapted version of the database or works the adapted database must also be offered under the [ODbL](#). If the database is redistributed then technological measures may be used to restrict the work as long as a version without such measures is also distributed.

The key point to note in this license is that the contents of the database are not covered with this license and that users should combine this license with others to protect the contents. ¹⁹

The [ODC](#) licenses are viewed to be a much more suitable method of licensing data from databases as they were created for use with databases and allow issues such as creating derived works to be clarified. These licenses are clear and accessible to understand and in terms of enabling data to be re-used, shared and linked on the web are a satisfactory alternative to the [CC](#) license.

5.7.9 Open Street Map Licensing

[OSM](#) have recently introduced the use of a new license called the [ODbL](#). This license allows users to freely share, modify and use the database while maintaining this same freedom for others. This license is one of the two licenses created by the [ODC](#)

[OSM](#) has chosen to adopt this license over other potential licenses due to their belief that their licenses are not suitable for the licensing of data and databases. These unsuitable licenses include the [GPL](#), the GNU free documentation license (GFDL) and the attribution share alike license (CC-by-SA) by the Creative Commons.

5.7.10 Licenses used on Data from the Linked Data Cloud

Table 5.6 illustrates the key datasets with the most links from the Linked Data Cloud and the licenses used. We can see from the table that the most commonly used license is the Creative Commons Attribution license. We note that some datasets do not specify licenses and although the data is free, the data still needs to hold and display some form of licensing in order for users of the data to be able to use it. We consider there

¹⁸<http://www.opendatacommons.org/licenses/pddl/1.0/>

¹⁹<http://www.opendatacommons.org/licenses/odbl/1.0/>

may also be a problem here with linking data which has different license terms. How will developers of applications give access to the applications if some parts of the data used are held under more or restricted licensing terms to the rest of the data in the application?

Data	How Licensed
MusicBrainz (http://musicbrainz.org)	Core data The core data of the database is licensed under the CC0 license Supplementary data - The remaining portions of the database are released under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 license
Ordnance Survey (http://data.ordnancesurvey.co.uk)	OS Open Data Licence compatible with Creative Commons Attribution License (cc-by)
DBPedia (http://dbpedia.org)	Creative Commons Attribution-ShareAlike License and the GNU Free Documentation License
RKB Explorer (http://www.rkbexplorer.com)	Non specified
The Gene Ontology (http://www.geneontology.org)	No licensing requirements
Freebase (http://www.freebase.com)	Creative Commons Attribution Only (CC-BY) license
Geonames (http://www.geonames.org)	Creative Commons Attribution 3.0 License , Creative Commons Attribution License (cc-by)

TABLE 5.6: Key Linked Datasets and their Licenses

5.8 Summary

In this chapter we have considered **PSI** and **GI** as the key subject of linking data in this thesis. We have explored the emergence of datasets from the public sector and how these differ from commercial data providers. We also showed further detail on **OS** and the products which they provide in **LD** format. We also detail a number of other public sector organisations who have begun to consider the use of **LD**.

Chapters 2, 3 and 4 reviewed the literature surrounding the key themes in this thesis. Chapter 5 introduced a chapter about **OS** which outlined its experiences with **LD**. From this literature we were able to highlight further key issues we believed to be necessary to investigate in more detail. Firstly we have acknowledged through the literature that there are two types of users of data online, these are commercial and leisure users. Do these users have different buying behaviours when making data purchases online? Following this if there are two types of data available, free and paid, will people be still willing to pay for data given that free alternatives may be available? Further to this we would like to be able to illustrate what a potential **LD** application look like with open data and data which requires payment.

We have raised a number of issues and questions in this section and the next chapters aim to answer some if not all the questions posed so far. Where we are able to answer these questions, backed up with empirical research, we have stated this, but where we feel that the area requires more investigation or the answer is unclear, we suggest further recommendations, which are outlined in the final chapter of this research.

Chapter 6

Requirements Elicitation

In chapter 2 we introduced the technical side of this research which included the technologies used for **LD** and the **SW**. Following this in chapter 3 we investigated the issues which would affect the business of linked data. Chapter 4 then outlines the business case and issues for **LD**. This review of the literature has established that there are a number of factors which require investigation in order to understand how to derive value from **LD**. Specifically, we have found that we need to take into consideration the opinions of users when considering potential revenue models and their willingness to pay for a new technology or concept such as **LD**. We also note that the characteristics of the product such as quality and value added also need to be taken into consideration.

The next three chapters use the findings from the literature in chapters 2 and 3 and 4 and 5 to inform a number of preliminary investigations to further understand the landscape for the business of **LD**.

We carried out three separate investigations, each one with a different aim. The first investigation looked at the types of users of geospatial information in order for us to understand who may be the users of the data. The second investigation aimed at bringing together the two communities in order to inform them of the possibilities of linked, geospatial data. Finally, we carried out a pilot study to investigate the appetite for **LD** applications. The aim of this experiment was to understand the reasons why people chose to use different types of data when there are free, paid and premium options available.

6.1 Ordnance Survey Open Space Investigation

Our first investigation looked at the consumers of **OS** OpenSpace. The **OS** OpenSpace API¹ is a free service which allows non commercial users to build web applications with

¹<http://openspace.ordnancesurvey.co.uk/>

OS data embedded in them. Firstly, we took the database of consumers who have registered with the OpenSpace API. We analysed the dataset containing registration information for 3275 users who had registered to use the OpenSpace API up until December 2009 and categorised them into 27 different groups. In order for us to separate the users into different groups we classified each user depending on how they registered for the API. Users who stated that they had some commercial relation or were experimenting with the data for commercial purposes were grouped into various different commercial groups. Next, users were put into separate groups if they came from a council or other local authority. All non-commercial users were identified and classed into respective categories and then finally all users who did not give any credentials were classified as undisclosed (See Table 6.1). The types of users and their applications is an important place to start when looking at the business model as it is the customers who will generate the income for the organisation, and without them the business model will be of no use.

We must note that given although only 3.5% of the total number of registrations was for councils, this could in fact represent a large potential user base compared with the numbers of individuals registering for personal use.

Once the user type had been grouped, we looked at the types of applications the users had registered for. The activity was classified into one of five different categories to determine the types of applications users required of the data (See Table 6.1). The five different categories were split up depending on whether or not a user had a specific experiment in mind. If the use type was specifically for experimentation, following this another classification was included which could expand on the type of experiment detailed. Users who gave no specific experimentation details were grouped into non-specific experimentation. Users who were more specific about their experimentation and those who were using the API for an educational purpose or for academic research were given separate groups. Many users who stated their use as non-commercial were grouped into service provision where they were experimenting with the API to provide a non-commercial service often for walking or other outdoor activities. The remaining users who did not fill in this section were classified as undisclosed.

6.1.1 Results

Of 3275 users, 44% did not disclose their organisation; specifically they did not enter anything in this field when registering for the API key, showing a high proportion of users were unwilling to divulge their identity. This could have been for a number of reasons. They may have been from rival organisations, or did not have an organisation and required the information for personal use only. 17% registered for the API for 'personal' use. These were people who identified their use was purely for a personal website and did not disclose any affiliation to an organisation. The uses for this category were mainly

User Type	Users	Percentage
Commercial	393	12%
Consultancy	122	3.7%
Developers	104	3.2%
Service Providers	43	1.3%
Other Commercial	124	3.8%
Governmental	115	3.5%
Borough Councils	22	0.7%
City Councils	6	0.2%
County Councils	23	0.7%
District Councils	14	0.4%
Local Authorities	24	0.7%
Town Councils	6	0.2%
Parish Councils	20	0.6%
Education	138	4.2%
Schools and Colleges	40	1.2%
Universities	98	3%
Non-Commercial	1144	35%
Other Non Commercial	6	0.2%
Charities	55	1.7%
Clubs and Societies	321	9.8%
Political Parties	18	0.5%
Developers	29	0.9%
Local Communities	85	2.6%
Open Source	9	0.3%
Personal	561	17%
Religious Organisations	23	0.7%
Emergency Services Official	19	0.6%
Emergency Services Volunteer	10	0.3%
Healthcare	8	0.2%
Undisclosed	1435	44%

TABLE 6.1: Classification Of Open Space API Users

Use	Users	Percentage
Educational Purpose	47	1.4%
Non-Specific Experimentation	567	17.3%
Specific Experimentation	1291	39.4%
(Non-Commercial) Service Provision	462	14.1%
Academic Research	105	0.3%
Undisclosed	766	23.3%

TABLE 6.2: Classification Of Open Space API Use

for route planning, interest in APIs or for blogging purposes including tagging pictures with a location.

These users often detailed their use for activities such as walking, cycling and other outdoor leisure pursuits. 9.8% of individuals registered for the OpenSpace API on behalf of a club or society. Again the uses given for this category were mainly for walking, cycling route planning or directions to other outdoor activities. 12% of users signed up for a commercial reason, and were grouped together for all types of commercial purpose. Just 3.5% of users were from the various different councils: town, parish and county councils. These small numbers of commercial and non-commercial users demonstrates that a considerable number of new users of LD could be attracted by a more suitable business model.

We also note that the low numbers of governmental users may well be due to the fact that they already have good access to premium data through their government agency licenses and agreements (for example the PSMA).

Table 6.1 illustrates the proportion of users who registered for the API depending upon the type of use they proposed. The majority of users registered with a specific experiment in mind. The most common experiment was for route planning. Location pinpointing was the second most popular use and was often associated with leisure activities or hobbies.

Furthermore, a significant number of users did not disclose the specific use of the data. This could be due to users experimenting with the data with no particular aim or due to not wishing to disclose their intent.

Of the users who stated a website attached to their registration details, we noticed that very few (less than 10%) had actually used the maps on their proposed website. We acknowledge that many of the users who registered for the API did not take their registration any further or in fact looked at the API and found it was no use for the purpose they signed up for. We also note that some of the reasons people signed up for the service may not have included the use of the maps on a website and therefore we would be unable to see their use from their registration details.

6.1.2 Discussion

The percentage of users who are non-commercial users versus the number of users who will make a commercial purchase in the future is significant in these results. This study was created after the release of Open Space but before the release of Open Data. These users can be noted as an untapped pool of potential paying users and should be taken into account with the consideration of potential LD products in the future. The number of users is over double for non-commercial purpose which gives evidence in support of

exploring the freemium model further as it will encourage more free users and potentially capture a wider commercial audience than previously. This will only be so if the pricing of the premium model is reviewed in correlation with the data available through the use of new technology. This investigation looked at OpenSpace and has not reviewed the use of OpenSpace Pro. To gain a clearer understanding of the types of users and types of use it would be beneficial to review the same details from OpenSpace Pro version of the API. We have made the assumption that the number of registrations with the API correlates with the API usage, as users will be unable to use the API without registering for it and that API usage is an indicator of the future of LD usage.

Having classified the users first detailed use, we felt it was interesting to see the other uses listed. Often the user stated it was for a non-specific experiment in which case their second use listed was 'none'. However, a number of users detailed exactly what the data would be used for and we were able to outline this further. In total 13% of users specified their intentions for route planning purposes and a further 6.3% specified the use of the API for location pinpointing. Of the remaining results 2% of users were experimenting with the API for comparison with other free data (such as OSM). This is a small proportion of users which could be increased to attract more customers if the data was available to them more readily and the business model was changed. The new model must be able to encourage new users and give them more opportunity to experiment. The results of the second use supplement the previous table demonstrating the highest proportion of users intending to use the data for route planning or other local investigations. This high proportion of users for non-commercial purposes show that users were utilising the API for personal use and there is potential here to gain a wider customer base if the data is made available to a broader range of customers.

The results of the preliminary experiments show that the greatest proportion of users interested in the API are personal users or from clubs and societies, despite the number of users who have detailed they are experimenting with the data with a potential to use or purchase further data in the future. The greater proportion of users in this group will not be prepared to pay for OS data which they require and will therefore look elsewhere for alternatives which are free. There were a number of users who have been clear that they are using this API as a comparison with other geographical data providers highlighting a key point. Giving away a proportion of the data encourages or stimulates demand for the data and may encourage users to make future purchases. The widespread adoption and use of OS data by free users stands to benefit both OS and non free users through the integration with third party data derived from OS linked GI.

OS define derived data as data created by the licensee that has used Ordnance Survey Digital Mapping Products in its creation ([Ordnance Survey, 2011](#)). Derived data must be considered when addressing the feasibility of the freemium model for linked GI because despite giving data away for free, the licensing issues surrounding derived data and GI may prevent the experimentation and development of applications.

The [PSMA](#) as we stated earlier enables other public sectors organisations to publish data without the restrictions previously experienced. We also note the introduction of free versions of [OS](#) data will, although hold licenses, will not hold the same strict restrictions as found with some derived data. It does, however, still require licenses to be included. We explored the possibility of licenses attached to the data in the previous chapter. We also suggest further use of licenses in the technical framework in chapter 8, which discusses a technical framework for the consumption of [LD](#).

6.2 Terra Future - Forging Links Seminar

In the previous section we carried out a short investigation to classify the types of users of the the [OS](#) OpenSpace API. This investigation provided us with details of the types of users of [OS](#) data and the uses which the data was put to.

Following our review of the literature surrounding [LD](#) and the relevant technologies and the previous investigation, we felt it was important to understand potential users' initial thoughts on the introduction and use of a new technology. One of the key findings from the literature in this section was that understanding of a technology was often a factor for not being willing to pay for a products or information online. Therefore, we felt it would be beneficial to create a situation where users could find out more about the technology and we could use this opportunity to find out more about their concerns and opinions of using a new technology. The main aim of the Terra Future - Forging Links Seminar was to bring two communities together, the [LD](#) community and the [GI](#) community. We saw the seminar as a way to answer some of the questions we raised in chapter one.

We understand that the [GI](#) community may have limited knowledge of the concept of [LD](#) or some may know nothing. Users of [GI](#) have an extensive understanding of the types of applications which would be useful and interesting. Comparably, [LD](#) practitioners have the technical understanding to put ideas for applications into practice. We felt that in order to gain insight into the types of applications which people hope to build, a seminar drawing the two communities together would enable us to interpret the obstructions each community faces when utilising a new technology.

6.2.1 The Approach

We organised the day into three parts. The first part of the day was a series of presentations by [LD](#) practitioners who explained the role of [LD](#). The presentations aimed to give the [GI](#) community an understanding of [LD](#) and how it may be applied. This session also gave the [LD](#) community an insight into the work currently being done.

The second part of the day took the form of two workshops geared specifically to the two communities. The first workshop demonstrated the potential of LD and the second workshop was targeted at the GI community giving both communities an opportunity to see the possibilities from each others perspective.

The final part of the day was a group idea generating session. The participants were split into 20 groups of 7, where half were from the GI community and half were from the LD community. Each table had a facilitator who helped encourage and guide the discussion to ensure that key questions were posed to the group to give us an understanding of the issues they face. The facilitator was briefed before the event and given a number of questions to put to the group.

We found a number of key topics arose during the afternoon discussions. These were:

1. Licensing
2. Cost
3. Technical ability

These discussions enabled users to give details of the issues which they felt were a priority having spent the morning learning about the possibilities of LD.

Many users voiced their concerns about how the data is licensed in terms of re-use within applications. Their concerns then turned to how much the data would cost if they wanted to use it for a commercial purpose.

Finally the non-technical users found that although they understood the concept of LD and had an understanding of the technology they were unsure if they would be able to replicate a LD application in the future without any further knowledge or training.

We found these round-table discussions useful to this research to inform us of the current issues of uptake of LD for non-technical users from other fields who may be interested in its implementation.

6.2.2 Conclusion

The Terra Future event aimed to introduce new users to the concept of LD and introduce users of LD to the issues of using GI in LD. From this event we were able to confirm the issues which users have regarding the use of LD as a new technology. We reiterate here that although applications built using LD would have LD in the background or be built on LD, users may not be aware of this and not notice any difference in the applications being used.

Following our review of the literature and our initial investigation into users of OS GI we have found that there are a number of issues which need to be addressed in order for us to understand the opportunities to derive value from linked geospatial information. These key areas are licensing, cost and revenue and technical ability.

6.3 Investigation - Information Quality Criteria Questionnaire

Initial work carried out in the previous section aimed at discovering the areas of concern noticed by potential consumers of LD. The results found that there are a number of factors which may prevent the full uptake of a potential LD product or service. Therefore, we decided that the next section of research should investigate the factors which may influence a decision to pay for data. We aimed to see which are the most important factors affecting consumers decisions to be able to inform data providers of the factors which they must consider when preparing to sell their data in LD market.

From our initial investigations into users and LD, we found that there were a proportion of participants who were not willing to pay for information, some who are willing to pay for information and some who were prepared to pay a premium price for information. Following this we felt that it was important to understand the specific attributes of the information which affect their decision to pay for the information.

As we mentioned earlier, there are users who will be unaware that the data they are consuming is in fact LD. We have chosen to test factors which are considered relevant for the consumption of ordinary data on the web as the same trust issues and concerns will be applicable. We have chosen the most commonly occurring factors from literature concerning information quality on the web (Eppler et al., 2003; Knight and Burn, 2005; Naumann and Rolker, 2000; Caro et al., 2005).

We wanted to investigate if there was any relation between the decisions made by participants and see if there was a preference between one particular criteria over others.

6.3.1 Experimental Design and Methodology

For the purpose of determining which factors are of most importance to consumers we decided that a survey would be the best way to establish these which of these criteria were most important. A survey was designed using Survey Monkey (See Appendix C) to determine the attributes of information people find most important to them when searching for information or data on the web.

From the literature review specifically in chapter 4, we have been able to identify that there are specific attributes of Information Quality (IQ) which may affect consumers

decisions to pay. We have narrowed these down to six criteria which we used to test which were most important.

We have defined each of the criteria below. We also note that this is our interpretation of the meaning of the terms and have specified this in the questionnaire to ensure that each consumer is aware the meaning we give for each term.

Accurate - The extent to which information is correct, reliable and verified free of error.

Consistent - The extent to which information is presented in the same format and compatible with previous information.

Secure - the extent to which access to information is restricted to maintain its security.

Timely - The extent to which the information is sufficiently up-to-date for the task at hand.

Complete - The extent to which information is not missing and is of sufficient breadth and depth for the task at hand.

Concise - The extent to which information is compactly represented without being overwhelming.

We decided to ask participants to select their preference for each criteria rather than ask them to order them in numerical order as different participants may not rank them using the same measure. For example, one person's view of a 5 on a scale of 1 to 10 may be different to another. We found that the most efficient way of accurately analysing the data was to use the Chi square test to test if each response was independent. Therefore we listed all of the possible pairs and in order to ensure that we did not introduce any bias into the questionnaire we made sure that the order of the data was randomised for each participant. This meant that the criteria would appear differently for each participant, ensuring that the results we generated from the study were not just the result of participants selecting the first option from a list.

The null hypothesis for this study is outlined below:

H_0 : There is no significant association between the variable A and variable B

- H_0 : Variable A and Variable B are independent.
- H_a : Variable A and Variable B are not independent.

When we refer to Variable A and Variable B we are illustrating the possible combinations of quality criteria. For instance Variable A - Accurate and Variable B Complete.

6.4 Participants

100 participants were recruited from a bank of participants held by uSamp.² We asked that the participants be geographically located to the UK only and could be of any age over 18. A total of 99 participants completed the study. When participants were recruited by uSamp, they came from a wide geographic area across the United Kingdom, to ensure that there was no bias towards the use of words from differing areas of the country. We restricted the study to only participants from the United Kingdom.

6.5 Results and Statistical Analysis

The results of the experiment are displayed below. Table 6.5 shows the results of the questionnaire. It shows the proportions of participants who selected each response.

Option	First Attribute	Second Attribute
Accurate or Timely	83	16
Accurate or Consistent	79	20
Accurate or Concise	75	24
Accurate or Complete	74	25
Consistent or Concise	73	26
Complete or Concise	69	30
Secure or Timely	66	33
Secure or Concise	66	33
Complete or Timely	63	36
Consistent or Timely	63	36
Secure or Complete	60	39
Accurate or Secure	59	40
Complete or Consistent	58	41
Secure or Consistent	55	44
Timely or Concise	47	52

TABLE 6.3: Results of Information Quality Questionnaire

Following the experiment a statistical analysis of the results was carried out using the SPSS statistics package. Table 6.5 shows the results of the statistical analysis. The table illustrates the option, the Chi square result, the degree of freedom and finally the asymptotic significance. Following this Table 6.5 details the results of the statistical analysis to clarify which responses had a significant association. See Appendix D for full details of the results.

²<http://www.usamp.com>

Option	Chi-Square	df	Asymp. Sig.
Accurate or Secure	3.646a	1	0.056
Accurate or Complete	24.253a	1	0
Accurate or Consistent	35.162a	1	0
Accurate or Timely	45.343a	1	0
Accurate or Concise	26.273a	1	0
Secure or Complete	4.455a	1	0.035
Secure or Consistent	1.222a	1	0.269
Secure or Timely	11.000a	1	0.001
Secure or Concise	11.000a	1	0.001
Complete or Concise	15.364a	1	0
Complete or Consistent	2.919a	1	0.088
Complete or Timely	7.364a	1	0.007
Consistent or Concise	22.313a	1	0
Consistent or Timely	7.364a	1	0.007
Timely or Concise	.253a	1	0.615

TABLE 6.4: Results of Statistical Analysis

Option	Significant Association.
Accurate or Secure	No
Accurate or Complete	No
Accurate or Consistent	No
Accurate or Timely	Yes
Accurate or Concise	Yes
Secure or Complete	Yes
Secure or Consistent	No
Secure or Timely	Yes
Secure or Concise	Yes
Complete or Concise	Yes
Complete or Consistent	No
Complete or Timely	Yes
Consistent or Concise	Yes
Consistent or Timely	Yes
Timely or Concise	No

TABLE 6.5: Significant Association of Results

6.6 Discussion

In the previous section we outlined the results of the questionnaire. From these results we see that out of the 15 pairs of quality criteria, 9 of the options showed a significant association. We chose to use the Chi Square test to determine if there was a significance between the choice each participant made for each pair of criteria. We wanted to determine if each participant just chose each result by chance or for a reason.

The results from this questionnaire disproves our hypothesis which says that there is no significant association between the two criteria. When we state significant association we suggest that the choices made were not due to chance alone and that there is a

significance in the results gained from the questionnaire. The remaining 6 options showed no significant association. We will discuss the significant results in further detail below.

When asked to rank accurate over another criteria it is clear to see that when put against any other option accurate is always the preferred choice. This backs up our findings from the literature which also suggests that quality is the most important factor to consumers when considering the selection of data online. We also note that when complete was ranked against others we found that it was the preferred choice followed by consistent. Table 6.6 below highlights the preferred criteria.

Option	Preferred choice
Accurate or Timely	Accurate
Accurate or Concise	Accurate
Secure or Complete	Secure
Secure or Timely	Secure
Secure or Concise	Secure
Complete or Concise	Complete
Complete or Timely	Complete
Consistent or Concise	Consistent
Consistent or Timely	Consistent

TABLE 6.6: Preferred Quality Criteria

We find that accurate, secure, complete and consistent are the criteria people find most important to them when considering data. From this we can draw our conclusion. The results from this questionnaire will be used to inform a further study regarding the actual use of data on the web.

6.7 Conclusion

Following our previous investigation which aimed to determine if consumers were willing to pay for data at all online, we decided that we also needed to ask questions to inform the factors which providers of data need to consider when publishing data from which they wish to receive revenue.

This questionnaire has enabled us to explore the key factors which consumers look for in the search for data online. We chose the criteria which we felt were the most relevant criteria for LD and from the questionnaire we were able to narrow the 6 criteria down to the most important.

1. Accurate,
2. Complete,
3. Consistent.

Through this investigation we have noticed that the term ‘accurate’ could be considered to be an umbrella term for the other attributes such as complete and consistent. We conclude that participants may have chose accurate when they in fact meant complete or consistent. We treat this result with caution in our further investigations as, although we gave specific definitions for each of the attributes we cannot be sure that the participants took the meaning of the criteria as we intended.

Security is one term which was ranked highly. Security of data or resources on the web specifically is a large area which could be potential basis of further research into LD in the future. For the purpose of this next experiment we did not choose to specifically model or test ‘secure’ data, as security is one feature which we would expect to be part of any online resource we use on the web. This does not however mean that it is not an important part of data.

6.8 Summary of Research Contributions

The requirements elicitation enabled us to begin our investigation into the potential to consume LOD and LCD. We wanted to determine the key features which would be desirable to consumers which are specifically non technical at this stage. The key areas we wanted to explore included the types of users, the potential to charge for data and the quality factors of the data in question.

Firstly we looked at the current users of the OS OpenSpace in order to find out the types of users of OS data. From this we found that there are three potential types of users of linked geospatial data and that a potential LD application needs to be able to support each potential consumer type.

We also not that with the introduction of a new technology, it is important to ensure that potential users are able to visualise how the technology will work and in turn, how it will affect them. We describe how we achieve this in the next chapter which explains the two specific LD experiments we implemented.

From the requirements elicitation we also found that there are certain quality factors which need to be considered to determine if they will directly affect a consumers willingness to pay for information. We use the factors highlighted here, in the design of the following experiments to test how consumers would act in a LD situation.

The key points we have raised so far through this research include:

1. Information quality is important when considering the building of applications using LD

2. Informing participants about the ease of the use of LD is essential to ensure that LD is accessible to all types of participants, not just technical participants and experts.
3. In order to understand the landscape for linked geospatial applications it is important to investigate the current applications and demonstrate the possibilities for future applications.
4. The type of participants of GI varies. This is important when targeting a new product or application, as some participants (commercial users) will be more prepared to pay a premium price whereas others (leisure users) may not be prepared to pay much more.

The following chapter describes the experiments we designed and implemented in order to find out more about the landscape for the consumption of [LOD](#) and [LCD](#).

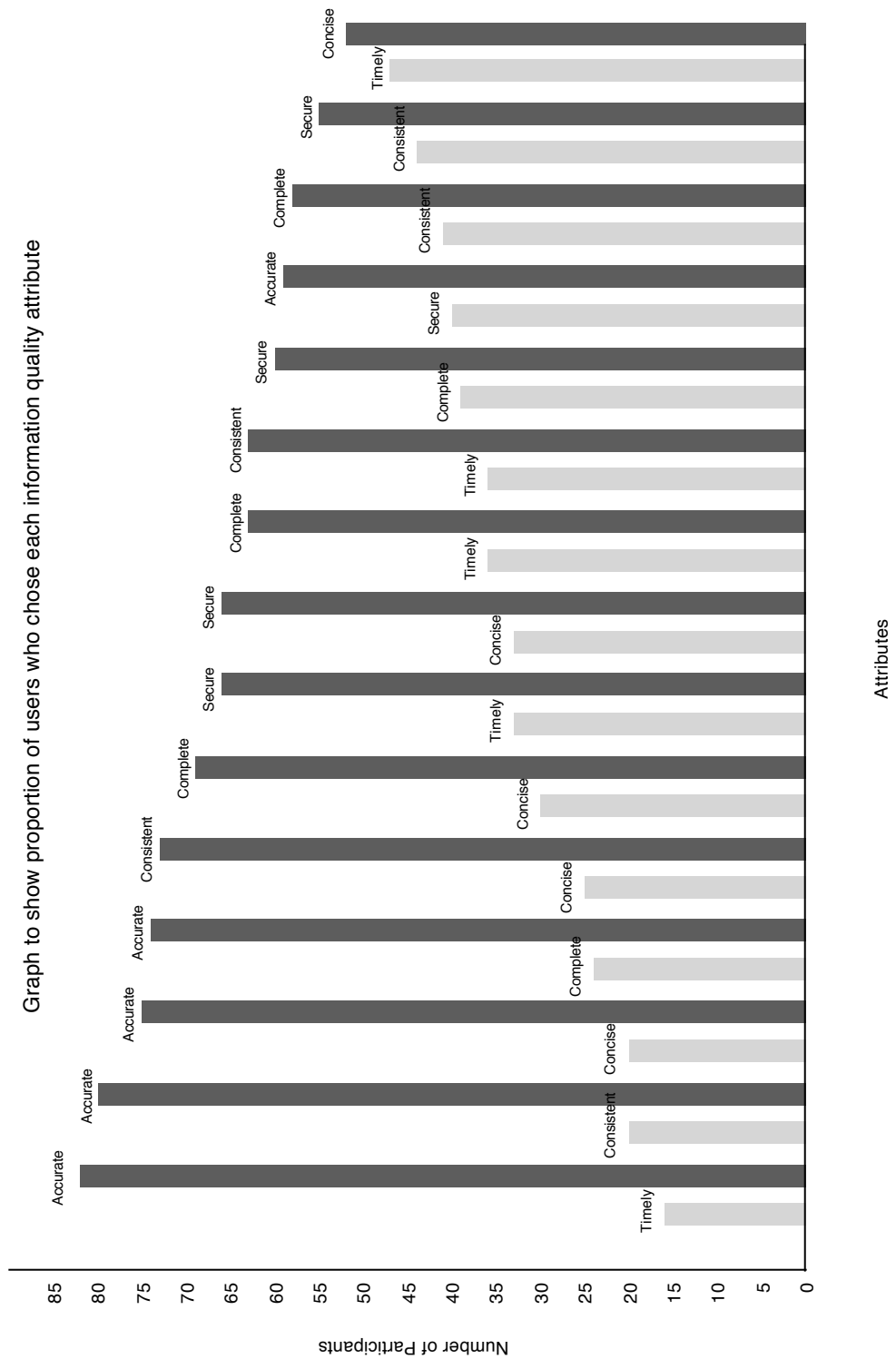


FIGURE 6.1:

Chapter 7

Linked Data Investigations

7.1 Linked Data Experiment 1

The literature in the previous sections and the investigation into the users of OS OpenSpace followed by the Terra Future event, highlighted a number of areas which we feel further investigation will contribute to the development of models for carrying out business and generate value from LD. The experiments were used to study how users interact with information provided by a LD application. More specifically we looked at the propensity of users to use the data. It should also be noted that as previously mentioned in section 2.3 LD is not intended for direct consumption by people as outlined by Berners-Lee et al. (2001), therefore we did not expose the users to the raw LD, or explain the technology behind the applications in detail. From our observations of how people use applications or consumer information based on LD, we draw conclusions on the potential purchase of LD.

The hypotheses for this experiment are:

H₁ Consumer willingness to pay for data is affected by information quality factors.

H₂ Willingness to pay for data will affect consumers product choice.

7.1.1 Experimental Design and Methodology

In order to answer some of the questions generated from the literature and preliminary work we have built a working prototype of a LD situation.

This study aimed to demonstrate the potential of LD and to open up thinking about the possibilities for LD applications on the web. Further to this, we aimed to discover the appetite for LCD by observing participants willingness to pay for premium data. We will explore the factors that affect the decision whether or not to pay for premium

data, or continue to use free data. The key question to answer in this study is: What criteria make participants chose premium data over free data?

The study was a server side application written in Python programming language, which was used to generate dynamic webpages. The script was executed on the server and so was not visible in the browser. We were able to customise our response based on the users input or response enabling different scenarios to be modelled.

This study was carried out online to ensure that all participants were able to participate regardless of their geographic location and also to ensure that the study simulated a real web browsing environment, similar to that in the study carried out by [Lopes and Galletta \(2006\)](#). The LD simulation enabled participants to browse through specific information using a hypothetical scenario, incorporating LD. (Appendix B for screen shots). The participants were provided with an initial bank balance. Premium data was given a value and the bank balance was reduced as they spent more. Premium data was considered in this case to be data which contained the most fruitful data. On purchase of premium data participants would be able to determine the optimal answer. Paid data would give the participants the correct answer but may have not contained all of the details. The data was prepared so that the optimal answer was most easily available by paying either for the premium data or the paid data which was not always complete.

The free version of the data had a number of factors which may have affected the participants choice. These factors included correctness, completeness and timeliness, which may have persuaded them to purchase the paid or premium data. The option was given to decide whether they would use free data, paid data or a premium dataset in the execution of their task.

Following the scenario, participants were asked to complete a short questionnaire to enable us to explore the user experience of the study and to record data about the numbers of users who may be prepared to pay for data (willingness to pay) and reasons why they chose to pay for premium data.

The data for this experiment was generated from a number of sources. The location of the car parks was taken from the Bournemouth Borough Council website.¹

This data was extracted manually from the website and collated into a .csv file and converted into RDF. The exact location of the car parks were then plotted onto Open-Street Maps and also OS Maps and the data relevant to each car park was added to each point depending upon the data type given. The free data did not contain many details about the car park, whereas the paid versions of the data contained more text and was displayed when a user clicked on the point. We note that at this stage the data available from the Bournemouth Borough Council website² was not in a LD format and

¹<http://www.bournemouth.gov.uk/PeopleLiving/Maps/CarParks/KingsParkCarParkMap.aspx#>

²<http://www.bournemouth.gov.uk/PeopleLiving/Maps/CarParks/KingsParkCarParkMap.aspx#>

required time to translate into data which is of a useful format for linking.

Monetary incentives were provided to encourage participation in the study. Participants were offered a base rate for participation and then a further incentive was offered for obtaining the best solution in the study, i.e. achieving the best data/result and having the highest bank balance at the end of the study.

In the design of this experiment we considered other willingness to pay for research efforts and have designed this user experiment, taking into account the findings from these experiments.

Although studies have been carried out regarding willingness to pay, there are no studies to date which have concerned the willingness to pay specifically for LD. We have also found that the most studies concerning willingness to pay relate to that of online auctions. We note that the factors which affect the consumers decision to pay or bid for items on sites such as eBay, have similar issues to the ones we have discovered for LD. Buyers are concerned with the trustworthiness of sellers and more specifically their feedback scores (Melnik and Alm, 2003). We take this into account as a potential way of ranking sellers of data online and investigate this in more detail in chapter 8, the technical architecture for LD.

Other choice experiments have been carried out in the past regarding physical goods and some for online goods. We note in particular a study carried out by Chyi (2005) which used interviews to find consumers willingness to pay for news online. We note that there are no similar studies for LD.

Although the use of LD in our application was effectively hidden from the participants, we believe that our choice experiment is the first such to use LD as the subject of a choice. As mentioned at the start of this chapter, LD is not intended for direct human consumption but the participants are nonetheless choosing between different sources of LD.

We note that there have been a number of experiments carried out which use false or pretend money and decided to use a pretend currency, in this case ‘map groats’ as opposed to pounds and pence, so that we did not bias the decision of participants by introducing the problem of price into our investigation. However, we suggest that a further development to this investigation should be carried out which investigates the different real prices consumers are willing to pay for online goods and in particular, data.

Every selection made during the study was emailed back to the researcher on completion of the study, enabling analysis to be conducted at the end of the study. The simulation enabled participants to interact with LD and to carry out a simple task which highlighted some of the key points of LD.

7.1.2 Participant Interaction with the Study

We have included screens shots of the experiment in Appendix B but we will explain how the participants interacted in the study in this section.

Firstly the participants were asked to enter their email address and click a send button which sent a personalised link to their chosen email address to ensure that we were able to trace the answers they gave.

The next page they viewed gave them the instructions for the study and then once they had read this page they were directed to the first page of the study.

They were shown a snapshot of a map of Bournemouth on the right hand side of the page and on the left was the scenario. They were asked to find the most suitable place to park in order to go shopping for a birthday present. They were given three criteria to take into account in their search for the correct space.

Following this they were asked if they would like to use free, paid for or premium data for their search for a suitable parking location in Bournemouth. They were able to see the data from all three and we recored which datasets they used to get their answer. Once they had made their choice and clicked next they were taken to a questionnaire about their experience with the scenario.

7.1.3 Participants

A wide sample of participants (displaying diversity in age, gender, occupation and location) were recruited to take part in the study. A number of participants had already shown an interest in taking part in research studies through Ordnance Survey and were contacted with details of this study.

Participants were also recruited from universities. This was advertised by email on behalf of the researcher by the participating universities.

As we stated, we had a wide sample of participants. In total 50 people took part in the study. However we noticed that the older participants in the study were less inclined to pay for the information available. We also noticed that some participants spent a minimal time completing the study. We acknowledge that this could introduce a bias in our results.

7.1.4 Results

The results from this experiment enabled us to begin to understand why or why not consumers would be willing to pay online for [LD](#).

Table 7.1.4 below outlines the number of users in each category who selected free, paid or premium.

On average we found that most participants spent 15 minutes searching for the optimal answer and then having understood what was required of them were able to answer the questionnaire following the study.

Data type	Number of Participants
Free	18
Paid	22
Premium	10

TABLE 7.1: Proportion of participants in each category

Table 7.1.4 illustrates the number of participants in each category who chose the correct answer which was location h. This was determined by the number of free spaces available. This could be found by looking at the additional information which was available through the paid or premium versions of the data. The optimal answer was not the closet geographically to the desired location but would fulfil the requirements of the study which was to find a space quickly at a peak time of day and by making a purchase they would have been able to make a decision with less effort.

Data type	Participants with the correct answer
Free	0
Paid	2
Premium	2

TABLE 7.2: Proportion of participants who chose each option

Participants were asked in each case to explain why they chose the answer they did. There were mixed responses to these questions.

Of the 18 participants who chose the free option, 9 participants said they were not prepared to pay for data and 7 participants said they would rather look for the information themselves.

14 participants said they would use the same data option in the future whereas 4 participants indicated that they may choose a different option in the future. This may be due to the fact that they had learnt that choosing a particular option may not result in the correct answer.

1 participant said that if they were using an application they would like to have seen more detail. Another user said they would choose a paid option on repeating the task, but, would not choose a premium option as they felt it was too costly for the task in question.

Of the 22 participants who chose to use paid data, 9 participants said they were prepared to pay a small amount for the data and 5 did not think that the free option would be

adequate.

9 participants said that they would be happy to use paid data again whereas 5 participants said they would make a different choice.

1 participant stated that they would not find more information in the premium option useful and 2 further participants stated that they would not be prepared to pay any more.

1 participant said that they would not want to put their financial details online and so would not use the paid option again. Another user said that they would like to check the free option first and then decide whether to use a paid option in the future.

Of the 10 participants who chose to use the premium option, 3 said that they chose this option as they would not be prepared to spend time searching for the answer. 6 participants said they would definitely pay for the premium option if they could guarantee it would give them the right answer.

1 participant stated that they would pay for the data only if they had little time and it was important.

7 participants stated that they would use the premium option again and 2 would choose a different one. 1 participant stated that they would like to see the other options before making a purchase.

7.1.5 Discussion

From the results (see Figure 7.1.5) we see that a high proportion, 64% of participants are in fact willing to pay for data. The remaining 44% of participants were not prepared to pay and 28% of participants who chose free data would use the same choice again. Those who are prepared to pay for the premium data do so because of convenience and time saving whereas the free participants are prepared to spend the time searching for data rather than spend extra for a service.

As a pilot study we only sampled a small number of participants. We feel that the situation participants were asked to simulate may have led them to make the decisions they did. Participants may have felt that if they chose the paid option they would get the ‘optimum’ result. Whereas some participants may have felt that by choosing the premium option they would get the best result. However, this may not have been a true reflection of their buying behaviour in everyday situations. We feel that a further study which demonstrates a number of different scenarios may be useful to determine whether the scenario has any effect on a user’s decision to pay from a premium service.

The currency used for the study did not reflect real money. We feel that this may have influenced the participants when they considered whether or not to pay for the data.

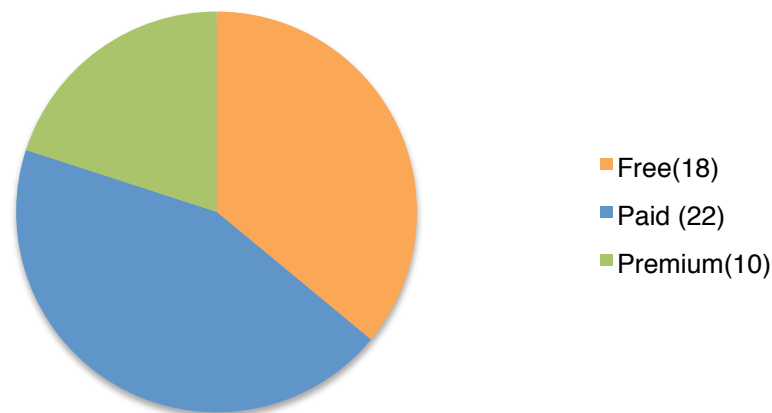


FIGURE 7.1: Data type selection.

Some participants may have not been able to relate the currency we chose with real money so may have spent it feeling that they were not making a real loss by spending. Or, others may not have spent as they may have felt that the currency was more valuable than it really was. Therefore a further study to consider using real currency and varied prices may be useful to help determine how price of the data affects a participant's decision to pay.

Having completed the study 4 participants went on to repeat it again. They chose the best option they found from using the premium data and selected free data the second time. We feel that this was for two reasons, the first as they wanted to achieve the best result and secondly they were not allowed to preview the data in the first place.

A number of participants suggested that in the future they would like to preview the free option before they decided whether they would make a purchase. For the purpose of this experiment we decided to make sure they were not able to preview the other options as this may have resulted in them still choosing a free option. Therefore a further study could look at offering a user a preview of all the options free, paid and premium and recording which one most participants chose.

One user stated that free data can be easily combined with other sources, these can include local people who would know the best car park. A paid for car park map may not include all relevant, up-to-date factors such as crime, car washing or accept credit card payments. We find this particularly interesting when looking at [GI](#). User generated maps such as Open Street Map have local knowledge attached to them and are often a key factor when people decide to use the map over an alternative. Therefore when looking at creating a linked [GI](#) application it is important to note the criteria which people consider when they choose to purchase data or an application. We consider this an important factor in the development of LD applications and will investigate the information quality factors in our further research.

We note here that the ‘free’ data was not considered of less quality, the premium data was considered ‘premium’ as it gave the same answer but with fewer steps involved. We aimed to demonstrate that the free data was acceptable to use, although may not have been able to answer all the criteria, but by purchasing premium data the optimum result for the situation could be achieved perhaps within a shorter time.

Only 4 participants chose the optimal answer given the data provided, which was ‘h’. We note that this was not the most obvious choice and most people chose the one which looked geographically closest, but only by purchasing the data would they find that this was not the most suitable place. The participants who had chosen the correct answer had chosen either a paid or a premium option. None of the participants who chose the free option got the correct answer as they were not given the relevant information only a scale and points with no additional details.

We aimed to test two hypotheses at the beginning of the experiment, the first;

H₁ Consumer willingness to pay for data is affected by information quality factors.

and the second;

H₂ Willingness to pay for data will affect consumers product choice.

We have proved through the results of this experiment that both hypotheses are true. There is a significant relation between consumer willingness to pay for data and different information quality factors. Secondly, that consumers purchasing behaviour will result in consumers choosing a cheaper alternative, even if the paid option will generate the required answer.

7.2 Linked Data Investigation 2

Following the questionnaire outlined in the previous chapter, we were able to inform a further study to test the top 3 most important criteria within a LD simulation. We wanted to confirm if in a LD situation, the factors outlined in the previous experiment would affect a participants decision to purchase the data in a LD situation. If data was incomplete, would participants then decide to make a purchase of data which was complete? Or, would they decide to purchase data without looking at the free options as they deemed this as time wasted.

The initial LD experiment in chapter 5 aimed to find out what people said they would do in a particular situation. This experiment aimed to test whether what people said in the previous experiment is true, factoring in the findings from the questionnaire which will make the experiment more substantial using criteria which has previously ranked.

We hypothesise for this experiment:

H₁ Context of a purchase situation affects a consumer's willingness to pay for information.

H₂ Willingness to pay for information is related to consumer requirement to reduce search times.

7.2.1 Experimental Design and Methodology

The same structure from the initial LD study was adapted to create an online interactive study (see Appendix E). This study was built around LD to demonstrate that it works and is implementable in a LD environment. It was again a server side application written in python, which was used to generate dynamic webpages. The script was executed on the server and so was not visible in the browser and all webpages were displayed in HTML. The data for each of the cameras was generated in RDF and displayed on the webpage using a SPARQL query which was executed to display different responses. We were able to customise our response based on the users input or response enabling different scenarios to be modelled. This was developed in order to give participants the chance to interact with different information. The information was tailored to reflect a real world situation where free reviews may have been biased, out of date, inaccurate and incomplete. Whereas paid reviews were accurate, complete, up-to date and concise. We wanted this experiment to illustrate how people can follow links which contain collaborations of data rather than typical web browsing.

The different quality reviews have been used to represent more complex reviews that could be composited with different qualities of LD. The free examples for instance were generated from sites using free data and could therefore be accessed via the web but being input in this application they displayed how they are able to be drawn together. The reviews which required payment, were reviews which would require payment on the web without necessarily being in LD. We gathered this information and translated it into RDF to structure it in a way which would enable us to show how data can be made more useful if it is in a structured format and may be something which consumers may wish to purchase.

The participants in this study would not necessarily be aware that they are using LD as such, however we have ensured that all of the data used in the experiment was in RDF and that the choices made by the participants were displayed by a SPARQL query, which again, they would not see but demonstrates how it is possible to create a LD application which can generate different sorts of applications.

Participants were invited to take part in the study at Ordnance Survey Headquarters in Southampton. They were asked to imagine they were looking to purchase a camera as a certain type of user (a point and shoot user, a keen amateur or a professional). We allocated them a user type and asked them to act with certain criteria in mind.

They were then asked if they would like to purchase a subscription to the paid data for a one off fee. If not, they were directed to the data where they were only given access to the free data. If they chose to they were able to purchase individual reviews.

They were asked to use the information provided to establish which camera they felt was the most suitable for their type of use.

After the interactive element of the study, users were asked to complete a questionnaire regarding their experience with the study (See Appendix F).

Following the study, participants were told of the purpose of the study and the reasons for the study were explained to them. We then gave them the correct answer and they were given their incentive payment.

7.2.2 Participants

30 Participants were recruited from a list of volunteers who signed up to take part in research projects through Ordnance Survey. Participants who signed up to this list have a keen interest in research at Ordnance Survey and an interest in maps especially. As the participants were mainly map enthusiasts, we chose to use a neutral topic of cameras rather than one specifically based on mapping to ensure that we did not introduce any bias in particular about brand loyalty. We also ensured that the data about the cameras in the study did not contain any brand names and we replaced these names with references to camera A, B, C, D etc.

The payment method we chose for participants was aimed to ensure that we gained the most natural response to the questions being asked. We advised participants that they would be given £5 to spend on data, they are not obliged to spend the money but if they get an incorrect answer they must return the unspent money.

In fact we paid the participants the full amount regardless of whether they achieved the correct answer. We felt that by asking them to perform the task in their own character meant that we would not find a bias in participants trying to earn the most incentive payment from the study.

7.2.3 Results of Linked Data Simulation

Of the 30 participants who took part, - 10 participants paid the subscription for the data from the start. - 3 participants chose to subscribe after viewing the free data - 13

participants never paid for the data and only used free data - 4 participants paid for individual data not under subscription - 11 participants achieved the optimal answer

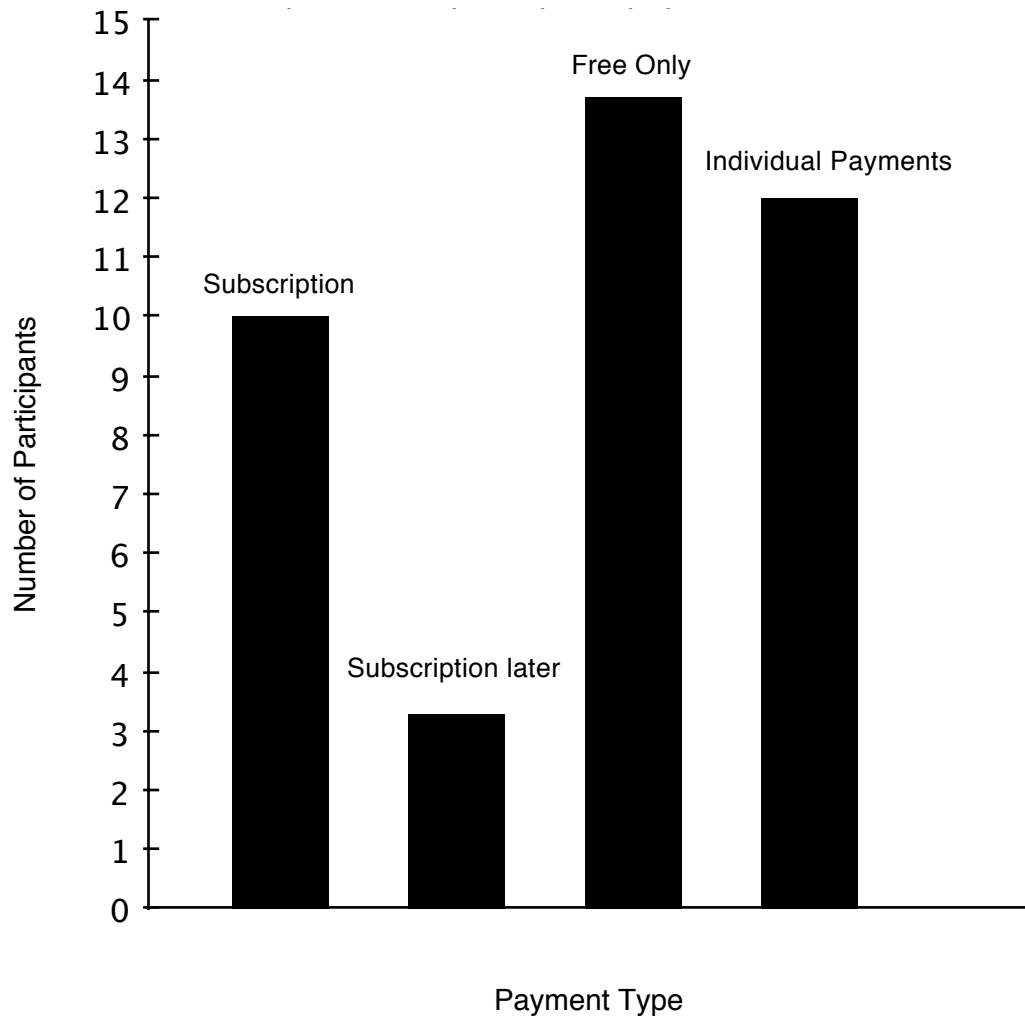


FIGURE 7.2: Graph to show proportion of users who chose each payment type

7.2.4 Results of the Post Simulation Questionnaire

28 users (98%) said that the type of hypothetical user they were allocated influenced whether they decided to pay for the data.

When asked which information quality criteria was most important to them when searching for information, 14 users said that accurate information was most important to them, 5 users said that complete information was most important to them and 11 users said that consistent information was most important to them.

When asked if they were to purchase information on the web, 12 users were concerned about credit card security, 14 users would prefer to look for the information themselves and 22 said they were concerned they could not guarantee the information would be useful.

10 users said they would prefer to spend a small fee for obtaining the answer instantly and 20 users would prefer to spend time looking.

When asked how much would they trust the checkout/payment process online. 7 users said they would trust a lot, 6 users said they would trust it a great deal, 11 users said they would trust it a moderate amount, 6 users said a little and 0 said none at all.

7.2.5 Participant discussion

Following the interactive part of the study and the questionnaire, users were asked for their comments on the study. Each participant was given the time to discuss the study and their views on paying for data online.

These discussions are summarised below:

Participant 1

Would never use a paid review as it may be biased, regardless of the situation.
Tends to find that the paid reviews do not give you all the information.

Participant 2

Would not usually pay for information but if they could not find the right answer they would pay for the data in the end.

Participant 3

No knowledge of cameras therefore obvious choice is subscription as £3 is reasonable for lots of information. For an expensive product like a television or camera, would spend a small amount for a trusted review as they felt that time is wasted searching for information. They felt that free reviews were no accurate or correct and they tend to trust reviews for which they pay.

Participant 4

How do you know the credentials of a paid for review? Would want to know the reasons why they should trust a paid review. For example would like a preview of the information/review before they purchase *but* a free trial which doesn't allow all the features would not be suitable. Depending on the type of user, depends on how fast you want the answer, i.e. a professional may need the answer faster for a deadline.

Participant 5

If searching for this information would spend longer looking and would be prepared to search rather than pay. Would however pay if a professional but not as an amateur.

Participant 6

Participant stated that there were not actually many 'paid' review sites around and therefore said the context may influence the decision more.

Participant 7

Participant specifically trusts the manufacturers more than other users as people tend to be biased and may review a product on personal preferences rather than on how the product really performs.

Participant 8

Stated that online shopping does not give the same experience as you get in the shop, therefore cannot get a true reflection of a product.

Participant 9

Context is very important on decision to pay

Participant 10

In the current economy participant felt that people may not be so prepared to pay for information as they see they can get the same answer by spending time searching.

Participant 11

If paying for information online would expect value for money, clear, accurate up-to date and trusted. Concerned about conflicting reviews, how do you know which one to trust?

Participant 12

Decided to pay for data as not convinced that the free data was correct.

Participant 13 The participant stated that regardless of the cost or accuracy of the reviews available online they would always spend their time searching through all free reviews available online to find the answer to the problem. They stated this would be their reaction in any situation or context. They also stated that they

would not be satisfied that all of the information would be available in a paid review and therefore would be unwilling to pay.

Participant 14 The main concern from this participant was providing their bank details online. They would like to pay for reviews, but would not pay for them as they would not want to disclose their bank details online.

Participant 15 The participant stated that they paid for the data as they were not convinced that the free data would be accurate, or trustworthy.

Participant 16 They viewed all of the data as they said that if they were going to pay for data they would want to get value for money.

Participant 17 The participant stated that they would rather spend time searching for data but if they did pay for the data they would like to make sure that it was accurate.

Participant 18 Swayed particularly by the opinions, therefore finds opinion useful and would be prepared to pay for this if they felt it was required. They also stated that context is important in deciding whether to pay for data.

Participant 19 Participant stated that their prior knowledge of cameras influenced their decision to pay, i.e. if they had little knowledge of the subject they would be more prepared to pay.

Participant 20 Participant showed concern for biased reviews and stated specifically that trust on the web was an issue for them.

Participant 21 Would like the option to make small ‘micro-payments’ for data but knows that their partner would never pay for information online, whatever the situation.

7.2.6 Discussion and Conclusion

The aim of this experiment was to answer a number of questions regarding the use of paid data online. By asking participants to do this, we wanted to find if the answers they gave in the first [LD](#) experiment detailed in chapter 5 were true responses. One of the critical issues we wanted to address is the reasons why they chose to make a payment for data.

We note that although this may not be a typical use for [LD](#) and camera reviews do not contain [GI](#) specifically. It was difficult to make strong inferences that the findings relate specifically to [LD](#) but we are able to demonstrate the capabilities of such a [LD](#) application and how it can be tailored for different needs.

The literature which we outlined in chapter 4 suggests that there are numerous factors which affect a consumers willingness to pay. We decided following our review of the literature, that we would consider the factors of quality. Stemming from the area of quality we were able to outline a number of other factors such as timeliness and accuracy. We tested each of the criteria with real participants and the interviews following the experiment enabled us to generate more details and accurate responses to our research question.

We developed two hypotheses for this experiment:

H₁ Context of a purchase situation affect a consumers willingness to pay for information.
and

H₂ Willingness to pay for information is related to consumer requirement to reduce search times.

We have found that both are significantly true. The results of the experiment show that there is a statistically significant relationship between the context and a decision to pay for information. We can see from the results that if the participant was asked to interact with the information as a professional user, then they were more likely to pay for the data as they felt that the product was expensive and they would want to make sure they had all the information available to them. However, they said that when they chose to pay, this was not necessarily the same choice they would make if they were a different user. For example if they were a point and shoot user, they would not necessarily have paid. Therefore, we note that there are different circumstances where people will choose to pay for information and this has a direct influence on willingness to pay for information online.

The second hypothesis is also true but only in some situations. We have found from this experiment that some consumers are willing to pay if they are limited by time, however some will never pay for data even if they are pushed for time.

We also noted that over two thirds of users would prefer to spend time looking for information rather than pay to obtain the information faster. This may be due to the issue of trust. People are cautious when searching for information online and when asked to make a payment for information they want to know that what they are paying for is worthwhile. This highlights an issue which needs further investigation to encourage users to interact more comfortably with linked information online.

We note that two thirds of participants who took part in the study were of retirement age. We suggest that there may be slight bias in their response, potentially as they may have more time to spend searching for data and are therefore less inclined to pay.

We have found that there is a clear divide between participants' responses. There are people who will always pay for information in any situation, there are people who will

never pay for information online (for a number of different reasons) and those who will pay depending on the situation/context. This illustrates different behaviours of users and suggests that until there is a satisfactory means of searching for information online people will always search for the information themselves. We find that this may be due to information having always been free on the web and similar to the introduction of charges for newspapers online. People are not willing to pay when there are free alternatives available. The issue here is how can organisations continue to operate online when free alternatives are available?

In terms of information quality, people mainly require accurate information followed by consistency. They are least concerned with completeness. When considering resources such as [OSM](#) we note that [OSM](#) is not a 'complete' resource as such and yet people still choose to use it. Perhaps this is due to consistency of the maps and on most occasions accuracy. If found not to be accurate users are able to contact [OSM](#) to make the relevant changes. The benefit of a user generated system allows people to contribute their local knowledge.

Trust was highlighted regularly in discussions with participants. The key issues highlighted were, if you purchase information online, how do you know that the author is trustworthy? For example, what is preventing them from publishing incorrect links to data and who verifies that the links are correct? How do you know that what you are going to purchase is what you require if you cannot see it before you buy? They also highlighted the issue of cost. For example how do you know that the cost of one piece of information for 50 pence, has the same quality of information of that which is £5? In a shop on the high-street you are able to hold and examine the product, and return it if it is not what you want. This is where brand recognition plays an important part on the web. We are more likely to make a purchase from a brand that we know has a reputation for good quality products than one which is unknown. But the key question is how do we model this for individual information providers who do not have 'brand recognition' status as such. We will investigate this further in the next section of this research.

We have made an assumption here that, although we know the meaning of the attributes listed, the participants may not have fully interpreted the meaning in the same way. This means the terms have an unstable meaning and therefore we suggest that the study could be repeated using a new set of participants to determine if they interpret the attributes in the same way. We could also use a control study to find out what the users understood by each term.

This leaves unanswered questions for information sellers on the web of - How much do you give away for free and how much do you charge for a paid version?

From the discussions with the participants following their interaction with the study we found that there were key criteria which also affected their decision to pay for information

online.

- Trust
- Accuracy
- Value for money (Range of opinions, quality and quantity)
- Bias
- Security
- Online payments

From the experiments described in this chapter, we have been able to establish a number of factors which we use to influence the user driven architecture for linked geospatial data which we detail in the next chapter.

Chapter 8

An Ordnance Survey Case Study Using Linked Geospatial Data

8.1 Introduction

The previous chapters have covered the literature surrounding the technical aspect of this research which included [LD](#) and the [SW](#). Following this we introduced the key theme for this research which was surrounding [PSI](#) and more specifically [GI](#). We then introduced the key economic and business issues for research which provided us with the foundation to the studies which we carried out in the remaining chapters of the thesis. The results from the empirical research we have carried out have enabled us to inform what we refer to as a technical framework for the consumption of [LD](#). By this we mean an informing policy which initially establishes how [LOD](#) and [LCD](#) may be consumed alongside each other and further, the possibility of granting access to data which may be closed via restrictive licenses or paywalls. The [Cobden et al. \(2011\)](#) paper which we co-authored and presented at the COLD (Consuming Open Linked Data) workshop at ISWC (International Semantic Web Conference) 2011, outlined a possible framework for [LD](#). However this left areas which were not addressed regarding the detail of the framework. We explore these areas in more detail and discuss here how a user interacts with the system. We also specify the architecture for each of the components of the system and how they may interact. This chapter is an examination of how [OS](#) is moving towards producing open data from a commercial perspective. In the absence of concrete information about the definition of pricing models, we use [OS](#) Open Space and [OS](#) Open Space Pro to illustrate a free and premium situation.

When we explore [LD](#) we notice that there are areas which are missing when incorporating a system for [LD](#). These include value of the data to the holding organisation and to the consumer and the willingness to pay from the consumer. Currently, systems for [LD](#) do not consider a link between a free dataset and a paid dataset. That is to say there are

direct methods to use free data and these are well utilised, searches can be carried out via SPARQL endpoints and data can be retrieved but where a situation arises where there is a paid dataset available which is an extension of a free one, there is no way of linking it to show it is related to the original. The free data is available shows no link to data which may require payment. The addressing products at OS are a premium product which are expensive to maintain and are seen by OS as core to its growth. There is a general belief that the more value can be realised by using the address data to act as a directory to 3rd party data, of which the value is unknown and is an ongoing area of research which is being investigated in the SPRITE project at OS. With a dataset which holds high potential value to its consumers and is a valuable product for the holding organisation, ways in which utilisation of this value can be exploited are an imperative area for further exploration. We take into account these shortcomings in the work to date and suggest the requirements for a potential system.

8.2 Ordnance Survey Case Study

8.2.1 Background

In the previous section we detailed the technology required for a LD architecture based on our findings from the literature and from our own investigations. We then use this section to describe how it can be applied around using addresses as a case study.

As we have noted earlier in the literature of this thesis, spatial data has a significance for LD. Hart and Dolbear (2006) highlights the importance of spatial data, stating that eighty percent of all data has a geographical component, meaning that many datasets contain data which has a direct or indirect link to a physical location. In order for us to illustrate the use of LOD with LCD we will outline a key focal point that we can use to link data to, this is an address.

We have chosen addressing as although spatial data does play an important part in many datasets as indicated above in very many cases this data does not contain explicit coordinate references but identifies location through the means of an address. In the UK the majority of these addresses are postal addresses as defined by the Royal Mail, although increasingly within local government administrative addresses obtained from local authority Local Land and Property Gazetteers (LLPG) are also used. Furthermore we note that addresses are useful in that they are typically used to directly identify and locate property and business, and indirectly (i.e. in conjunction with other data) to identify and locate people. Addressing therefore has significant economic value.

OS has a range of products called AddressBase® which contain Royal Mail Post Code Address File (PAF) addresses, both commercial and residential matched to the local authority Unique Property Reference Numbers (Unique Property Reference Number

(UPRN)). We have chosen the AddressBase products as these products allow the identification of property and features and the older address products from OS are being phased out and therefore this product range is the most suitable to base our architecture on.

AddressBase is available from OS in .csv and GML formats. It contains Royal Mail PAF addresses, both commercial and residential matched to the local authority Unique Property Reference Number (UPRN) and also addresses included by OS identifying a number of non-postally addressable objects such as structures and certain natural features such as ponds. There are three levels of Addressbase available. The first level, known simply as AddressBase contains just the postal address, the second level contains the postal address, the OS address and the local authority address and the third level contains all of these and some additional attribute contain information such as alternative addresses.

We note that the AddressBase range is a commercial product of value and is not available free of charge. It is also only provided through bulk data supply and there is service provision. By contrast the Royal Mail provide PAF, a commercial product containing postal addresses and very granular coordinates as both a data offering and also as a service. The service enables 15 lookups per day free of charge. The service thus provides very limited free use in a manner that does not threaten its commercial venture but provides the general population with a means to look-up the odd address. This shows a form of Freemium model which we outlined earlier in this thesis.

The AddressBase products support addresses that are compliant to both the PAF address format and also to British Standard BS7666 Spatial Data-Sets for Geographic Referencing. The later standard is used by Local Government to construct LLPGs and also in aggregated form to construct the [National Land and Property Gazetteer \(NLPG\)](#) that forms part of AddressBase Plus and Premium. We also note that whereas PAF simply contains a list of addresses, the LLPGs and NLPG contain references to property identified by [UPRN](#) that have addresses associated with them.

For the purposes of this exercise we will use AddressBase Premium as it is the richest of the three products. Most importantly this product has more records than AddressBase as it includes objects without postal addresses, such as subdivided properties, places of worship and community centres, and richer attribution than AddressBase Premium as it also has alternative addresses.

The current licensing for AddressBase is outlined in the licensing section in Chapter 5 under the Discover, Evaluation and Developer license.

AddressBase® is not currently available as part of OS OpenData and therefore requires translation into [LD](#) format in order to process it in a [LD](#) application.

We have stated previously that addresses derived from [LLPG](#) are based on the BS7666 and that this standard covers not just addressing but also properties. We therefore now

outline the nature of BS7666 and the way that it describes properties and addresses. For properties, the standard is based on the concept of a land parcel unit known as a Basic Land and Property Unit **Basic Land and Property Unit (BLPU)**. A **BLPU** is defined in BS7666 part 2, as an area of land in uniform property rights or, in the absence of such ownership evidence or where required for administration purposes, inferred from physical features, occupation or use¹. Each **BLPU** has a unique reference number **UPRN**, a spatial reference (grid co-ordinate) and one or more **Land and Property Identifiers (LPI)**.

The standard identifies two types of **BLPU**: a Primary Addressable Object **Primary Addressable Object (PAO)** and Secondary Addressable Object **Secondary Addressable Object (SAO)**. A **PAO** typically references a property at building level and a **SAO** identifies a property within a building. **SAO** are therefore referenced to the corresponding **PAO**.

The **LPI** is the address of the **BLPU** in a standard format that uniquely identifies the **BLPU** in relation to a street as defined and held in the **National Street Gazetteer (NSG)**. The principal components of the **LPI** are the **UPRN** from the **BLPU**, the **Unique Street Reference Number (USRN)** from the National Street Gazetteer **NSG** and sufficient elements from the hierarchy of PAO Name **Primary Addressable Object Name (PAON)** and SAO Name **Secondary Addressable Object Name (SAON)** necessary to uniquely identify the **BLPU**. We notice that **BLPUs** are therefore equivalent to features and **UPRNs** equivalent to **TOIDs** within the **OS OS MasterMap** product range. A **BLPU** can have one or more **Unique Property Identifier (UPI)**s as a **UPI** is effectively an address and a **BLPU** can have multiple addresses.

Addresses under BS7666 may or may not be identical to postal addresses.

For example the postal address for a location is

```
5 Picture Close,  
Warsash,  
SOUTHAMPTON  
SO31 9AJ
```

but the BS7666 address is

```
5 Picture Close  
Warsash,  
Hampshire  
SO31 9AJ
```

¹<http://www.iahub.net/docs/1183553456634.pdf>

We can see from this example that in the postal address Warsash is associated with the Post Town Southampton (Royal Mail convention is to capitalise the Post Town component). We also note that Warsash falls neither geographically nor administratively within Southampton. However for postal purpose mail directed to Warsash is first routed to the main area sorting office in Southampton. The postal address is therefore very functional. By comparison the BS7666 or administrative address relates Warsash to the largest administrative area in which it falls, the county of Hampshire.

We note that AddressBase Premium therefore offers an immediate benefit in that it cross-references the postal and administrative addresses to the associated [BLPU](#).

Even within BS7666 a [BLPU](#) may have multiple administrative addresses as we demonstrate below. The Address for Ordnance Survey is:

SAON	Ordnance Survey
PAON	Explorer House
Street	Adanac Park
Locality	Nursling
Town	Southampton
Administrative Area	
Postcode	SO16 0AS

Note that the admin area is not Hampshire as administratively Southampton is not in Hampshire!

An alternative address for Ordnance Survey is:

SAON	Ordnance Survey
PAON	4
Street	Adanac Park
Locality	Nursling
Town	Southampton
Administrative Area	
Postcode	SO16 0AS

We note that due to a lack of clarity within the standard the following is also a valid address form:

SAON	Ordnance Survey
PAON	4
Street	Adanac Park
Locality	Nursling
Town	Southampton
Administrative Area	Southampton
Postcode	SO16 0AS

Here Southampton is referenced as both a town/city and administrative area. We note that whilst this address form is somewhat clumsy it has nevertheless been used by some local authorities adding further to the complexity of an already complex data form.

We can express the relationship between the address elements and to the [BLPU](#) using [LD](#) but to do so need to assign [URIs](#) to the various components. The [URI](#) of the [BLPU](#) can be based on the [UPRN](#). However there is no equivalent unique id for the address component and so we will have to create one. We can base the URI for Streets on the [USRN](#) defined by part 1 of BS7666, Locality and Town from the URIs contained with the OS 50K Gazetteer and the URI for the Admin area can be obtained from the [OS Boundaryline Gazetteer](#).

We can describe the BS7666 structure ontologically using OWL as shown below using the Manchester Syntax:

```
Prefix: : <http://www.semanticweb.org/ontologies/2013/addressing-ontology>
Prefix: dc: <http://purl.org/dc/elements/1.1/>
Prefix: owl: <http://www.w3.org/2002/07/owl#>
Prefix: rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
Prefix: xml: <http://www.w3.org/XML/1998/namespace>
Prefix: xsd: <http://www.w3.org/2001/XMLSchema#>
Prefix: rdfs: <http://www.w3.org/2000/01/rdf-schema#>
```

```
Ontology: <http://www.semanticweb.org/ontologies/2013/addressing-ontology>
```

```
ObjectProperty: isOn
  SubPropertyOf:
    owl:topObjectProperty
```

```
ObjectProperty: owl:topObjectProperty
```

```
ObjectProperty: isIn
```

```
ObjectProperty: isAssigned
```

```
ObjectProperty: identifies
```

```
ObjectProperty: isPartOf
```

```
Class: PrimaryAddressableObject
```

```
SubClassOf:
  isOn min 1 Street,
```

`BasicLandAndPropertUnit``Class: CityOrTown``SubClassOf:``isIn exactly 1 AdministrativeArea``Class: SecondaryAddressableObject``SubClassOf:``isPartOf exactly 1 PrimaryAddressableObject,
BasicLandAndPropertUnit``Class: AdministrativeArea``Class: Street``SubClassOf:``isIn max 1 Locality,
isIn exactly 1 CityOrTown``Class: Locality``SubClassOf:``isIn exactly 1 CityOrTown``Class: Postcode``SubClassOf:``identifies some BasicLandAndPropertUnit``Class: BasicLandAndPropertUnit``SubClassOf:``isAssigned exactly 1 Postcode`

The ontology sets out certain requirements for the address. We have a POA which must be located on exactly one street and that the SOA is part of exactly 1 POA. A SOA is usually a business or flat within a property (POA) and therefore enables us to illustrate where there are more than one businesses located within one building or who share the same address. The locality shows areas within urban areas and is an optional feature and may not be contained within every address.

We illustrate this example using Figure 8.1. Although we do not need to explicitly state each of the requirements (i.e. every address has a POA) we suggest that it makes it clearer and more accessible to easily understand the address. The dashed lines between each part of the address illustrates whether each element is required or optional and also allows the address to be easily put back together in a [LD](#) format.

Firstly we will talk through the relationships between the data. Then we will go onto explain how freemium models based around addressing can be made to work according to the suggested [LD](#) architecture giving examples for free and premium data.

- A [BLPU](#) has a Post Code (URI based on UPRN)
- A SAO is a [BLPU](#) and is also part of a PAO
- A PAO is a [BLPU](#)
- A PAO is on a Street (URI based on USRN)
- A street is in a locality (not always required) and is also in a town URI taken from OS 50K Gazetteer
- A locality is in a town/city (URI taken from OS 50K Gazetteer)
- A town/city is in and administrative area (URI taken from BoundaryLine Open)

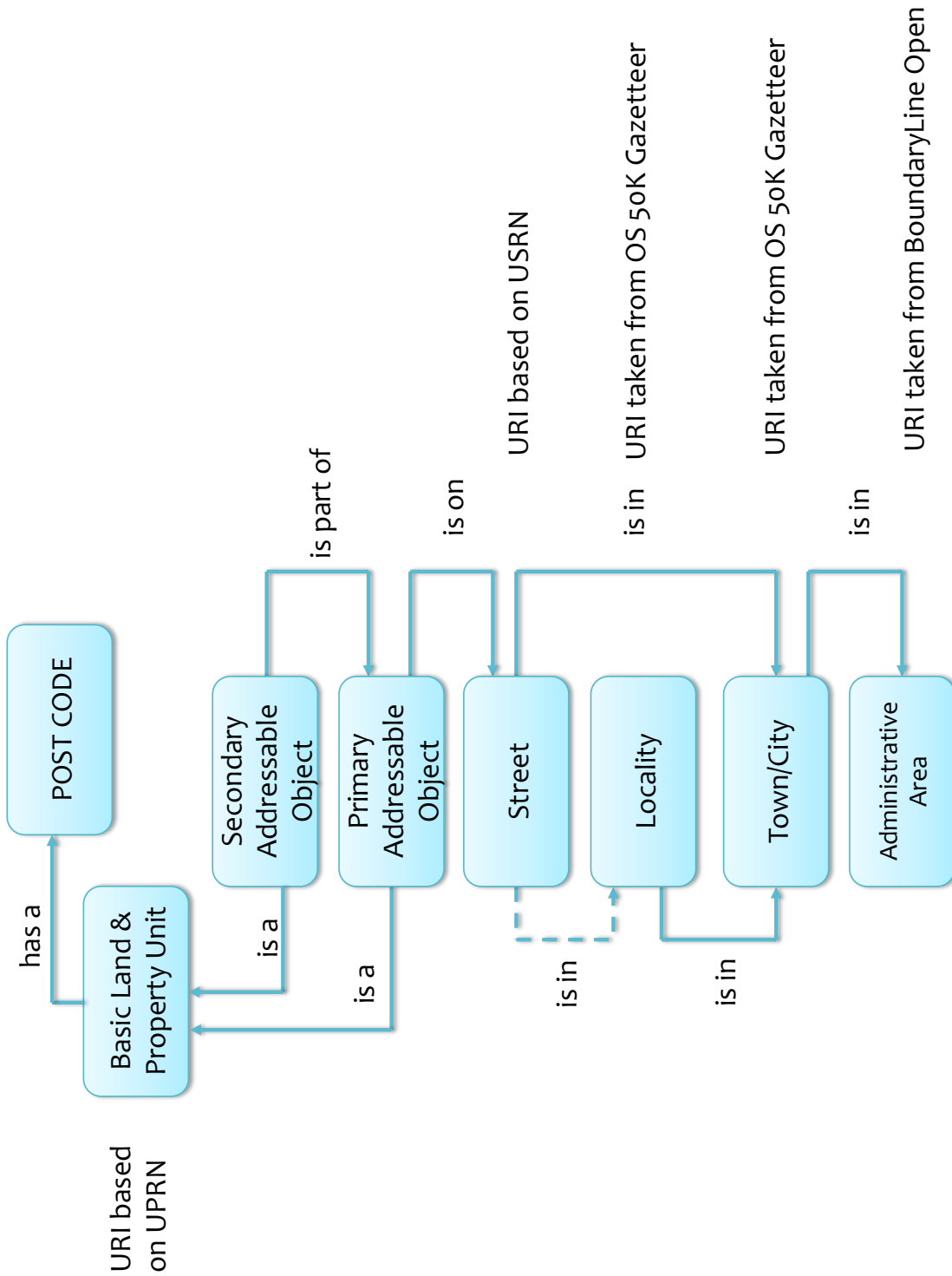


FIGURE 8.1: Illustration of the Addressing Example

8.3 Analysis and Requirements

As we stated earlier, the experiments carried out were designed to specifically identify the users requirements for the consumption of LD. Each experiment produced results which were used to contribute to the framework and that we discuss in more detail within this chapter.

We have identified that there will be scenarios of LCD alongside LOD. These situations may occur where there is a free version of data and a premium version of data. There may also be a situation where a ‘mashup’ is made up of data from various different sources, some of which may be closed and require payment. In order for us to manage this, a system must be in place to support each scenario, open data and closed data, or a combination of both.

The first investigations aimed at looking at the consumers of existing OS data and the proportions of leisure users and the types of purpose for which they used the data. From this we were able to establish there are different levels of leisure users. Following this we invited consumers of OS products to join in roundtable discussions with technically minded users to discuss their concerns with the introduction of a new technology. The results of this gave us the background for the framework to ensure we established a system which would be suitable for different types of users. Following the clarification of the types of users we decided to test how these differing types of users would react to different levels of data and if they would be prepared to pay. From this we were able to clarify that there is a definite proportion of users who may pay for data and therefore the framework we established needed to cover different levels of data. Further to this we extended the factors which would influence the decisions to pay and we began this investigation using a questionnaire and followed this up by creating a further study which queried if the decisions the participants said they would make in the first experiment were in fact true following a simulation of a LD situation. This study also clarified that the system would need to handle payments and access to data and manage the access through the use of licensing.

To summarise, the overarching requirements for the framework include:

- Search for data or navigate from other data
- Authentication and access control
- Pricing (Freemium data = Free + Premium)
- Licensing
- Payments processing

8.4 Actor Description

We have identified four key actors within the framework. These actors are detailed below:

The users

Role: Searches for data, enters login details, enters payment details, consumes data

The data provider(s)

Role: Provides data, publishes data in RDF on the web.

The data access control host

Role: Authenticates users for access to data

The payment handler

Role: Following authorisation from access control host authorises payments for data.

Figure 8.2 shows the actors we have identified within the framework and how they interact.

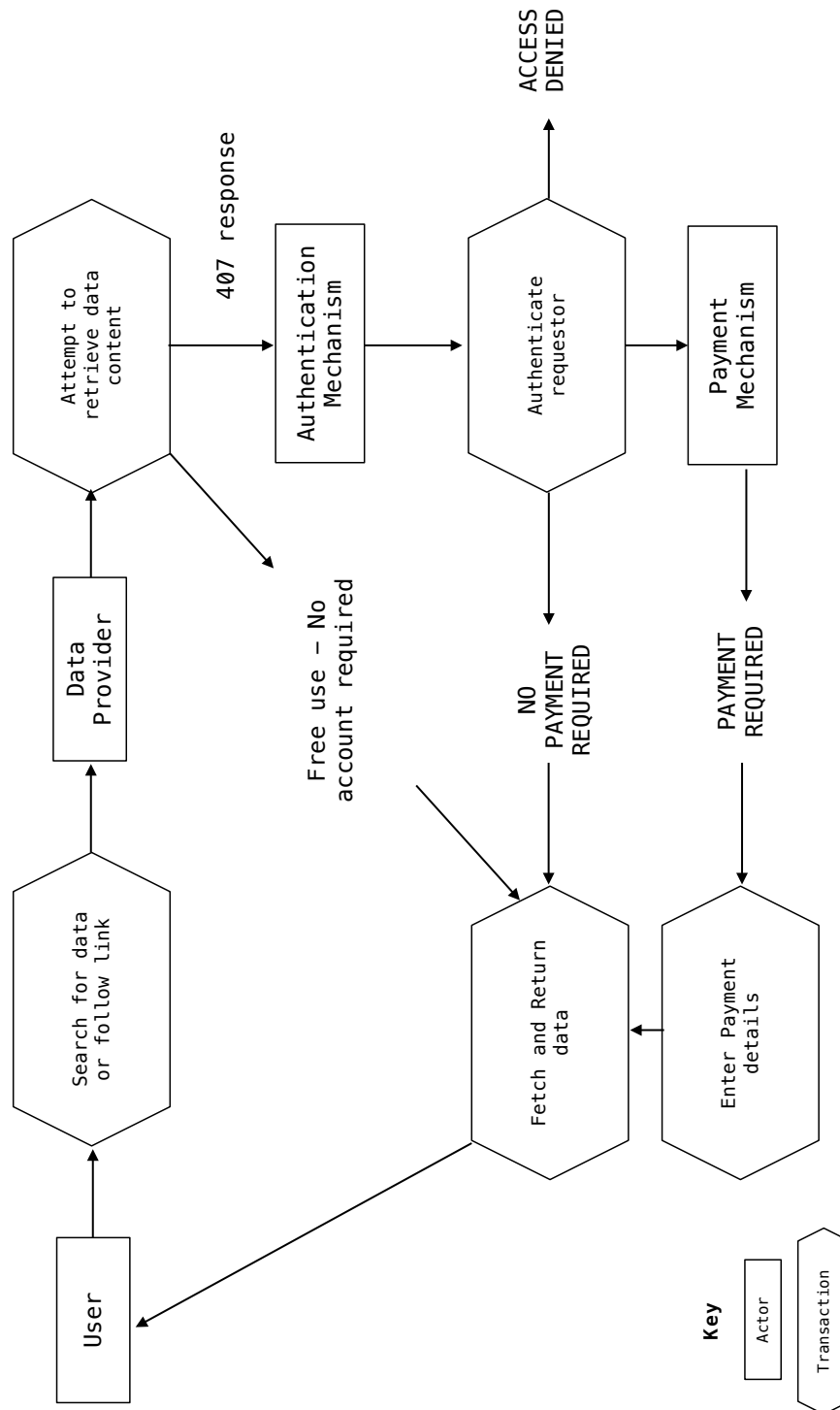


FIGURE 8.2: Actors directly involved in the framework and the transactions between them

8.4.1 Context - How does Addressing apply to our suggested architecture?

There are a number of possible scenarios that can be implemented into the architecture we have outlined. First we consider how we might implement a service based around OSs current pricing policy where addresses are charged and not free. Here OS could provide either a subscription model or pay-per-use. For the example we will work through we shall describe a subscription scenario. Here we assume a service user wishes to obtain the addresses related to a Post Code. The process is summarised in 8.2 and we will use this as the structure for the walk-through. In order to protect the data OS must set up their data-store such that any request for a URI referring to any address element other than the Post Code will return a 402 response Payment required.

This next section gives the step by step directions for how the user will interact with the data in the system which was illustrated in Figure 8.2.

1. A SPARQL query is executed by the requestor for the addresses required that relate to a specific Post Code such as AX13 1PQ.
2. The query is passed to OS who execute the query against their data.
3. Attempts to access the data content through their URIs will result in a 407 Authentication Required request being generated. This will be handled by a service such as OAuth and if the requestor is authorised to access the data through the subscription service then the data will be returned as requested. If the requestor is not authenticated the process terminates.

Figure 8.3 shows the actors we have identified within the framework and how they interact.

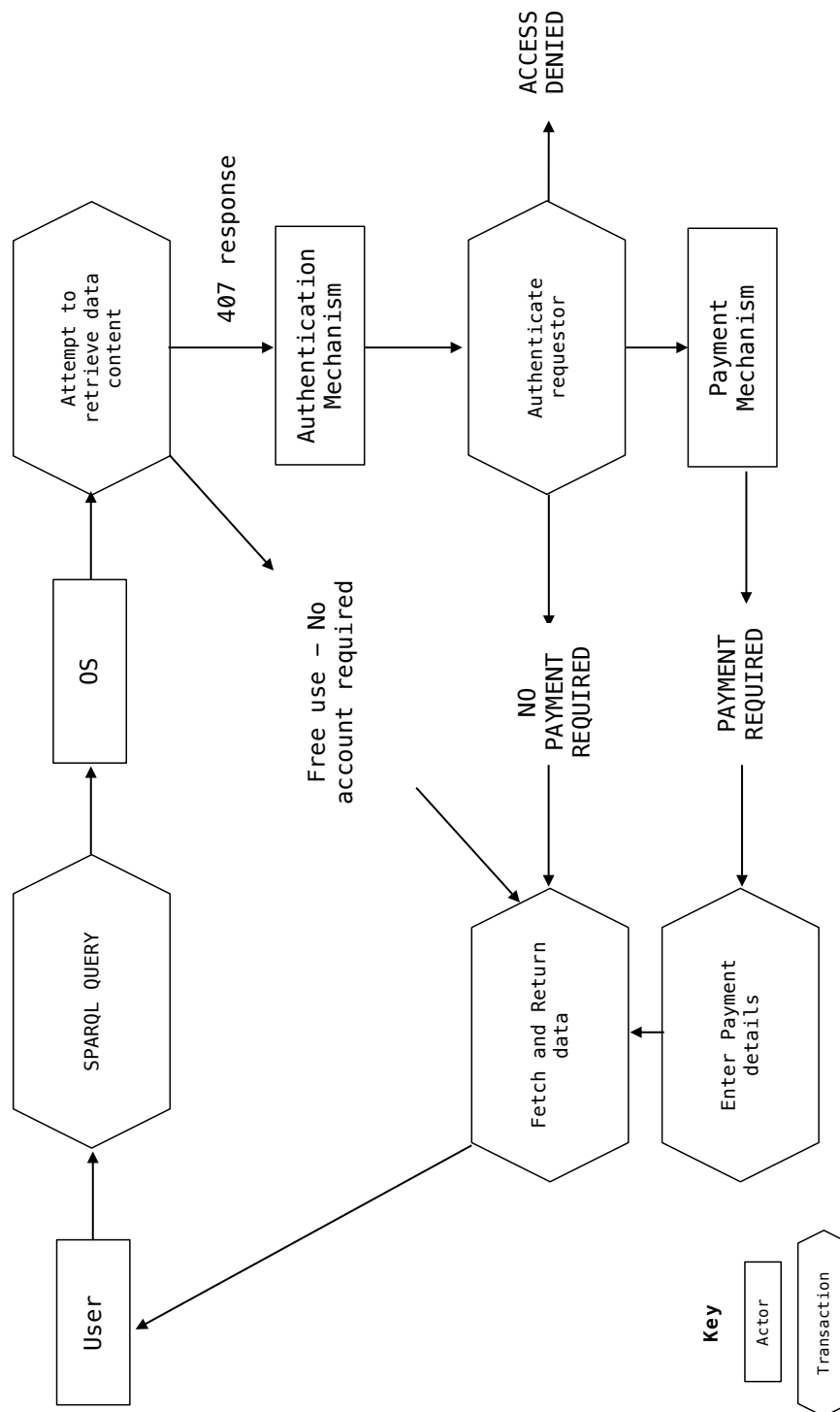


FIGURE 8.3: Stage one through the framework

Due to the nature of the system we are able to make very quick changes to the types of data which is paid for by changing the access rights on the links and not on the code/software.

So if the data provide decides to change the policy and they decide to give away some more data for free, all that is required of the organisation is the access rights to the links which direct them to the data, not the code in the software. Similarly we can see that it is simple to extend the data content by simply include the extra data (expressed as triples) and linked to the existing data. A conventional relational database would be required to alter the database schema to absorb new data as it does not support the simple and uniform triple model but rather requires data to be held within structured tables. Using this method it is therefore easy to see how it can be applied to support our second scenario, where some address components are free and others charged for.

We will suppose that OS and the Land Registry wish to operate a service to provide property information where some of the data will be free and some premium. Let us also suppose that they decide that information down to street level will be provide free for example average house price for a town, locality or street will be provided free but that specific details about individual properties require the user to be subscribed to the premium service. It is decided that OS will be the service provider. OS therefore take data from the Land registry (we assume here that Land Registry have their data as LD, if not it will need to be converted) and link it to their data: matching street data to street, locality to locality, town to town and property (BLPU) to property. If the Land Registry data remains hosted by the Land Registry then this matching would generate sameAs relationships otherwise we note that OS will simply add additional properties to their existing content.

In order to support this model OS change their pricing model so that they now do not charge for address elements and related properties below and including the Street. This is achieved merely by changing the access permissions to the appropriate URIs.

Now if a user requests the average property price relating to a street the SPARQL query will not initiate a process that results in a 407 Authentication Required response but will simply retrieve the requested data. However, we note that a 407 response will be generated if the user request data held by the Land Registry relating to a specific BLPU.

8.4.2 SPRITE

Sprite is OS prototype system exploring the concept of providing data and services based not just on OS data but also on third-party data. It use semantic web technologies to store and serve the data as LD. It currently demonstrates the principals through a demonstrator system that supports a service to enable people wishing to purchase a house to investigate the area of interest using OS and government Open Data including

data on schools, hospitals, social deprivation, amenities and house sales. All the data is either supplied or converted to LD before storing in a triple store.

The primary reason for adopting LD has been because of the ease in importing and storing new data enabling the system to grow in terms of information richness.

The Sprite example we give here does not use the Freemium model at present and in fact ignores service charging completely due to its focus being a technical demonstrator. However, we have noted that the potential for such a system to generate revenues through a charging model has been noticed by both the Products and Sales and Market Development Groups. We in turn note the ease with which a Sprite could be extended to demonstrate the freemium charging model and indeed following discussions that we have had with the Sprite Programme Lead this has now been added to the areas for further demonstration in the third incarnation of Sprite. (Sprite 1 was an initial demonstrator, Sprite 2 is currently being developed as a more sophisticated version holding a wider range of data and explicitly supporting address level searching).

8.4.3 Application - Why use Linked Data for this?

We note that it can be argued that the use of LD is not essential to the implementation of a freemium model in any of the examples given. We concede this observation but argue that the use of LD in this way offers two significant advantages over the use of conventional implementations.

Firstly we note that LD operates by placing access restrictions on the links between data. The architecture implementing the LD solution is itself entirely neutral to the nature and content of the data and no software is involved in setting restrictions to specific data elements. Therefore adding additional data that have may have very different content to existing data and providing the necessary charging merely involves adding the data, linking it to existing data and setting the chargeable restrictions on links either from existing data to the new data and, or within the new data. No code is required to be changed. The only assumption is that the data can be represented as LD which is very likely to be true. A more conventional approach is will not only require schema changes to its databases each time a new dataset is added but also code changes to handle it. Furthermore even changing the charging policy may involve code changes.

Secondly, we note that although it can be argued that conventional methods could be used to construct a data-driven system that would not require software to be altered given new data this would be done in a non-standard way, each implementation being unique. LD by comparison is based on international standards and a wealth of software already exists to support it. Implementation costs will therefore be significantly lower.

8.5 Technologies

Based on the literature we detailed in Chapter 2, we outline the technology and consider its suitability below.

8.5.1 Data Format

Many formats of data will be available on the web include .csv and .pdf. According to the five star ranking system outlined by the W3C² as long as data is available, even in a non-machine readable format, it achieves one star, but in order for it to achieve a higher ranking, (four stars), it needs to be in an open standard such as RDF to enable others to link to it. We suggest that in order to comply with this requirements of the W3C, RDF is the data format of choice for this purpose.

RDF, as detailed in the earlier sections of this thesis, is a language used for representing information about resources in the World Wide Web.³ RDF is a machine readable format which means when queries are executed the data can be processed by machines and processing by humans is not required. RDF is the language used as part of the architecture as it specifies the format of the triples being used and without which we would not be able to form the graphs containing the data. The W3C outlines the specification for RDF which is found on the W3C website.⁴

8.5.2 Query Language

In order for the data within a repository to be queried, it needs a specific query language. In this instance it is SPARQL. A simple example of a SPARQL query is shown below. The query is searching for the title of a book from the data. It is formed in two parts:

SELECT - identifies the variables to appear in the result.

WHERE - provides the basic graph pattern against which the data will be matched.

```
SELECT ?title
WHERE
{
  <http://example.org/book/book1> <http://purl.org/dc/elements/1.1/title> ?title .
}
```

²<http://www.w3.org/DesignIssues/LinkedData.html>

³<http://www.w3.org/TR/2004/REC-rdf-primer-20040210/#basicconcepts>

⁴<http://www.w3.org/TR/2004/REC-rdf-primer-20040210/>

The result of this query would be displayed as:

title
"SPARQL Tutorial"

The full specification for SPARQL is outlined by the W3C and full details of this specification are found from the W3C website.⁵

8.5.3 Authentication

If we are to consider a system which handles payments of an type, we must include an authentication mechanism to support this. We suggest that a possible solution to this would be the us of API keys or OAuth which we have detailed in chapter 2 in more detail. These mechanisms will enable small payments to be made and access granted to the data available.

The following section details two possible authentication models which could be used. There are other authentication systems available but the two outlined below give detail of how two differing systems work.

8.5.3.1 API Keys

In chapter 2 we introduce an [API](#) as an interface used to enable different software components to communicate with each other. In order to control access to these [APIs](#), organisations may wish to use something called an [API key](#) to approved users. An [API key](#) is basically a strong password with an account identifier (or name). ([Farrell, 2009](#)) outlines the use of [API keys](#) and suggests that the functionality and security of [API keys](#) is variable and suggests that the OAuth specification is a more suitable way to interact with protected data.

8.5.3.2 OpenID

OpenID, similar to OAuth aims to create a decentralised authentication system. This system enables users to consolidate their digital identities. Users create accounts with preferred identity providers which are then used to sign into other websites. The identifier is transferred into a unique URI which is sent to a provider which handles access.

There are a number of security issues surrounding OpenID which should be considered. These include areas where a user may be directed to a bogus authentication page. Similarly as a web page based system, the possibility of the webpage being intercepted by

⁵<http://www.w3.org/TR/rdf-sparql-query/>

an unauthorised person is a threat. We consider this to be a concern when dealing with payments however a number of companies such as Yahoo, AOL and Google incorporate the use of OpenID and demonstrate its security.

The difference between the two systems being that OpenID asks the users for their identity, whereas access to OAuth is requested directly from the application via the token system. For the instance of a LD system OAuth allow limited access tokens to be granted which will facilitate differing versions of data. For example, free, paid and premium.

We do note however there is discussion about the possibility of combining the features of OpenID and OAuth into a hybrid model. This would be particularly beneficial to LD as it would enable features from OAuth such as limited access to be incorporated into an OpenID system (Balfanz et al., 2009).

8.5.4 Licensing

As outlined in chapter 3, there are a number of licenses which are available for use with data publication. Specifically they comply with the principles detailed by the Open Knowledge Foundation.⁶ With many different datasets being linked together, restriction with one particular licence would not be suitable, therefore depending upon the type and nature of the data relevant licenses could be applied.

8.6 Protocols

The framework we describe below outlines the consumption of Linked Open and Closed data from search execution through to purchase, download and use. Figure 8.2 helps to illustrate the system.

A consumer will be required to enter search terms in the system via a linked data search engine which will execute a SPARQL query as exemplified earlier.

Following discovery of the required data, the user will click on the resource which will return one of a number of different status codes. These codes are set out by the W3C⁷ which could be used following a users request (Fielding et al., 1999).

- 303 See Other,
- 401 Unauthorised,
- 402 Payment Required,

⁶<http://opendefinition.org/licenses/>

⁷<http://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html>

- 403 Forbidden,
- 404 Not Found,
- 405 Method Not Allowed option to transfer to 402,
- 406 Not acceptable
- 407 Authentication Required.

Figure 8.4 shows the codes with the possible outcomes. Each of the codes can be used for different purposes. Table 8.1 outlines the use we outline for each code.

Status Code	Use
303 See Other	Free Data
401 Unauthorised	Restricted access
402 Payment Required	Premium Data
403 Forbidden	Restricted access
404 Not found	
405 Method Not Allowed	Option to choose premium
407 Authentication Required	For confidential resources, only be accessible via login

TABLE 8.1: Status Codes and Their Uses

As we outlined in the literature in chapter 2, we suggest that although there is the potential for large amounts of data to be available as ‘Open’ data, we have to take into account datasets which are not necessarily ‘paid for’ datasets, but for security reasons have restricted access. For example organisations who want to adopt LD but not necessarily publish it openly on the web.

The first option: 303 see other, the user is authorised to access the data as it is available free of charge. The user is sent directly to the data and the process is complete. The second option 402 payment required. The user will be transferred to the payment mechanism where payment is made and confirmation is sent back to the authorisation mechanism where access to the data is granted. This process is completed by returning a 200. The third case, 403 method not allowed. This enables restrictions to be made due to licensing restrictions or because the data provider has restricted access to the system. There is the option within this to provide an alternative such as a redirect to an authorised login page but this is an option which can be entered if required. The last case 401 authentication required enables data providers to allow users access to the data, but with the restriction that they must provide login details which enables the provider to trace them.

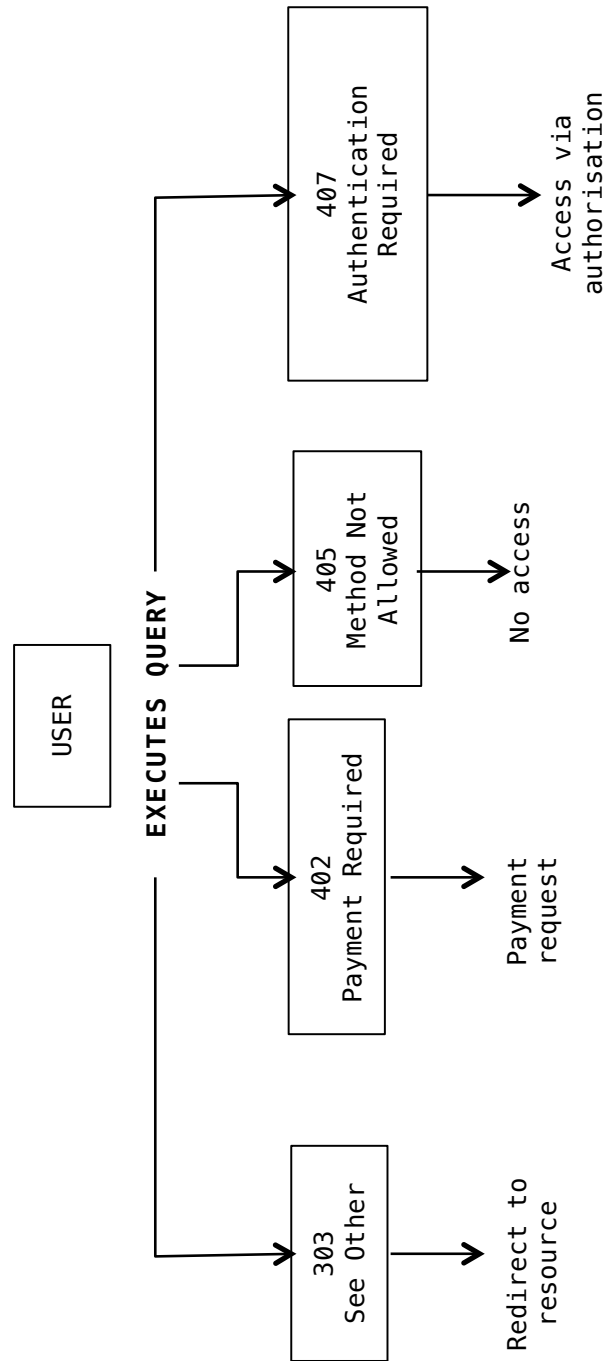


FIGURE 8.4: Possible outcomes from SPARQL Query

The subscription model is a model where users will be required to pay a subscription to a service in order to gain access to it. This will enable users to gain unlimited access to a system or a dataset or access specifically one dataset. This system is particularly suitable to a situation where users want to use or access a large quantity of data and thus makes a micropayment in this instance unsuitable. The subscription model also enables users to subscribe to a service where they can pay for additional features if they require (based on the freemium model). A subscription model will be suitable for handling a premium version of the data or to implement the premium version of a freemium model. However we need to ensure there is a way of handling other payments for smaller portions of data.

Micro payments for payments online are suitable for individual small payments for data to allow consumers to purchase as much or as little of the information as they require. From our earlier research, it shows that most people are unlikely to subscribe to a service and would prefer a service which they can opt in to purchases freely.⁸

There are a number of different micropayment systems available, for instance the W3C lists a number of different micro payment handlers for example Paypal, Clickshare and Cartio.

PaypalTM(or a similar technology) hides users details, thus enabling the user to share their details once rather than repeatedly for each purchase. This reflects concerns made by participants in the Linked Data study . The most re-occurring issue highlighted in the post study discussion was the concern with sharing payment details online. A number of participants stated that they may wish to purchase additional data but would not be prepared to enter their details online. A secondary supporting factor for using a Micropayment system similar to Paypal is that the system is fast. Paypal only requires a login ID (typically an email address) and password and therefore minimises the amount of data users need to enter into a system to make a purchase. For instance if this micropayment feature was incorporated into a mobile phone application, a user would discover a link to further data and then be transferred to a payment gateway if they choose to click through to make a purchase.

8.7 Other Considerations

8.7.1 Trust and Reputation

Following the investigations carried out in this research we note the need for users to trust the information or data which they wish to purchase.

⁸<http://www.w3.org/TR/WD-mptp-951122>

In the literature we investigated earlier, we considered the issue of brand awareness. Consumers are more willing to pay for data from a brand they recognise and remember as opposed to an unknown individual. We must consider ways of addressing this issue. Our first recommendation is to utilise the metadata surrounding the data and use this to provide a ranking system. This will include details surrounding the data provider and the person who made the link. We have noticed that it is not just the data itself which contains value but also the link (Bonatti et al., 2011). The link can be detrimental to the value of data if it links to something which is incorrect or inaccurate. The metadata could also be used to detail the number of links people have made and how many they have made which are inaccurate. The freemium model creates the best environment for the user to determine the quality and trustworthiness of a dataset as they are able to experience the free version before making a purchase. The issue still remains with convincing people to make the jump from a free dataset to a paid one. This leads us back to prior research which highlighted organisations such as OS the differing types of consumers; leisure users and commercial organisations. We also directed our research in the LD experiment to investigate the behaviours of three different types of users. Point and Shoot (Leisure users), keen amateurs and professional users who would all react differently to purchase of data online.

8.7.2 Willingness to pay for data

From the research carried out in the previous chapters we note that there is the issue of how to convince users to pay for data when there is a free option available.

We suggest that again we utilise the metadata attached to the data. This could be made available in a form which allows users to check that it covers all of the areas they would expect the data to contain. I.e. Date, coverage, format, etc. This could be a pop up box which is displayed when the user has the option to choose free or paid for data.

8.8 Conclusion

In this chapter we have established a framework for the consumption of LOD and LCD online. We have highlighted the key technologies which we suggest to be most suitable for the framework and detailed the key actors and processes. We also describe two different revenue models for which LD can be sold through and the licenses to which can be attached to the data to ensure that the data provider maintains control of data which they own. Following this we have addressed the issues which are most cause for concern and have proposed possible solutions. We concluded with an example of the use of LD for OS and how different types of data could be integrated into such a LD application.

From the architecture we have outlined in this chapter we have demonstrated the different options available using the addressing problem as an example. The addressing issues is a good example for LD as there are many elements within the address and each have different potential values. We have demonstrated how the architecture can provide certain elements of an address for free and then redirect users to a paid resource where they can enter payment details to gain access. We also illustrate how we can deny access to resources should the data provider deem this necessary.

In the next section we will conclude this research and propose future directions to extend the research in the future.

Chapter 9

Conclusions

The aim of this research was to answer a number of specific questions regarding the derivation of value from geospatial [LD](#).

In our research we outline the architecture of the World Wide Web and the key technologies used, we then go on to identify the role of [LD](#) and the SW and the affordances for both. In Chapter 3 we outlined the relevance of PSI and the role of data providers for [LD](#). We then looked at GI and established the importance of GI for [LD](#). Chapter 4 draws upon the third theme for this research, which is the business case. This chapter also introduced potential revenue models for [LD](#) and looked at similar digital content industries. Following our review of the literature we began Chapter 6 with our requirements elicitation which contained investigation into consumers of OS products and then drew together potential consumers of [LD](#) to identify factors which were of concern to them with a new technology. We then looked at users response to specific information quality criteria factors in order to prioritise them for investigation.

Chapter 6 described two specific [LD](#) experiments which we designed and implemented to discover the willingness of consumers to pay for [LD](#) when free alternatives are available. We then, from this experimental work, devised an architecture for linked geospatial data.

Chapter 7 brings together all of the findings from our preliminary investigations into a suggested user driven architecture for linked geospatial data.

Each chapter contributed to answering the questions we posed in Chapter 1. Chapter 6 addressed the user driven aspect of the research where we interacted with potential users of [LD](#) to explore finer niceties regarding information quality and how these were applicable to [LD](#). The experiments in Chapter 7 looked at the actions users took when interacting with a simulated [LD](#) application.

The main contributions made in this thesis are:

1. Identification of the different types of consumers of LD. We established that there were either professional or leisure users.
2. Identification of the factors which affect a users willingness to pay. We tested through the experiments outlined in Chapter 7.
3. Determine the quality factors of data which affect a users decision. These factors were also tested through the experiments in Chapter 7
4. A use case for the consumption of LD. We have detailed this use case in Chapter 8.

The research we have carried out to date has supported the hypotheses we outlined in Chapter 1.

h_1 – Does criticality of purpose affect a user’s decision whether or not to pay for premium LD when free alternatives are available?

h_2 – WIs willingness to pay for premium data positively influenced by consumers’ perceptions of its value?

The specific findings from the research are detailed in the following sections.

9.1 User Requirements

Chapter 6 looked at the specific requirements of technical and non-technical users for the implementation of a new technology and specifically to LD. The first investigation discovered the current users of OS data and the ways they facilitated the data. We found that there were three different types of users; personal leisure users (who used GI to map routes for walks and personal interest), a second type of leisure user (who used the data for clubs and societies to give information to its members about planned events and activities, and the third type of user (those with a commercial interest who used the data to see if they valued it enough to make a purchase of further data). This is where we began to think about the potential for LOD and LCD.

Once we had determined the types of consumers we then went on to look at finding potential users to give us feedback about the possible concerns they may have with the implementation of a new technology. We did this by organising the Terra Future event whereby technical users and GI specialists were brought together to learn about the technology and to discuss. We found that the key areas which gave concern to potential users were; the licensing of the data, the costs involved and technical ability of potential users. We took these concerns into account and used them to inform the LD experiments detailed in the later chapters.

The final stage of our requirements elicitation involved a questionnaire which was sent to 100 participants. They were asked to rank the data in order of preference and were shown two options at each stage. We found that there were definite information quality factors which users rated more important over others. These included; accuracy, completeness and consistency. Using this information we were able to inform two LD simulations which we carried out in Chapter 7.

9.2 User Interaction with Linked Data

Chapter 6 was used to determine the requirements of the users for a LD system which we chose to design and implement in Chapter 7. Our first experiment in Chapter 7 aimed to test the findings we had established from our requirements elicitation, in which we wanted the users to experience a LD environment and see the potential of LD. Our second experiment clarified the different types of user; the first who will always pay for data regardless of the situation, the second type of user who will pay for the information if they deem it to be useful and the third type of user who will never pay for data online. We found that there were varying reasons given for willingness to pay and how often concerns related to security, mainly disclosing payment details online. We used the results of these two experiments and our user requirements elicitation to inform the architecture for Linked Geospatial data.

The results generated from the user interaction experiments were used to put together a user driven architecture for linked geospatial data. We outline the architecture in the next section.

9.3 Architecture for Linked Geospatial Data

Following the experiments of Chapters 7 we were able to establish an architecture for the consumption of LOD and LCD. We proposed a system which would support the consumption of data which has sensitive or restricted content such as information which holds a monetary value to the holding organisation or is sensitive due to its content, or PSI which is free and available to all. We suggest the use of status codes in order to direct the user to the specific end, which may involve entering a username and password, entering payment details, restrict access or allow access. The work we have carried out to date is a suggestion for the consumption of LD and is open to further development which we explain in the next section.

9.4 Future Work and Research Directions

Following the conclusions from the experiments we have carried out and the outline of the user driven architecture, we are able to highlight areas which will require further investigation in order to envisage a working model for the consumption of [LOD](#) and [LCD](#).

The aim of the current research was to establish a potential frame-work for the consumption of data. We suggest that it would be beneficial for a working prototype of the proposed architecture for linked geospatial to allow potential users to interact with the system to ensure that it works in a manner which is acceptable for both a supplier and a consumer of [LD](#). There are a considerable number of elements included within the framework and before conclusions are drawn regarding each element, we suggest that testing should be carried out into areas such as; authorisation and authentication, payment and security to determine their suitability.

9.4.1 Implementation of the suggested technical architecture

The aim of the research we have carried out to date was to establish a potential framework for the consumption of data. We suggest that a working prototype of the proposed architecture for linked geospatial data would be beneficial to allow potential users to interact with such a system and ensure that it works in a manner which is acceptable for both a supplier and a consumer of [LD](#). There are a considerable number of elements included within the framework and before conclusions are drawn regarding each element, we suggest testing should be carried out into areas such as authorisation and authentication, payment and security to determine their suitability.

9.4.2 Usability and HCI

One of the key features for [LD](#) is that it is accessible, therefore we suggest that the sites used to view data and download specific datasets are compatible with all platforms from Windows, Linux and OSX to hand held devices, which incorporate mobile platforms, such as mobile phones and tablets. This will enhance the usability and accessibility of the data for all consumers and will also include the development of applications which will have the potential to create mashups of data by potential non-technical users.

Whilst our experiments were aimed at introducing consumers to the potential of [LD](#), we are aware that there are certain usability and human-computer interaction issues which may have an effect on the usability of the data. For example, such as how the consumer will access the data through SPARQL endpoints and then view the data on

screen. These should be addressed in order to make potential LD applications attractive and easy for use non-technical users.

9.4.3 Return on Investment

The key question asked by companies looking to implement LD as a product is, what is the return on investment? This is both the revenue which can be generated and which also has the potential for non-monetary returns. These include the consumer engagement with potential products. As more data is available and tools are developed to link this data, consumers will be able to interact not only with the data via the creation of their own applications but also through interaction with applications developed by others. For a company, this uptake in use of its products and data, will in turn increase brand awareness. As can be seen from the LD cloud diagram which we show in Chapter 2, the amount of data published on the web is ever increasing. A presence on the web as a data provider in a world where data has become a raw material, holds potential for ROI from both the data and potential applications.

We look at OS as an example for other companies who may consider LD as a product. We recommend that initially creating a free set of data which can be released as a way of understanding how the technology works and to iron out the technical issues which may arise. Following the release of this data the company should monitor the usage to see who its users are and the ways in which the data is used. Once this has been established further more valuable, paid for content can be added and the usage quantitatively measured. This can then be used to establish the qualitative return on investment.

9.4.4 Test for speed of data discovery

As aforementioned, the key question for companies relate to the return on investment. We suggest that testing and development of the speed of data discovery is important to demonstrate to the consumer that the use of data can solve issues of time spent searching for data. The faster that consumers can gain access to an organisations data, the greater the likely satisfaction from achieving the desired outcome. It is important for the speed to be tested to be able to determine that the use of LD applications is faster than simple web browsing and that it will help contribute to answering the question regarding return of investment.

9.4.5 Pricing

We have established a significant amount of data which will be available for free and thus call this LOD. We also note that there will be data which is not available for free

which we call LCD. This LCD may not just be closed due to licensing and privacy issues, but also to a company not wishing to, or not being able to, financially provide it for free. We also established that although there are consumers who will not pay for data, there are some consumers who would be willing to pay. Therefore, in order to charge for data, we need to establish the prices which consumers are willing to pay for such data.

This will include a variety of pricing structure as identified earlier in this research to include single payments and subscription options. Empirical research carried out by [Melnik and Alm \(2003\)](#) suggests that reputation has a direct effect on a consumers willingness to pay for goods online and suggests that a negative feedback score on an online auction site such as eBay will directly affect the willingness to pay for a purchase. We suggest that incorporating brand awareness and a form of ranking data providers by openly revealing their reputation as a data provider is incorporated into further investigations. Furthermore, we also note that it would be beneficial to discover if more trusted organisations or those with a better reputation are able to sustain charging higher prices for information or whether consumers will always rather choose an inexpensive or free option in the presence of charged items.

We observe from our research that there is the potential for more value to be had from the sale of applications using linked data than from the sale of the data itself. We suggest that this is a consideration which should be investigated further to help companies decide if being a data provider will generate required income or if further developments should be made to build applications.

9.4.6 New Classes of Data

Due to the complexity of the GI, the development of GI into LD was a huge challenge. We have observed that there is a considerable amount of PSI readily available and that some organisations such as OS are publishing data and investigating the potential for LD. However, we also recognise the need to investigate the linking of new and different classes of data such as the linking of personal information too which has adds a new dimension to LD.

9.4.7 Awareness of the Potential of LD

In order for LD to gain further momentum in the LD movement, we suggest that it would be of particular benefit to generate more awareness to the potential of LD and its applications. We note that with the start of the Open Data Institute¹ and its involvement with government, potential companies and users will be made more aware of the capability of the technology.

Appendix A

Terra Future Invite

The Terra Future Invites are reproduced on the following pages.



Forging links seminar

It is well known that geography forms an important medium to enable data from different sources to be associated. It does this through the fact that location is often a common theme across datasets. This power to unite data has been central to a number of initiatives, including the Digital National Framework and, more recently, as a central core to the UK Location Strategy. And it has also formed the glue that enables many of the more informal mash-ups to be generated.

If the focus is moved away from geographic information (GI) and one looks for wider trends in the information economy, then a new way of organising information is emerging – the linked data web. Although relatively young and an offshoot of the semantic web, the linked data web has begun to grow quite rapidly. It has been recognised by link data practitioners that location based information can provide a valuable means to assist the growth of the linked data web, and indeed some of the core datasets currently published on the linked data web are rich in location data. However, it is also clear that in the majority of cases where location is exploited, it is not always fully done so. Similarly, the majority of people in the GI community will either know only vaguely about the nature of the linked data web or will have not heard of it at all.

This special Terra future™ seminar aims to bring the two communities together: to inform the GI community about the power of linked data, to inform the linked data community about the power of GI, and, most importantly, to **forge links** between the two communities. If linked data is about modern ways to interconnect information, and if part of the power of GI lies in enabling data to be linked, then Terra future is about creating links between communities.

Who should attend?

Those in the GI community wishing to find out about the linked data web and those in the linked data community wishing to find out how to exploit geographic data.

Where, when, how much?

The event will be held in the Ordnance Survey Business Centre ([map](#)) on 10 March. Better still, it's free!

What is the plan for the day?

A provisional agenda is outlined below:

- | | | |
|-------------|--|--|
| 09.15–10.00 | Registration and coffee | |
| 10.00–10.15 | Welcome and introduction | – Peter ter Haar, Ordnance Survey |
| 10.20–10.40 | Linked data in a nutshell | – Tom Heath, Talis® |
| 10.40–11.00 | The power of GI | – Liz Ratcliffe, Ordnance Survey |
| 11.00–11.15 | GeoVation™ | – Chris Parker, Ordnance Survey |
| 11.15–11.30 | Coffee break | |
| 11.30–11.50 | Linked data and <i>the Beeb</i> | – Tom Scott and Silver Oliver, BBC® |
| 11.50–12.10 | Linked data in government | – speaker TBC |
| 12.15–13.00 | Lunch | |
| 13.00–13.45 | Linked data technical workshop and panel discussion | – Hugh Glaser, Sameas.org and University of Southampton® |
| 13.45–14.30 | Geographic information workshop and panel discussion | – Brian Higgs, Dudley Metropolitan Borough Council and Digital National Framework; Ian Holt, Ordnance Survey |
| 14.30–14.45 | Coffee break | |
| 14.45–15.45 | Group idea generation | – Delegates form a number of discussion groups with equal numbers from each community in facilitated discussion around practical joint uses of linked data and GI. |
| 15.45 | Closing address and summary | – Glen Hart and Liz Ratcliffe, Ordnance Survey |



How do I register?

To register please fill out the form below and send it to us at terrafuture@ordnancesurvey.co.uk.

Please state:

- Name

- Organisation (if applicable)

- Whether you are a member of the GI community, Linked data, both or neither

- Any special dietary considerations?

- Whether you will require parking

Any other comments

Places will be limited to around 140 (approximately 70 from each community), so book early to avoid disappointment!

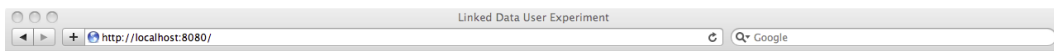
Please feel free to forward this invitation to interested parties.

Appendix B

Parking Experiment

The following images illustrate the screens from the [LD](#) simulation.

Screen 1 - Homepage



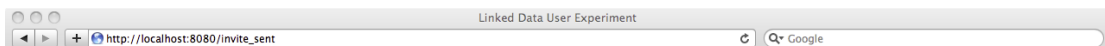
Welcome to the Ordnance Survey Linked Data Experiment



Begin by entering your **email address**.

Please note: By entering your email address you agree to participate in this study.

Screen 2



Please check your email and follow the link in the email to start the experiment

Screen 3 – Instructions page

Instructions for Study

http://localhost:8080/instructions

Study Instructions

The aim of this study is to find the best location using the options provided.

You will have 3 different data options, Free, Paid and Premium.
Select these by clicking the box next to it. Use these datasets to find the answer to the problem.

Can you find the best answer?

You are looking for a some example data

Please consider:

- You want to know there is an available space.
- You want to know that there are shops nearby.
- You would like to know the price of the parking before you leave.

Please select the maptype and data from the options below and browse through the map to find the best solution.

Map and Data Type	Summary	
Free Data: Free	Displays the data you get for free	<input type="radio"/>
Paid Data: 50 Map Groats	Displays the data you get for 50-map Groats	<input type="radio"/>
Premium Data: 100 Map Groats	Displays the data you get for 100-map Groats	<input type="radio"/>

Balance:
Map Groats 500

Next

Screen 4 – First Scenario

Parking

http://localhost:8080/parking

Can you find the best place to park on a Saturday afternoon?

You are looking for a parking space in Bournemouth Town centre as you are going shopping for a birthday present.

Please consider:

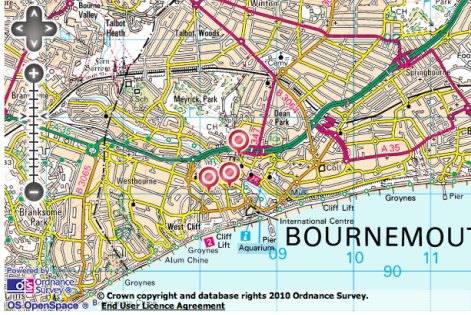
- You want to know you there is an available space.
- You want to know that there are shops nearby.
- You would like to know the price of the parking before you leave.

Please select the maptype and data from the options below and browse through the map to find the best solution.

Map and Data Type	Summary	
Free Data: Free	Car park locations	<input type="radio"/>
Paid Data: 50 Map Groats	Car park location, Post Code, Total number of spaces	<input type="radio"/>
Premium Data: 100 Map Groats	Car park location, Post Code, Total number of spaces, Current spaces, Nearby Shopping Centres	<input checked="" type="radio"/>

Balance:
Map Groats 500

Next



Screen 5 – Scenario Question Page

The screenshot shows a web browser window with the address bar displaying 'http://localhost:8080/parkingquest'. The browser's menu bar includes 'Spitefire', 'Experiment', 'Hotmail', 'ECS Mail', 'C4 OD', 'Google', 'LIBRARY STUFF', 'Jen', 'Horse', 'RESEARCH', and 'SOTON'. The page content is as follows:

Thank you for completing the first stage

Please answer the following questions regarding your search for parking data:

Question	Please select your answer
1. Which dataset(s) did you use to get your answer?	Free Data <input type="radio"/>
	Paid Data <input type="radio"/>
	Premium Data <input type="radio"/>
	Free, Paid & Premium Data <input type="radio"/>
	Free & Paid Data <input type="radio"/>
	Free & Premium Data <input type="radio"/>
	Paid & Premium Data <input type="radio"/>
<hr/>	
2. If you chose to use only free, why did you chose this data?	I preferred the map background <input type="radio"/>
	I would not be prepared to pay for data <input type="radio"/>
	I would rather look for the information myself <input type="radio"/>
	I did not find the additional information useful/important <input type="radio"/>
	I would not be prepared to disclose my payment details online <input type="radio"/>

[Submit]

Final Page – Deception Statement

The screenshot shows a web browser window with the address bar displaying 'http://localhost:8080/finish'. The browser's menu bar includes 'Survey Complete' and 'Gmail - Completed Survey - linke...'. The page content is as follows:

Thank you for completing the experiment

You have successfully completed this study your reward will be forwarded to you.

Deception Statement

Although you were told at the beginning the purpose of this experiment was to discover your willingness to pay for a mobile phone application, the real reason was to discover if you would be willing to pay for linked data online.

In order for us to obtain a true understanding of your willingness to pay for data it was important we didn't ask you directly as the could have altered your response to the different scenarios.

Your responses will still remain anonymous and if you have any questions please contact the researcher via email jlb08r@ecs.soton.ac.uk or linkeddataexperiment@gmail.com

Following the experiment the participants were directed to a questionnaire. The questions from this questionnaire are shown below.

1. Which dataset(s) did you use to get your answer?
 - (a) Free Data
 - (b) Paid Data
 - (c) Premium Data
 - (d) Free, Paid and Premium Data
 - (e) Free and Paid Data
 - (f) Free and Premium Data
 - (g) Paid and Premium Data

2. If you chose to use only free, why did you chose this data?
 - (a) I preferred the map background
 - (b) I would not be prepared to pay for data
 - (c) I would rather look for the information myself
 - (d) I did not find the additional information useful/important
 - (e) I would not be prepared to disclose my payment details online
 - (f) Other (please state)

3. If you chose to use premium data or a combination of data, why was this?
 - (a) It was quicker to use premium data
 - (b) I preferred the map background
 - (c) The premium data gave me all the information I needed
 - (d) Other

4. Any other comments about this scenario/data?

Appendix C

Information Quality Questionnaire

The questionnaire which was sent to participants online is displayed on the following pages.

Information Quality Assessment

Thank you for offering to take part in this study.

The study should take no longer than 10 minutes to complete.

1. Which category below includes your age?

- 17 or younger
- 18-20
- 21-29
- 30-39
- 40-49
- 50-59
- 60 or older

Information Quality Assessment

Instructions

We are looking into the criteria which affect your decision to use free information or pay for information on the web.

There are a number of different information sources which you can use on the web. Some, such as buyer reviews on Amazon and manufacturer websites will be free. Others, such as those from professional reviewers (for example Which?) will charge. Each of these sources have certain attributes which make them appealing to you as a consumer.

Please consider this scenario:

Imagine you wish to buy an expensive new digital camera. Before purchasing you want to compare a number of different potential cameras online.

Information Quality Assessment

Information Selection Criteria

We have identified 6 different attributes of information and would like you to consider these attributes when looking at information online.

ACCURATE

The extent to which information is correct, reliable and verified free of error.

Example: inaccurate data might incorrectly state the size of the camera's sensor (in megapixels).

CONSISTENT

The extent to which information is presented in the same format and compatible with previous information.

Example: an inconsistent review might list different information about a particular camera, so that it will be difficult to make a comparison between cameras..

SECURE

The extent to which access to information is restricted appropriately to maintain its security.

Example: a insecure review site might require you to register with your personal information (such as name and email) and take insufficient care to protect your personal information from others.

TIMELY

The extent to which the information is sufficiently up-to-date for the task at hand.

Example: an untimely review might contain information about a camera which is no longer available.

COMPLETE:

The extent to which information is not missing and is of sufficient breadth and depth for the task at hand.

Example: an incomplete review of a camera might omit particular information, such as the size of its sensor.

CONCISE

The extent to which information is compactly represented without being overwhelming

Example: an in-concise review of a camera makes it hard to find particular information about the camera due to poor organisation (putting key facts in a long paragraph).

You will be shown pairs of attributes and will be asked to choose the most important to you.

Information Quality Assessment

***2. For the following attributes please choose the most important to you**

	First Attribute	Second Attribute
Accurate or Complete	<input type="radio"/>	<input type="radio"/>
Accurate or Consistent	<input type="radio"/>	<input type="radio"/>
Consistent or Timely	<input type="radio"/>	<input type="radio"/>
Accurate or Secure	<input type="radio"/>	<input type="radio"/>
Accurate or Concise	<input type="radio"/>	<input type="radio"/>
Secure or Consistent	<input type="radio"/>	<input type="radio"/>
Consistent or Concise	<input type="radio"/>	<input type="radio"/>
Accurate or Timely	<input type="radio"/>	<input type="radio"/>
Timely or Concise	<input type="radio"/>	<input type="radio"/>
Complete or Consistent	<input type="radio"/>	<input type="radio"/>
Complete or Timely	<input type="radio"/>	<input type="radio"/>
Complete or Concise	<input type="radio"/>	<input type="radio"/>
Secure or Timely	<input type="radio"/>	<input type="radio"/>
Secure or Concise	<input type="radio"/>	<input type="radio"/>
Secure or Complete	<input type="radio"/>	<input type="radio"/>

Appendix D

Information Quality Questionnaire Significance Results

- A 2 x 2 Chi-square test was carried out to examine the association between Secure and Accurate. There was no significant association between Secure and Accurate [Chi-square (1, N = 99) = 3.646, $p > .05$ (computed $p = 0.056$, which is just outside of significance) , Cramer's V = 0.19].
- A 2 x 2 Chi-square test was carried out to examine the association between Accurate and Complete. There was no significant association between Accurate and Complete [Chi-square (1, N = 99) = 24.253, $p > .05$ (computed $p = 0.0005$, which is just outside of significance) , Cramer's V = 0.49
- A 2 x 2 Chi-square test was carried out to examine the association between Accurate and Consistent. There was no significant association between Accurate and Consistent [Chi-square (1, N = 99) = 35.162, $p > .05$ (computed $p = 0.0005$, which is just outside of significance) , Cramer's V = 0.6].
- A 2 x 2 Chi-square test was carried out to examine the association between Accurate and Timely. There was a significant association between Accurate and Timely [Chi-square (1, N = 99) = 45.343, $p < .001$ (computed $p = 0.0005$), Cramer's V = 0.68].
- A 2 x 2 Chi-square test was carried out to examine the association between Accurate and Concise. There was a significant association between Accurate and Concise [Chi-square (1, N = 99) = 26.273, $p < .001$ (computed $p = 0.0005$), Cramer's V = 0.52].
- A 2 x 2 Chi-square test was carried out to examine the association between Secure and Complete. There was a significant association between Secure and Complete

[Chi-square (1, N = 99) = 4.455, $p < .05$ (computed $p = 0.035$), Cramer's V = 0.21].

- A 2 x 2 Chi-square test was carried out to examine the association between Secure and Consistent. There was no significant association between Secure and Consistent [Chi-square (1, N = 99) = 1.222, $p > .05$, Cramer's V = 0.11].
- A 2 x 2 Chi-square test was carried out to examine the association between Timely and Secure. There was a significant association between Timely and Secure [Chi-square (1, N = 99) = 11.000, $p < .005$ (computed $p = 0.001$), Cramer's V = 0.33].
- A 2 x 2 Chi-square test was carried out to examine the association between Concise and Secure. There was a significant association between Concise and Secure [Chi-square (1, N = 99) = 11.000, $p < .005$ (computed $p = 0.001$), Cramer's V = 0.33].
- A 2 x 2 Chi-square test was carried out to examine the association between Complete and Concise. There was a significant association between Complete and Concise [Chi-square (1, N = 99) = 15.364, $p < .005$ (computed $p = 0.0005$), Cramer's V = 0.39].
- A 2 x 2 Chi-square test was carried out to examine the association between Complete and Consistent. There was no significant association between Complete and Consistent [Chi-square (1, N = 99) = 2.919, $p > .05$ (computed $p = 0.088$, which is a trend towards significance) , Cramer's V = 0.17].
- A 2 x 2 Chi-square test was carried out to examine the association between Complete and Timely. There was a significant association between Complete and Timely [Chi-square (1, N = 99) = 7.364, $p < .01$ (computed $p = 0.007$), Cramer's V = 0.27].
- A 2 x 2 Chi-square test was carried out to examine the association between Consistent and Concise. There was a significant association between Consistent and Concise [Chi-square (1, N = 99) = 22.313, $p < .01$ (computed $p = 0.0005$), Cramer's V = 0.47].
- A 2 x 2 Chi-square test was carried out to examine the association between Consistent and Timely. There was a significant association between Consistent and Timely [Chi-square (1, N = 99) = 7.364, $p < .01$ (computed $p = 0.007$), Cramer's V = 0.27].
- A 2 x 2 Chi-square test was carried out to examine the association between Timely and Concise. There was no significant association between Timely and Concise [Chi-square (1, N = 99) = 0.253, $p > .05$, Cramer's V = 0.05].

Appendix E

Linked Data Study Screen Shots

The following images are screen shots of the screens from the Linked Data Study.

Welcome to the Information Quality Study

Please read each statement and check each box to agree to take part in the study.

- a. I confirm that I have read and understand the project description.
- b. I understand that my participation is voluntary and that I am free to withdraw at any time and keep any incentive payment.
- c. I understand that at the end of the study the data collected will be stored at the University of Southampton and used for research studies. (All personal information will be deleted on completion of the analysis of data.)
- d. I agree to take part in the above study.

Please enter your participant number:

In association with:



SEC Number: N/11/10/01

Instructions

Imagine you wish to buy an expensive new digital camera. Before purchasing you want to compare a number of different potential cameras.

There are a number of different information sources which you can use: for example, buyer reviews on Amazon and manufacturer websites will be free; others, such as those from professional reviewers for example Which will charge.

You are given £5 to spend on information about the cameras. You do not have to buy any information in which case you will only have access to free information, or you may choose to purchase information on an individual basis or subscribe to all the information.

If your choice of camera best matches the camera we think the information indicates is best then you keep all of the money that is unspent.

If however, you do not select the best camera you have to return any unspent money.

You will always keep the £10 we pay to compensate for your time and also your expenses will be covered.

Customer Selection

Please select the customer you were allocated.

- Keen Amateur This customer would like a camera which takes good quality pictures with manual controls, budget might be an issue.
- Point & Shoot user This customer would like a camera which has fully automatic controls which are considered simple to use.
- Professional Photographer This customer would like a camera which will take high quality photographs and has full functionality, price is less of an issue.

Continue

Pre-purchase a subscription?

A subscription to a collection of 'premium' camera reviews is available.

The subscription will cost £3.00.

You can choose to subscribe later, at any time during the study.

Yes No/Later

You have a starting balance of £5

£5.00

Camera Reviews

Balance: £5.00

When you have reached a decision please click here to enter you answer

Make Selection

b

Best Buy	£1.00
Manufacturer's Data	free
Expert Paid Review	£1.00
Best Buy Plus	£1.00
Free customer review	free
More free customer reviews	free

c

Best Buy	£1.00
Manufacturer's Data	free
Expert Paid Review	£1.00
Best Buy Plus	£1.00
Free customer review	free
More free customer reviews	free

a

Best Buy	£1.00
Manufacturer's Data	free
Expert Paid Review	£1.00
Best Buy Plus	£1.00
Free customer review	free
More free customer reviews	free

Camera B

November 2011



Camera B is a fraction cheaper than others and has a 12.9-megapixel CMOS sensor compared with B's 15.1-megapixel sensor.

Screen and Battery

We have no similar concerns about the body itself. Weighing 884g including the battery and lens, Camera B is solid in the hand and feels like it will take the odd knock well. The manufacturer has attempted to set it apart from the crowd of consumer DSLRs by adding a hinge to the 2.7in LCD on the back. We found little practical use for it, though – it's theoretically handy for shooting over the heads of a crowd or taking shots low to the ground, as the screen can be angled up or down, but there's no way to shoot around corners, for instance.

Customisation

Like its more expensive stablemate, Camera B features an accurate 11-point focus system – a clear upgrade from the previous versions three-point system. It is also faster than others on the market, offering a maximum of four frames per second (fps) in continuous-shooting mode. Although this impressive speed is only maintained for the first 10 shots, Camera B also includes auto-ISO, which allows you to set the slowest permissible shutter speed at which the ISO should be raised, as well as the effective D-Lighting setting, which makes images appear to have greater contrast.

Speed

Camera B takes gorgeous images and it's experience in the sub-£1000 DSLR market shows: Camera B's high-ISO performance has to be seen to be believed. At its highest extreme, it can be pushed to one stop over ISO3200 – ISO6400, in other words. Performance at ISO3200 was superb, and we'd be happy to use the Camera B at ISO800 for nearly every kind of shot. For noise-phobic users, Camera B's lowest ISO is 200, and can even be pushed a stop below that.

The kit lens, budget 18.55mm (f/3.5-5.6 VR – isn't going to set anyone's pulse racing, although we were pleasantly surprised by the lack of distortion at wide angles. There was occasionally a lack of sharpness, although nothing that a few minutes in Aperture wouldn't fix. Of slightly more concern was the lens' all-plastic construction – even the mount that connects the lens to the camera is plastic. The zoom motion is less than perfectly smooth, and purists will lament the lack of distance numbers on the focus ring.

Video

Please select your chosen Camera

Please select your chosen camera from the list below:

(Please note, you will not have been shown all of the cameras listed below)

- Camera A
- Camera B
- Camera C
- Camera D
- Camera E
- Camera F
- Camera G

Submit

Appendix F

Linked Data Questionnaire - Post Study

The questionnaire which participants of the Linked Data study completed following the study

Linked Data Post Study Questionnaire

Post Study Questionnaire

Having completed the study, please fill in this questionnaire regarding your experience of the study.

1. Please select the type of user you were allocated

- Keen Amateur
- Point and Shoot
- Professional

2. Did you find that the type of user you were allocated influenced the decision you made whether to purchase information?

- Yes
- No

3. Did you choose to subscribe to the information from the beginning?

- Yes
- No

4. If no did you pay for information later?

- Yes
- No

5. If you didn't subscribe to the information did you use free or both free and paid?

- Free only
- Paid only
- Both

6. Which of these criteria was most important to you when you were looking for information

- Accurate
- Complete
- Consistent

Other (please specify)

Linked Data Post Study Questionnaire

7. 7. If you had to purchase information on the web, what are your reasons for not paying for data online? (Please choose all that apply)

- Concern about credit card security
- Would prefer to look for information myself
- Cannot guarantee the information will be useful

Other (please specify)

8. Would you prefer to spend time searching for data or pay a small fee for getting an answer instantly?

- Small Fee
- Spend Time

9. How much would you trust the checkout/payment process online?

- A great deal
- A lot
- A moderate amount
- A little
- None at all

Bibliography

- Aichholzer, G., 2004. Electronic access to Public Sector Information: Some key issues. *Lecture Notes in Computer Science* 3183, 525–528.
- Aichholzer, G., Burkert, H., 2004. Public sector information in the digital age: between markets, public management and citizens' rights. Edward Elgar Publishing.
- Alani, H., Dupplaw, D., Sheridan, J., Hara, K. O., Darlington, J., Shadbolt, N., Tullo, C., 2007. Unlocking the Potential of Public Sector Information with Semantic Web Technology. *Lecture Notes in Computer Science* 4825, 708–721.
- Alexander, K., 2008. Rdf/json: A specification for serialising rdf in json. *Scripting for the Semantic Web (SFSW)*.
- Allan, R., 2009. Power of Information Taskforce Report. Study Report, UK Government Cabinet Office.
URL <http://poit.cabinetoffice.gov.uk/poit/>
- Allemang, D., 2010. Semantic Web and the Linked Data Enterprise. In: Wood, D. (Ed.), *Linking Enterprise Data*. Springer US, Boston, MA, pp. 3–23.
- Anderson, C., 2006. *The Long Tail: Why the Future of Business is Selling Less of More*. Hyperion.
- Anderson, C., 2009. *Free: The Future of a Radical Price: The Economics of Abundance and Why Zero Pricing Is Changing the Face of Business*. Random House Books.
- Antoniou, G., van Harmelen, F., 2008. *A Semantic Web Primer*. The 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 (2), 58–71.
- Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z., 2007. Dbpedia: A nucleus for a web of open data. *The Semantic Web*, 722–735.
- Auer, S., Lehmann, J., Hellmann, S., 2009. LinkedGeoData: Adding a spatial dimension to the Web of Data. In: *Lecture Notes in Computer Science*. Vol. 5823. Springer, pp. 731–746.

- Bailey, N., 2006. Guide to the Establishment and Operation of Trading Funds. Report, Financial Reporting Policy Team - HM Treasury.
- Balfanz, D., de Medeiros, B., Recordon, D., Smarr, J., Tom, A., 2009. OpenID OAuth Extension.
URL http://svn.openid.net/repos/specifications/oauth_hybrid/1.0/trunk/openid_oauth_extension.html
- Barker, E., Duncan, C., Guadamuz, A., Hatcher, J., Waelde, C., 2005. The Common Information Environment and Creative Commons: a study on the applicability of Creative Commons licences. Report, Intrallect.
URL http://www.eduserv.org.uk/research/studies/~media/Foundation/pdf/CIE_CC_Final_Report%20pdf.ashx
- Beaumont, P., Longley, P. A., Maguire, D. J., 2005. Geographic information portals-a UK perspective. *Computers, environment and urban systems* 29 (1), 49–69.
- Becker, C., Furness, P., Jan 2010. Linking Spatial Data from the Web. *Journal of Direct, Data and Digital Marketing Practice* 11 (4), 317–323.
- Belam, M., 2010. What is the value of Linked Data to the news industry?
URL <http://www.guardian.co.uk/help/insideguardian/2010/jan/25/news-linked-data-summit>
- Benslimane, D., Dustdar, S., Sheth, A., 2008. Services mashups: The new generation of web applications. *IEEE Internet Computing* 12 (5), 13–15.
- Bergman, M. K., Aug 2001. White Paper: The Deep Web: Surfacing Hidden Value. *Journal of Electronic Publishing* 7 (1), 1–34.
- Berners-Lee, T., Hendler, J., Lassila, O., 2001. The Semantic Web. *Scientific American* 284 (5), 34–43.
- Beuscart, J., Mellet, K., 2008. Business models of the web 2.0: Advertising or the tale of two stories. *Communications and Strategies (Special issue)*, 165.
- Bizer, C., 2009. The emerging web of linked data. *IEEE Intelligent Systems* 24, 87–92.
- Bizer, C., Boncz, P., Brodie, M., Erling, O., 2012. The meaningful use of big data: four perspectives—four challenges. *ACM SIGMOD Record* 40 (4), 56–60.
- Bizer, C., Heath, T., Berners-Lee, T., 2009. Linked data—the story so far. *International Journal On Semantic Web and Information Systems* 5 (3), 1–22.
- Bizer, C., Heath, T., Idehen, K., Berners-Lee, T., 2008. Linked Data on the Web. In: *Proceedings of the 17th International Conference on World Wide Web*. pp. 1265–1266.

- Bonatti, P. A., Hogan, A., Polleres, A., Sauro, L., Jul 2011. Robust and scalable linked data reasoning incorporating provenance and trust annotations. *Web Semantics: Science, Services and Agents on the World Wide Web* 9 (2), 165–201.
- Caro, A., Calero, C., Caballero, I., Piattini, M., 2005. Data quality in web applications: A state of the art. In: *IADIS International Conference WWW/Internet*. Vol. 2. pp. 364–368.
- Chakrabarti, S., Dom, B., Indyk, P., 1998. Enhanced hypertext categorization using hyperlinks. In: *ACM SIGMOD Record*. Vol. 27. ACM, pp. 307–318.
- Chan-Olmsted, S., 2004. Introduction: Traditional media and the internet: The search for viable business models. *International Journal on Media Management* 6 (1), 2–3.
- Cehade, A., Jan 2011. Dominant revenue streams in the web 2.0 era. Tech. rep., Southern Illinois University Carbondale.
URL http://opensiuc.lib.siu.edu/cgi/viewcontent.cgi?article=1179&context=gs_rp
- Chesbrough, H., Rosenbloom, R., 2002. The role of the business model in capturing value from innovation: evidence from Xerox Corporation’s technology spin-off companies. *Industrial and Corporate Change* 11 (3), 529–555.
- Chi, E., Jan 1996. Evaluation of micropayment schemes. Tech. Rep. HPL-97-14, Hewlett Packard Labs.
URL <https://www.hpl.hp.com/techreports/97/HPL-97-14.pdf>
- Chilton, S., 2009. Crowdsourcing is radically changing the geodata landscape: Case study of openstreetmap. In: *24th International Cartographic Conference, Chile*. Retrieved from <http://bit.ly/H8PfTw>.
- Chyi, H. I., Aug 2005. Willingness to pay for online news: An empirical study on the viability of the subscription model. *Journal of Media Economics* 18 (2), 131–142.
- Clay, K., Goettler, R., Wolff, E., 2003. Consumer learning about experience goods: Evidence from an online grocer. Tech. rep.
- Cobden, M., Black, J., Gibbins, N., Shadbolt, N., Jan 2010. Consuming Linked Closed Data. eprints.ecs.soton.ac.uk.
URL <http://eprints.ecs.soton.ac.uk/21686/>
- Cobden, M., Black, J., Gibbins, N., Shadbolt, N., 2011. A Research Agenda for Linked Closed Data. In: *Workshop on Consuming Open Linked Data at The 10th International Semantic Web Conference*. <http://eprints.soton.ac.uk/272711/3/position.pdf>.
- Coote, A Smart, 2010. The Value of Geospatial Information to Local Public Service Delivery in England and Wales. Tech. rep., Local Government Association Analysis and Research, www.lga.gov.uk/GIresearch.

- Crockford, D., 2006. The application/json media type for javascript object notation (json). Rfc4627, Internet Engineering Task Force.
- Dai, X., Grundy, J., Lo, B., 2001. Comparing and contrasting micro-payment models for e-commerce systems. In: Proceedings of the 2001 International Conference on Info-tech and Info-net (ICII2001). Vol. 6. pp. 35 – 41.
- Davies, J., Fensel, D., Van Harmelen, F., 2003. *Ontology-driven Knowledge Management - Towards the Semantic Web*. John Wiley & Sons.
- Davies, S., Donaher, C., Jan 2011. Making the semantic web usable: interface principles to empower the layperson. *Journal of Digital Information* 12 (1).
- de Vries, M., Kapff, L., Achiaga, M., Wauters, P., 2011. Pricing of Public Sector Information Study. Tech. rep., The European Commission.
URL http://ec.europa.eu/information_society/policy/psi/docs/pdfs/report/11_2012/summary.pdf
- Dolata, U., 2011. The music industry and the internet: a decade of disruptive and uncontrolled sectoral change. *Research Contributions to Organizational Sociology and Innovation Studies Discussion Paper*.
- Donker, F., 2009. Public Sector Geo Web Services: Which Business Model Will Pay for a Free Lunch? In: *SDI Convergence Research, Emerging Trends, and Critical Assessment*. Netherlands Geodetic Commission, pp. 35–52.
URL <http://www.gsdi.org/gsdi11/papers/pdf/143.pdf>
- Dou, W., 2004. Will internet users pay for online content? *Journal of Advertising Research* 44 (04), 349–359.
- Dubosson-Torbay, M., Pigneur, Y., Usunier, J., 2005. Business models for music distribution after the p2p revolution. In: *Proceedings of the Fourth International Conference on Web Delivering of Music, 2004. WEDELMUSIC 2004*. pp. 172–179.
- Eppler, M., Algesheimer, R., Dimpfel, M., 2003. Quality criteria of content-driven websites and their influence on customer satisfaction and loyalty: An empirical test of an information quality framework. In: *Proceedings of the Eighth International Conference on Information Quality (ICIQ-03)*.
- Farrell, S., 2009. API Keys to the Kingdom. *IEEE Internet Computing* 13 (5), 91–93.
- Feigenbaum, L., Herman, I., Jan 2007. The Semantic Web in Action. *Scientific American* 297 (6), 90–97.
- Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P., Berners-Lee, T., 1999. Hypertext transfer protocol–HTTP/1.1. Tech. rep., RFC.

- Flanagin, A., Metzger, M., 2008. The credibility of volunteered geographic information. *GeoJournal* 72 (3), 137–148.
- Flores, N. E., Carson, R. T., 1997. The relationship between the income elasticities of demand and willingness to pay. *Journal of Environmental Economics and Management* 333 (3), 287–295.
- Fraser, C., 1996. On the provision of excludable public goods. *Journal of Public Economics* 60 (1), 111–130.
- Gallaughar, J., Auger, P., BarNir, A., 2001. Revenue streams and digital content providers: an empirical investigation. *Information & Management* 38 (7), 473–485.
- Gaustard, T., Aug 2002. The problem of excludability for media and entertainment products in new electronic market channels. *Electronic Markets* 12 (4), 248–251.
- Giff, G., van Loenen, B., Zevenbergen, J., 2008. PSGI Policies in Norway and England: Are they within the Spirit of Recent EU Directives. *International Journal of Spatial Data Infrastructures Research* 3, 118–145.
- Giles, J., 2005. Internet encyclopaedias go head to head. *Nature* 438 (7070), 900–901.
- Gillespie, A., 2007. *Foundations of Economics*. Oxford University Press, USA.
- Goodchild, M., 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal* 69 (4), 211–221.
- Goodchild, M., 2012. A review of “map of a nation: A biography of the ordnance survey”. *Annals of the Association of American Geographers* 102 (1), 244–245.
- Guel, F. L., Rochelandet, F., 2006. The Willingness To Pay For Online Music In The Presence Of Copying: An Empirical Investigation. Tech. rep., ADIS working paper, Université de Paris-Sud.
- Gunaratne, S., 2010. Demise of newspapers and the rise of cyberspace. *Asia Pacific Media Educator* (20), 33–36.
- Haklay, M., 2010. How good is volunteered geographical information? a comparative study of openstreetmap and ordnance survey datasets. *Environment and planning. B, Planning & design* 37 (4), 682.
- Haklay, M., Basiouka, S., Antoniou, V., Ather, A., 2010. How many volunteers does it take to map an area well? the validity of linus’ law to volunteered geographic information. *Cartographic Journal, The* 47 (4), 315–322.
- Halb, W., Raimond, Y., 2008. Building linked data for both humans and machines. In: *Proceedings of the WWW2008 Workshop on Linked Data on the Web*.

- Hammer-Lahav, E., 2010. The oauth 1.0 protocol. Tech Report, Internet Engineering Task Force (IETF).
- Harris, K., 2010. Selling and Building Linked Data: Drive Value and Gain Momentum. In: Linking Enterprise Data. Springer, pp. 65–76.
- Hart, G., Dolbear, C., 2006. So what’s so special about spatial? Terra Cognita workshop.
URL http://www.sange.fi/~humis/iswc2006-cd/Workshops/Terra_Cognita_-_Geospatial_Semantic_Web/Hart.pdf
- Hausenblas, M., 2009. Exploiting linked data to build web applications. Internet Computing, IEEE 13 (4), 68–73.
- He, B., Patel, M., Zhang, Z., Chang, K., 2007. Accessing the deep web. Communications of the ACM 50 (5), 94–101.
- 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), 1–136.
- Heath, T., Goodwin, J., 2011. Linking geographical data for government and consumer applications. In: Wood, D. (Ed.), Linking Government Data. Springer New York, pp. 73–92, 10.1007/978-1-4614-1767-5_4.
- Heimer, M., 2011. The theory of access replacing ownership on the example of spotify. Master’s thesis, Kalmar University.
- Heipke, C., 2010. Crowdsourcing geospatial data. ISPRS Journal of Photogrammetry and Remote Sensing 65 (6), 550–557.
- Hendler, J., 2010. Web 3.0: The Dawn of Semantic Search. Computer 43 (1), 77 – 80.
- Hendler, J., Golbeck, J., 2008. Metcalfe’s law, Web 2.0, and the Semantic Web. Web Semantics: Science, Services and Agents on the World Wide Web 6 (1), 14–20.
- Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P., Rudolph, S., 2009. OWL 2 Web Ontology Language Primer. Recommendation, W3C, <http://www.w3.org/TR/2009/REC-owl2-primer-20091027/>.
- Hougaard, J., Tvede, M., 2010. Selling digital music: business models for public goods. Netnomics 11 (1), 85–102.
- Hyland, B., 2010. Preparing for a Linked Data Enterprise. In: Linking Enterprise Data. Springer, pp. 51–64.
- Ihlström, C., Palmer, J., 2001. Revenues for online newspapers: owner and user perceptions. Electronic Markets 12 (4), 228–236.

- Kahin, B., Varian, H., 2000. *Internet Publishing and Beyond: The Economics of Digital Information and Intellectual Property*. A Publication of the Harvard Information Infrastructure Project. Mit Press.
- Kanliang, W., 2004. Pricing strategies for information products: a review. In: *Proceedings of the IEEE International Conference on E-Commerce Technology for Dynamic E-Business, 2004*. pp. 345–348.
- Kind, H.J., N. T., Sorgard, L., 2009. Business Models for Media Firms: Does Competition Matter for How They Raise Revenue? *Marketing Science* 28, 1112–1128.
- Kirchhoff, S. M., 2009. The US newspaper industry in transition. Tech. rep., Washington, DC: Congressional Research Service, http://digitalcommons.ilr.cornell.edu/key_workplace/634 ”.
- KK Breitman and M A Casanova and W Tuskowski, 2007. *Semantic Web: Concepts, Technologies and Applications*. Springer Verlag.
- Knight, S., Burn, J., 2005. Developing a framework for assessing information quality on the world wide web. *Informing Science: International Journal of an Emerging Transdiscipline* 8, 159–172.
- Kobilarov, G., Scott, T., Raimond, Y., Oliver, S., Sizemore, C., Smethurst, M., Bizer, C., Lee, R., 2009. Media meets semantic web—how the bbc uses dbpedia and linked data to make connections. *The Semantic Web: Research and Applications*, 723–737.
- Kowatsch, T., Maass, W., 2009. The use of free and paid digital product reviews on mobile devices in in-store purchase situations. In: *The 4th Mediterranean Conference on Information Systems (MCIS 09) Athens, Greece*.
- Kreitz, G., Niemela, F., 2010. Spotify—Large Scale, Low Latency, P2P Music-on-Demand Streaming. *IEEE Tenth International Conference on Peer-to-Peer Computing (P2P)*.
- Lassila, O., Hendler, J., 2007. Embracing ”Web 3.0”. *IEEE Internet Computing* 11 (3), 90–93.
- Latif, A., Saeed, A. U., Hoefler, P., Stocker, A., Wagner, C., 2009. The Linked Data Value Chain: A Lightweight Model for Business Engineers. In: *Proceedings of the 5th International Conference on Semantic Systems (I-SEMANTICS ’09)*. pp. 568–575.
- Li, X., 2006. *Internet newspapers: the making of a mainstream medium*. Routledge.
- Lin, A., 2005. Understanding the market for digital music. *Stanford Undergraduate Research Journal* 4, 50–53.
- Linde, F., 2009. Pricing information goods. *Proceedings of Scholarship in the New Information* 18 (5), 379–384.

- Longhorn, R., Blakemore, M., 2007. *Geographic Information: Value, Pricing, Production, and Consumption*. CRC.
- Longley, P., Goodchild, M., Maguire, D., Rhind, D., 2005. *Geographical information systems and science*. Wiley.
- Lopes, A., Galletta, D., Oct 2006. Consumer perceptions and willingness to pay for intrinsically motivated online content. *Journal of Management Information Systems* 23 (2), 203–231.
- Love, J., 1995. Pricing government information. *Journal of Government Information* 22 (5), 363–387.
- Lowe, J., 2005. A Geospatial Semantic Web. *Geospatial Solution* 15 (6).
- Ltd, K. N., 2009. *Film Market Market Review 2009*. No. ISBN 978-1-84729-471-5. Key Note Ltd.
- Luczak-Roesch, M., 2009. Linked data authoring for non-experts. In: *Proceedings of the WWW2009 Workshop on Linked Data on the Web*, Madrid, Spain.
- Lytras, M., Garcia, R., 2008. Semantic Web applications: a framework for industry and business exploitation—What is needed for the adoption of the Semantic Web from the market and industry. *International Journal of Knowledge and Learning* 4 (1), 93–108.
- Macdonald, E., Sharp, B., 2000. Brand Awareness Effects on Consumer Decision Making for a Common, Repeat Purchase Product: A Replication. *Journal of Business Research* 48 (1), 5–15.
- Masinter, L., Berners-Lee, T., Fielding, R., 2006. Uniform resource identifier (uri): Generic syntax. Tech Report RFC 3986, Internet Engineering Task Force, <http://www.ietf.org/rfc/rfc2396.txt>.
- McCole, P., Ramsey, E., Williams, J., 2010. Trust considerations on attitudes towards online purchasing: The moderating effect of privacy and security concerns. *Journal of Business Research* 63 (9), 1018–1024.
- McNutt, P., 1999. *Public goods and club goods*. Edward Elgar Publishing/University of Ghent.
- Meeks, W., Dasgupta, S., 2004. Geospatial Information Utility: An Estimation of the Relevance of Geospatial Information to Users. *Decision Support Systems* 38 (1), 47–63.
- Melnik, M. I., Alm, J., 2003. Reputation, information signals, and willingness to pay for heterogeneous goods in online auctions. *Southern Economic Journal* 72 (2), 305–328.
- Mendes, P., Mühleisen, H., 2012. Sieve: Linked data quality assessment and fusion. *Proceeding of the 2nd International Workshop on Linked Web Data Management (LWDM 2012)*, Berlin.

- Meyer, A., 2010. *Public Goods, Private Goods: The Merging in Global Space*. Rosedog Press.
- Milhollin, K., 2012. Dynamic semantic publishing for news organizations, http://semanticweb.com/dynamic-semantic-publishing-for-news-organizations_b29439.
- Miller, E., Manola, F., 2004. *RDF Primer*. Recommendation, W3C, <http://www.w3.org/TR/rdf-primer/>.
- Miller, P., Styles, R., Heath, T., 2008. Open data commons, a license for open data. *Proceedings of the 1st Workshop about Linked Data on the Web (LDOW2008)* 369.
- Mitchell, I., Wilson, M., March 2012. *White Paper: Linked data - Connecting and exploiting big data*. White Paper, Fujitsu.
- Mooney, P., Corcoran, P., Winstanley, A. C., 2010. Towards quality metrics for openstreetmap. In: *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*. GIS '10. ACM, New York, NY, USA, pp. 514–517.
- Morales-Arroyo, M., Sharma, R., 2009. Deriving value in digital media networks. *International Journal of Computer Science and Security (IJCSS)* 3 (2), 126.
- Naumann, F., Rolker, C., 2000. *Assessment Methods for Information Quality Criteria*. Professoren des Inst. für Informatik.
- Newbery, D., Bently, L., Pollock, R., 2008. *Models of Public Sector Information Provision Via Trading Funds*. Study Report, Department for Business, Enterprise and Regulatory Reform (BERR) and HM Treasury.
URL <http://www.berr.gov.uk/files/file45136.pdf>
- Novak, T., Hoffman, D., 2001. *Profitability on the web: Business models and revenue streams*. Owen Graduate School of Management, Vanderbilt University, eLab Position Paper, January.
- Oestreicher-Singer, G., Zalmanson, L., 2009. Paying for Content or Paying for Community? The Effect of Social Involvement on Subscribing to Media Web Sites. In: *Proceedings of the IEEE International Conference on Computer and Information Science*, China.
- OFT, 2006. *The Commercial Use of Public Information*. Study Report OFT861, The Office of Fair Trading.
URL http://www.offt.gov.uk/shared_offt/reports/consumer_protection/oft861.pdf
- Oh, H., 2000. The effect of brand class, brand awareness, and price on customer value and behavioral intentions. *Journal of Hospitality Tourism Research* 24 (2), 136–162.

- Omitola, T., Zuo, L., Gutteridge, C., Millard, I., Glaser, H., Gibbins, N., Shadbolt, N., 2011. Tracing the provenance of linked data using void. In: Proceedings of The International Conference on Web Intelligence, Mining and Semantics (WIMS'11).
- OPSI, 2009. United Kingdom Report on the Re-Use of Public Sector Information 2009. Report, The Office of Public Sector Information.
URL <http://www.opsi.gov.uk/advice/psi-regulations/uk-report-reuse-psi-2009.pdf>
- Ordnance Survey, 2011. Derived Data.
URL <http://www.ordnancesurvey.co.uk/oswebsite/aboutus/foi/questions/docs/PanGovtAg.pdf>
- OXERA, 1999. The Economic Contribution of Ordnance Survey GB. Report, The Oxford Economic Research Associates Ltd.
URL <http://www.ordnancesurvey.co.uk/aboutus/reports/oxera/oxera.pdf>
- Peuquet, D., 2002. Representations of Space and Time. Guilford Press.
- Picard, R., Jan 2000. Changing Business Models of Online Content Services - Their Implications for Multimedia and Other Content Producers. The International Journal on Media Management 2 (11), 60–68.
- PIRAInternational, 2000. Commercial Exploitation of Europe's Public Sector Information. Study Report, European Commission, Directorate General for the Information Society.
URL http://www.epsiplus.net/content/download/23804/314839/version/1/file/media_672+full+report.pdf
- Pollock, R., 2008. The Economics of Public Sector Information. Study Report, University of Cambridge.
- Price, G., 2001. The Invisible Web: Uncovering Information Sources Search Engines Can't See. Information Today.
- Raghu, T., Sinha, R., Vinze, A., Burton, O., 2009. Willingness to pay in an open source software environment. Information Systems Research 20 (2), 218–236.
- Raimond, Y., Scott, T., Oliver, S., Sinclair, P., Smethurst, M., 2010. Use of semantic web technologies on the bbc web sites. Linking Enterprise Data, 263–283.
- Rajala, R., Nissilä, J., Westerlund, M., 2007. Revenue models in the open source software business. Handbook of Research on Open Source Software: Technological, Economic, and, 541.
- Rappa, M. A., 2004. The utility business model and the future of computing services. IBM Systems Journal 43 (1), 32–42.

- Rizk, A., Streitz, N., André, J., 1990. Hypertext: Concepts, systems and applications. In: Proceedings of the First European Conference on Hypertext, INRIA, France. Vol. 5. Cambridge University Press.
- Sansone, S., Rocca-Serra, P., Field, D., Maguire, E., Taylor, C., Hofmann, O., Fang, H., Neumann, S., Tong, W., Amaral-Zettler, L., et al., 2012. Toward interoperable bioscience data. *Nature genetics* 44 (2), 121–126.
- Scharrenbach, T., Bischof, S., Fleischli, S., Weibel, R., 2012. Linked raster data. In: Seventh International Conference on Geographic Information Science.
- Servant, F.-P., 2008. Linking Enterprise Data. In: Proceedings of the 1st Workshop on Linked Data on the Web (LDOW2008), 17th International World Wide Web Conference (WWW2008). Vol. 369.
- Shadbolt, N., Hall, W., Berners-Lee, T., 2006. The Semantic Web revisited. *IEEE Intelligent Systems* 21 (3), 96–101.
- Shapiro, C., Varian, H., Jan 1999. *Information rules: A Strategic Guide to the Network Economy*. Harvard Business Press.
- Sheridan, J., Tennison, J., 2010. Linking UK government data. In: Proceedings of the WWW2010 Workshop on Linked Data on the Web, USA.
- Shuen, A., 2008. *Web 2.0: A Strategy Guide Business thinking and strategies behind successful Web 2.0 implementations*. O'Reilly Media, Inc.
- Sliwinski, A., 2004. Toward Perceived Value-based Pricing of Geographic Information Services. In: Proceedings of the 7th AGILE Conference on Geographic Information Science. pp. 541–549.
- Stahl, F., Schafer, M.-F., Maass, W., 2004. Strategies for selling paid content on newspaper and magazine web sites: an empirical analysis of bundling and splitting of news and magazine articles. *The International Journal on Media Management*, 6 (1 & 2), 59–66.
- Stallman, R., Gay, J., Lessig, L., 2002. *Free software, free society: selected essays of richard m. stallman*.
- Strong, D., 1996. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems* 12 (4), 5–33.
- Sylvie, G., 2008. *Developing an online newspaper business model: Long distance meets the long tail*. University of Texas at Austin 24.
- Teece, D., 2010. Business models, business strategy and innovation. *Long Range Planning* 43 (2-3), 172–194.
- The Stationery Office, 2012. *Open data white paper - unleashing the potential*. White Paper CM8353, TSO.

- Thurman, N., Herbert, J., 2007a. Newspapers' e-business models: A survey of attitudes and practice at UK news websites. Paper presented to the International Symposium on Online Journalism, Austin, Texas.
- Thurman, N., Herbert, J., 2007b. Paid content strategies for news websites: An empirical study of British newspapers' online business models. *Journalism Practice* 1 (2), 208–226.
- Troncy, R., 2010. Bringing the iptc news architecture into the semantic web. In: *Proceedings of the Seventh International Semantic Web Conference*. Springer-Verlag, pp. 483–498.
- Turnor, R., 2007. Is the print version of the sunday times in terminal decline and set to be replaced by the internet?
URL <http://webjournalist.com.au/sundaytimesresearch.pdf>
- Usery, E., Varanka, D., 2011. Design and development of linked data from The National Map. *Semantic Web* 3 (4), 371 – 384.
- Varian, H. R., 2000. *Internet Publishing and Beyond: The Economics of Digital Information and Intellectual Property*. Palgrave Macmillan.
- Vedamuthu, A., Orchard, D., Hirsch, F., Hondo, M., Yendluri, P., Boubez, T., Yalçinalp, U., 2007. Web services policy 1.5-framework. W3C Recommendation 4.
- Veglis, A., 2004. New production models for newspaper organizations. *WSEAS Transactions on Communications* (1), 218–222.
- Wang, C., Ye, L., Zhang, Y., Nguyen, D., 2005. Subscription to fee-based online services: What makes consumer pay for online content. *Journal of Electronic Commerce Research* 6 (4), 304–311.
- Wang, H., Chin, A., 2011. Social Influence on Being a Pay User in Freemium-based Social Networks. *IEEE International Conference on Advanced Information Networking and Applications (AINA)*, 2011, 526 – 533.
- Weiss, P., 2004. Borders in cyberspace: Conflicting government information policies and their economic impact. Study Report, European Public Sector Information (PSI) Platform.
URL <http://www.epsiplus.net/content/download/568/3839/file/Borders%20in%20Cyberspace.pdf>
- Wiercinski, J., Mason, J., 2010. Music in the digital age: Downloading, streaming and digital lending. *CAML Review* 38 (1), 5–16.
- Wright, A., 2008. Searching the deep web. *Communications of the ACM* 51 (10).
- Ye, L., Zhang, Y., Nguyen, D., Chiu, J., 2004. Fee-based online services: Exploring consumers' willingness to pay. *Journal of International Technology and Information Management* 13 (2), 133–141.

- Yu, L., 2011. Linked Open Data. In: A Developer's Guide to the Semantic Web. Springer Berlin Heidelberg, Berlin, Heidelberg, Ch. 11, pp. 409–465.
- Zeithaml, V., 1988. Consumer perceptions of price, quality, and value: a means-end model and synthesis of evidence. *The Journal of Marketing*, 2–22.
- Zwemer, S., Xu, A., Pang, C., Aguirre, N., von Weisberg, M., 2010. From print to portal: Pricing strategies in the online news realm.
URL <http://albertsun.info/misc/PricingStrategiesOnlineNews-MKTG288-Penn.pdf>