The Data Divide: A Society Divided by Open Data

Peter G. West
Web and Internet Science Research Group, Electronics and Computer Science
University of Southampton
Southampton, UK
pw9g09@ecs.soton.ac.uk

ABSTRACT

Open Data has been positioned as a way to create accountability and transparency, to empower people to make more informed choices, and to engage individuals in greater civic participation. Governments and industries have made available a deluge of datasets. Subsequently, interest has risen in measuring and validating the effects that this investment in Open Data has had on society. However, a number of challenges have made it difficult for citizens to make effective use of this data. The vast majority of the potential consumer base of Open Data does not have the necessary expertise to effectively manipulate, interrogate, or interpret the datasets in their particular forms.

In this report, we investigate several factors pertaining to difficulties that individuals have in using Open Data. Such difficulties have caused the majority of individuals to become reliant on a data-capable, data-literate minority who have the skills and tools necessary to interrogate it. We then propose potential avenues to mitigate and reduce this "Data Divide", including learning from the difficulties experienced thus far, taking a multidisciplinary approach to Open Data, and lowering technical barriers.

Categories and Subject Descriptors

H.1.2 [**Models and Principles**]: User/Machine Systems—*Human factors*; H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Data sharing*

General Terms

Human Factor, Design, Standardization, Verification

Keywords

Open Data, Government Data, Digital Divide

1. INTRODUCTION

The recent pushes towards releasing public data as Open Data have resulted in governments, scientists and businesses releasing vast quantities of datasets for free [7]. There have been strong efforts to ensure data is released as machine-readable linked-data to form a *Web of Data* [10, 1], but little attention has been paid to how humans may consume it. Halford et al. suggest that without the involvement of users of data during these early stages, the web for them may become *less* transparent and usable than it is now [23]. However, these users rarely possess the required knowledge to engage with the necessary technologies. They risk being neglected by those developing the Web of Data [21]. Without prompt action, current focuses on achieving technically-perfect data could be leading towards a "Data Divide", where society is divided into those who have the knowledge to use data in their everyday lives, and those who do not.

In this article, we will consider how data has been crucial to the progress of a diverse range of disciplines, and how the involvement of these disciplines in the construction of the Web of Data is of pressing importance. The roots of Open Data were not out of a need for computation – the Victorians that founded social statistics were from varied disciplines: mathematics, philosophy, and medicine [32]. Today, the list of disciplines that need data is all-encompassing. Data is changing society – it is the "raw material of the 21st century" leading to transparency and democratisation [7, 6]. It is a valuable resource for everyone. Traffic data helps people avoid congestion, government data allows efficient allocation of tax payers' funds, and data from research institutes offers necessary materials for scientists to conduct research.

We will discuss how the strong focus on technical approaches could be making this Data Divide a reality. It is an echo of the Digital Divide; in modern developed society, it is assumed that individuals understand modern technologies. For example, a newspaper may display a URL, even though a minority of individuals have no access to the web, nor have ever used it [25]. Already, this minority of individuals are unable to access data through digital means.

In concluding, three suggestions will be made as to how the Data Divide could be mitigated. First, through understanding the Digital Divide, parallels may be drawn such that we can understand the causes and how we may prevent them. Second, a multidisciplinary approach to Open Data should be taken to allow different disciplines to become involved in the development of the nascent Web of Data. This could be facilitated through the new discipline of Web Science [27]. Finally, we discuss how lowering the technical barrier is an effective way to engage those without technological expertise. This has been demonstrated though the *Lowercase Semantic Web* – a set of techniques that makes it easier to publish structured data [20].

2. BACKGROUND

In this section, we briefly discuss how government data has become so important to modern society of the developed

world. We will start by investigating the roots of statistics and how this led to the collection of data to make improvements in healthcare and medicine. We will then give an overview of more recent movements, including efforts to make huge quantities of government data available to the public as Open Data.

2.1 Roots of Statistics in Society

The use of government data has roots in early 19th century England. During this time, rapid industrialisation was leading to vast social changes. Huge divisions in classes led to slums, where overcrowding fostered endemic disease [32]. Physicians, angered with the conditions that the urban poor had to live in, had strong motivation to better understand social conditions. However, there was little relevant information available. The government did not have records of how many poor people received relief or even how much money was in circulation. There were only rough estimates of the population [32].

It was Belgian mathematician Adolphe Quetelet who, at Cambridge in 1833, presented his statistical work that highlighted the need for a statistical society. Quetelet analysed data on French crime rates, revealing an association between demographics and crime [32]. He established that approximate crime rates were predictable with regard to sex, age, education, climate and season. Quetelet, who is now widely regarded as the founder of modern social statistics, later formed Belgium's Central Statistical Bureau, influencing other countries to make similar efforts [15].

Quetelet's work interested a number of famous figures in England, including Charles Babbage, and in 1834, the Royal Statistical Society was formed. The Society would go on to urge the government to form the General Register Office - Europe's first office dedicated to producing demographic records, including births, marriages and deaths. Today it is recognised as one of the key elements that addressed social unrest in Victorian England [28].

2.2 Nightingale's Healthcare Reforms

Quetele's efforts in statistics influenced Florence Nightingale's work during the Crimean War that led to huge reforms in military hospitals [15]. By recording and analysing detailed records of sick patients during the war, Nightingale identified that for every soldier that died from injury, 7 died from preventable disease [33].

During this time, government understanding of statistics was poor. However, through effective graphical presentation of the data – of which Nightingale became a pioneer of – Nightingale was able to encourage the government to establish subcommissions to carry out reforms that hospitals were in dire need of [33, 15]. Nightingale's work has lead to the collection of detailed statistics that workers in healthcare are now dependent on.

2.3 1854 London Cholera Outbreak

The use of statistics in healthcare has since been vital. One famous example is Dr John Snow's use of mortality data during the 1854 cholera epidemic in Soho, London. Snow, having studied previous incidents of cholera outbreaks, believed

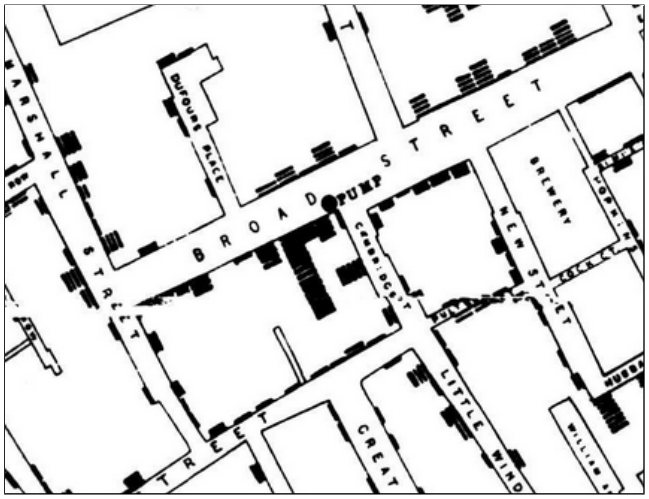

Figure 1: John Snow's spot map of cholera cases in Soho, London, illustrating a correlation of number of deaths to proximity to the Broad Street Pump.

that cholera was transmitted by drinking water, and not, as was more commonly believed, via atmospheric causes. The General Register Office was able to provide Snow with the locations of deaths. This, in conjunction with door-to-door surveys, led Snow to identify that the majority of people who died drank from one pump: Broad Street Pump [8, 40].

Snow had identified that the water from this pump was provided by a water company that sourced water from a heavily polluted area of the Thames (and later discovered that sewage was leaking into the pump). By plotting the deaths on a spot map (as shown in Figure 1), Snow was able to effectively illustrate the relationship between mortality and proximity to the pump. This subsequently convinced the council to remove the pump handle, rendering it unusable. Today, his efforts are famous and are credited as a major contribution towards fighting cholera [14].

A number of observations can be made from Snow's work and later interpretations. First, it was through use of government data – records on deaths from the General Register Office and records of water companies – that Snow was able to identify the locations of deaths, where these individuals drank from and ultimately the source of cholera within the area [14]. Without this data, it would have been impossible for Snow to have identified the correlations that led to his findings.

Secondly, before collection of this data, Snow had already worked with cholera outbreaks and had a hypothesis: cholera was spread by polluted drinking water. His spot map showed a correlation of deaths in close proximity to the Broad Street pump – a correlation that helped confirm Snow's hypothesis [8]. However, in refusal to accept the "germ theory" that Snow proposed, the government used this spot map to claim that it was more likely that cholera was spread by atmospheric causes, where overfilled cesspits were the primary cause [14]. This highlights the importance of fully understanding data and its context – data can be misleading to those without the necessary knowledge to understand

it. The government subsequently replaced the handle [34].

2.4 The Web and Open Data

Early work such as Snow's has strengthened the clear need for the availability of government data. It has allowed a better understanding of crime and has helped revolutionise medicine. Through a better understanding of society, it has helped form the modern world [38].

Throughout the 20th century, techniques for collecting data have advanced considerably [38, 41]. Advances in statistical methods allow a better understanding of data, with greater observations being able to be made [41]. In the last two decades, with the advent of the World Wide Web, data is now able to be shared and accessed quickly and easily on an international scale. These factors have led to data being described as "the raw material of the 21st century" [7], with its use ubiquitous in a wide variety of disciplines [12].

More recently, governments, businesses and scientists have been encouraged to release data on the web as "Open Data" – data that is released under an open license and may be used without restrictions [4]. Spearheaded by Tim Berners-Lee, this movement is argued as a way to boost the economy, increase transparency and democratisation, improve public sector innovation, and make life easier [7, 6].

Governments have led the way [6, 29] in releasing Open Data, with the UK government, for example, releasing over 9000 datasets on data.gov.uk[1]. More recently, the UK government funded *Open Data Institute* has been established, which promises to find ways to use data effectively [42, 7]. In the United States, Barack Obama, on his first day in office as President, announced that his administration would "establish a system of transparency" and "openness will strengthen our democracy and promote efficiency and effectiveness in government" [35]. More governments will implement Open Data strategies, releasing more data and opening up other opportunities of Open Data, including, for example, its use in law enforcement [29].

3. FUTURE OF OPEN DATA

In this section, we will consider where the future of Open Data lies. In particular, we will discuss the importance to big data and the difficulties it presents. It will then be outlined how providing data as linked data can give data context, making data more intuitive to understand and navigate.

3.1 Big Data

Data collection has become so easy and ubiquitous that the amount of data collected for particular purposes can get so large it becomes difficult to handle [16]. However, these datasets have huge potentials, as they can be combined to answer extremely niche questions. The uses for this facility are numerous. For example, a drug study may want to find 54-year-old women with high blood pressure and who dropped out of school [11].

The uses of big data have already been well demonstrated.

[1]UK government Data portal - `http://www.data.gov.uk` (Accessed: 3 May 2013)

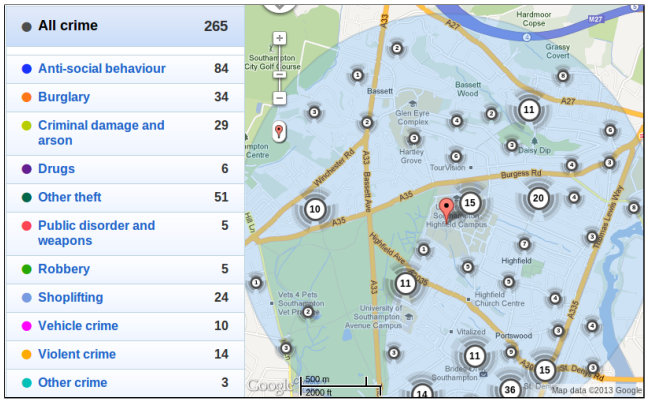

Figure 2: Police.uk use data data about crimes to show where crimes have been committed in the last month. This service may be used throughout the UK.

In reducing crime, huge quantities of UK crime data have been condensed and plotted on a map to let citizens understand where crime has been committed[2] (see Figure 2). In the United States, police have used, to some success, statistical analysis of huge crime datasets to predict where and when crime will take place [5]. In the context of environmental issues, energy data can be used to increase awareness of changes that will reduce environmental damage. For example, data generated by thousands of electricity usage meters was used to calculate that as much of 5% of electricity in the US is consumed by DVRs while they are on standby [39]. A simple modification to existing DVRs could resolve this.

Combining such heterogeneous datasets (data from different sources with different data representations) is a challenging task. Datasets may use different vocabularies, where entities are represented in different ways. Furthermore, the quality of data is difficult to establish, especially as anyone can publish to the Web [11]. To address these issues, there has been a strong focus on trying to find technological solutions. These solutions have been focused in the area of linked data, where data can be interlinked with semantic relationships [10].

3.2 Linked Data

With the ever-growing body of data, researchers are working towards linking resources together to form a *Web of Data* [10]. This involves the publishing of structured data using the Resource Description Framework (RDF) and connecting entities in datasets from different sources via hyperlinks [10, 9]. If multiple entities point to a common entity, then that entity can be used to link those data sources. The primary unit of the web will become real world things, such as people, buildings and events, in which each thing has a Uniform Resource Identifier (URI) [9, 23].

The practical uses of Linked Data for publishing Open Data have been well demonstrated with DBpedia, a structured dataset formed from Wikipedia [4]. It links in other datasets

[2]Police.uk crime maps – `http://www.police.uk/crime` (Accessed: 9 May 2013)

Figure 3: DBpedia allows sophisticated querying against Wikipedia data.

to data extracted from Wikipedia articles to allow sophisticated queries to be made, as shown in Figure 3. Such a resource is valuable to tools that require access to specific data – for example, looking up the history of a city by a user's location. Furthermore, Auer et al. argue that DBpedia could be used to identify disparity between datasets – for example, mismatching population figures [4] – such that data can be made more reliable and consistent.

Interlinking data is not without difficulty – the nature of heterogeneous data makes it difficult to identify where entities are the same, and therefore where datasets may be linked. Without these mappings being manually identified, machines are unable to reliably understand the meaning of datasets. For example, Doan et al. used a machine learning approach to match ontologies, which accurately matched up to 66-97% of nodes [17]. To combat this, Tim Berners-Lee has encouraged data to be published as Open Linked Data [10]. The five-stars of Open Data[3] are a suggested metric for determining how "open" data is, from data presented in propriety forms to data in an open, linked form.

4. PROBLEMS WITH OPEN DATA

In this section, we identify and discuss the challenges and opportunities in Open Data. First, we discuss the usability challenges of data representation. In particular, we consider how raw structured data is useful to machines, but inherently difficult for humans to understand. Second, we identify problems pertaining to data representations, including how data ambiguity leads to multiple interpretations. Here, we also discuss philosophical challenges of data, such as ontology design. Without careful design, data representations may be difficult or impossible to use. Finally, we consider the social factors involved in Open Data, including how understanding data requires specialised knowledge from other disciplines. Without this knowledge, data may be misinterpreted with disastrous consequences.

[3]5 Stars of Linked Data, Is your Linked Open Data 5 Star?, W3C – `http://www.w3.org/DesignIssues/LinkedData.html` (Accessed: 18 April 2013)

Table 1: How the Open Data rating affects the non-technical usability [23]

Rating	Data characteristics	Usability
⋆	under an open license	⋆ ⋆
⋆ ⋆	structured format (e.g., Excel)	⋆ ⋆
⋆ ⋆ ⋆	non-proprietary format	⋆ ⋆ ⋆
⋆ ⋆ ⋆ ⋆	uses URIs to identify things	⋆
⋆ ⋆ ⋆ ⋆ ⋆	links to other data for context	⋆

4.1 Data Usefulness vs. Usability

Linked Data adds context to data, creating a resource that can have sophisticated queries made against it [4, 22]. However, data published as linked data is notably unprepared for human consumption [22, 23, 20, 26]. Halford et al. suggests that Open Data is increasingly mediated by technical structures (such as RDF, OWL and ontologies), and that 4- and 5-star data is particularly difficult for non-technical use [23]. This is illustrated in Table 1.

For such data to be understood, users must have the skills to work with the data directly or be able to use tools that can interpret the data. However, it has been argued that few possess the skills required to engage with development [21]. Furthermore, tools used to interpret data are built by the few who *are* skilled enough to interact with data, mostly by organisations and institutions who can afford the costs of working with such technologies [30].

Huynh et al. argue that authors of niche data – such as hobbyists – are particularly important to consider [30]. The importance of such diversity can be demonstrated by considering The Long Tail – products in low demand can collectively rival the market of the relatively few high demand products (see Figure 4) [3]. Huynh et al. argue that too many Semantic Web projects are targeted at information domains with huge popularity or size, and too little at small information domains. Niche data authors cannot commit the time, cost and effort to learning linked data technologies, and so are unable to publish data suitable for the Web of Data [30].

It has further been demonstrated that users of data have difficulty with tasks that involve using multiple information sources, such as planning an event using shopping and maps websites [43]. The Digital Divide means that some people are unable to even access these tools [25]. Van Kleek et al. state that this modern requirement to use multiple data sources to make informed decision demonstrates a clear need for data integration.

4.2 Data Representation

The diversity of disciplines and disagreement on how data should be presented leads to heterogeneous data [2, 37]. Datasets about the same subject may have different ways to express structure, properties and values. Research has shown that heterogeneity between datasets is extremely common [43, 19, 4, 17, 37]. Reconciling these datasets has been a challenge particularly within database research, where techniques have been developed to address this problem [2]. However, it has been observed that mixing heterogeneous

subjects has contingent and unpredictable outcomes [23].

Establishing data representations requires computational expertise, so efforts in this area are governed by technical experts [23]. However, Halford et al. state that designing data structures is not simply a technical issue. In order to address the problem of heterogeneity, the development must involve other disciplines to form standards and practices for representing data [37].

One challenge within linked data is how to indicate when two data representations are about the same entity. In this case, there is multiple URIs for the same thing [19]. This problem is exacerbated by a lack of standards in URI design between data providers [43]. One proposed solution is *owl:sameAs*, which pairs URIs that identify the same entity [24]. However, Halpin et al. argues that philosophically, *owl:sameAs* is used incorrectly. It is used too commonly in place of where a weaker relationship should be used. A study showed that as few as 51% of *owl:sameAs* operations are used correctly [24].

A further challenge within linked data is how to link structured data with unstructured data. Unstructured data has insufficient semantic information to make meaningful links [22]. This puts a dependency on data-authors to offer their data as structured data, rather than as formatted documents or proprietary formats that they have already released their data as [30, 23].

4.3 Social Factors

For data to be used effectively, it must be understood by those working with it. Specialised data should be handled by specialists in that discipline. Otherwise, there is risk that data may be misinterpreted, particularly when made accessible to unspecialised individuals, such as through visualisation to demonstrate correlations. In medicine, Brody et al. states that where visual correlations are presented, "the resulting claims about disease 'hot spots' may create unjustified worry and detract attention from solid but more visually appealing lines of research" [14].

One example that has been extensively studied is the correlation between drinking and wealth. It is commonly theorised that drinking has a harmful effect on the economy due to related health conditions. However, Peters and Stringham describe the outcome of a study which shows a positive correlation between drinking and wealth [36]. They hypothesise that through drinking, individuals enhance their social capital – social skills and number of acquaintances – leading to higher earning. However, an opposite hypothesis to this is that people with social capital drink less [45]. These conflicting hypotheses are formed from the same correlations – it is unclear if there is a relationship at all.

Raw data in particular may be misleading. Without an understanding of the particular discipline, data may be open to misinterpretation. In 2013, a Leeds hospital closed its heart surgery due to a startlingly high mortality rate in the data. However, this data had yet to be analysed, and decisions were made based on data before it was fully understood. The hospital was subsequently reopened [13]. Cases such as these must be considered by data-users, such that such events are prevented in the future.

5. NEXT STEPS

In this section, we describe three steps that should be taken to prevent the Data Divide. First, parallels should be drawn with the Digital Divide to understand the causes and how it may be prevented. In particular, we consider current efforts that are going into bridging the Digital Divide, and how such efforts may be used within Open Data. Secondly, it is suggested that a multidisciplinary approach to Open Data be taken. This involves encouraging other disciplines than computer science to get involved with the development of technology for the Web of Data. Finally, the technical barrier should be lowered, such that the learning curve for becoming involved in publishing quality Open Data is smaller, reducing the cost, effort and time required to do so.

5.1 Learn from The Digital Divide

Throughout the last two decades, computers have become hugely pervasive. Almost 80% of people in North America have access to the internet[4] and over 40% of US citizens own a smartphone[5]. This is leading to individuals becoming depended on these technologies, and governments, companies and media expecting individuals to have access to it – for instance, the phrase "see our website for details" is found on huge numbers of products and media. However, there is a small minority of individuals who do not have access to such technologies. There is a growing digital divide between the individuals who have access to these technologies and those who don't [25, 44].

Those who can use digital technologies have a significant advantage in modern society [44]. These technologies form an increasingly important part of modern society, ranging from how we communicate to how we can gather information. Government services are online, most TV broadcasting is now done digitally, and an increasing number of people depend on social networks to communicate with one another [44]. With these aspects, it becomes increasingly difficult for someone to cross the divide to data-literate.

Current work to reduce this divide is focused in the area of educating those who cannot access these technologies. These individuals are mostly the older generation, who have not been exposed to digital technologies so pervasively [25]. Without careful consideration, there is a risk that Open Data could lead to an increase in this divide, or lead to a "Data Divide". Thus, it is crucial that causes of the Digital Divide are well understood to ensure that the Data Divide can be mitigated [25].

5.2 Take a Multidisciplinary Approach

For the Web of Data to be built for use by multiple disciplines, it should involve these disciplines in the early stages

[4]World Internet Usage And Population Statistics for June 30, 2012 – `http://www.internetworldstats.com/stats.htm` (Accessed: 9 May 2013)

[5]comScore, Smartphone Usage, March 2013 – `http://www.comscore.com/Insights/Press_Releases/2013/5/comScore_Reports_March_2013_U.S._Smartphone_Subscriber_Market_Share` (Accessed: 9 May 2013)

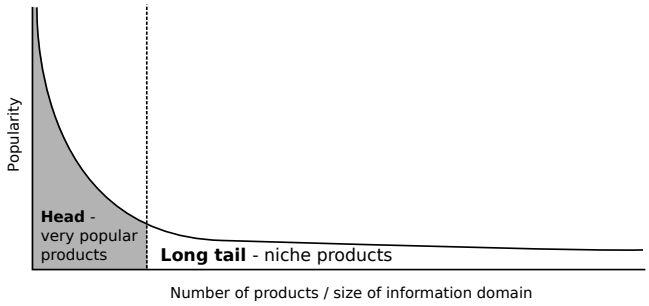

Figure 4: The Long Tail – products in low demand can collectively rival the market of the relatively few high demand products.

of development. Reichman et al. argue that the discipline of Ecology should be involved in the development of the nascent semantic web [37]. This involvement has also been considered a necessity by other disciplines, including Sociology [23], Medicine [18] and Geography [21].

Without an approach that involves other disciplines, there is a strong risk of the Web of Data becoming a single tool that tries to do everything. If we consider the Long Tail (see Figure 4), then efforts will primarily be in large information domains by organisations who can afford this work [3]. It is crucial then that other disciplines become involved to build tools for the smaller information domains – for instance, all types of leaves. In this way, instead of one large tool, there will be a large constellation of tools, each a product of collaboration with a specialised purpose.

This call for a multidisciplinary approach is not a new one – social sciences have worked with technological experts for data analysis since 1960s [38]. However, the rapid changes with data practices necessitate the involvement of other disciplines if they are to take advantage of Open Data [23]. Brodie argues that a major challenge within big data integration is multi-disciplinary [11].

Web Science is a new discipline to foster this multidisciplinary approach, aiming to identify and nurture the common goals between disciplines [27]. This will involve combining the technical aspects of the web with other disciplines such as Law, Economics and Sociology. Through this, Web Science aims to identify the limitations of the web such that it may be developed to become more suited to specialist areas, encouraging the Web's growth [27].

5.3 Lower the Technical Barrier

For data to be linked effectively, it must be machine-readable structured information [23, 4, 10]. To move towards a Web of Data, there is a strong push to throw away existing web publishing practices, in favour of techniques that generate machine-readable structured information [20]. However, this is a large commitment for both companies and individuals, who must invest time, effort and resources into such technologies. Subsequently, there have been calls to lower the technical barrier, such that less technological expertise is required to become involved in the semantic web [26, 37, 23].

```
<div xmlns:v="http://rdf.data-vocabulary.org/#" typeof="v:
     Person">
  <div property="v:name">Peter West</div>,
  <div property="v:title">undergraduate</div> at
  <div property="v:affiliation">University of Southampton</div>
  <a href="http://www.peter-west.co.uk" rel="v:url">www.peter-
       west.co.uk</a>
</div>
```

Figure 5: How a person can be represented using HTML5's Microdata. These Semantic annotations do not affect the visual display of the HTML, but provide a way for structured data to be extracted.

Technologies that lower this technical barrier focus on prioritising human-readability over machine-readability. These are often classed as part of the *Lowercase Semantic Web* [20, 31], and include Microformats, Microdata and RDFa. By allowing the mark-up of structured data within existing information on the web, these technologies reduce the amount of knowledge required to author Linked Data and do not require the author to generate supplementary structured data. An example of using Microdata to form structured information about a person is shown in Figure 5.3. This simple and concise nature has led to their widespread adoption by individuals and companies, particularly amongst blogs [20].

One problem that has been identified with the Semantic Web is that the outcomes of such investments are not immediately obvious [30]. This particularly pertains to individuals who maintain small websites, where producing RDF does not directly affect them. Exhibit[30] is one approach to generating RDF for existing website through providing structured browsing, such that static websites may become faceted. It has been demonstrated as an easy-to-install software package that gives websites the capabilities of structured browsing – an incentive for individuals to use it to generate structured data [30].

The usability of authoring Semantic Web is paramount. Lowercase technologies and Exhibit both demonstrate methods that are more usable to those who author and consume data, reducing the level of data-literacy required to interact with the Web of Data. The Lowercase Semantic Web is a crucial stage to building up the nascent Semantic Web.

6. CONCLUSION

Open Data is paving the way to creating accountability and transparency. It is giving people information that allows them to make more informed choices and better engage in civic participation. There are hundreds of thousands of datasets that have been made available by governments and industries. With this investment, interest has risen in measuring the effects that Open Data has on society.

In this report, we have reviewed the events that have led up to this stage of Open Data and have considered its future. In doing so, we have identified a number of problems which must be addressed for Open Data to be sustainable in the future and within other disciplines. Namely, we have identified that a vast majority of potential Open Data consumers do not have the necessary skills to manipulate, interrogate

or interpret datasets in the form that they are available as.

These problems stem from a focus on finding "technical perfection" to the Web of Data. Current technological solutions represent data in forms that are difficult to understood by humans. Furthermore, tools to access these datasets are too few, too difficult or too specific. It has been proposed that the technological skills required to access data, either in its raw form or through tools, is creating a Data Divide, between those who can access and use data, and those who can't.

Three stages to mitigating this divide have been proposed. First, by drawing parallels with the Digital Divide, we can better understand how such a social divide comes about, such that we can reduce the effect. Second, by taking a multidisciplinary approach to Open Data, the design of data representations and tools may better cater to niche information domains. Finally, by lowering the technical barrier, data consumption may become accessible to more people.

Pursuing these avenues could ensure that the future of Open Data is more sustainable. In turn, this will demonstrate its value to data publishers, encouraging the publication of further data, and better conformation to standards and practices. It will make Open Data more accessible to the masses, helping people make better informed decisions and better understand governments and organisations. The next stages in Open Data will reveal its use as a pervasive tool in everyday life, but also as a way to divide society.

7. ACKNOWLEDGMENTS

I'd like to thank my supervisor, Dr. Max Van Kleek, for his guidance throughout this research. His insights have been very valuable during this project. I'd also like thank my colleagues, in particular Bahman Asadi and Jim Skinner for thoughtful discussions related to this research.

8. REFERENCES

[1] K. Aberer, P. Cudré-Mauroux, and M. Hauswirth. The chatty web: emergent semantics through gossiping. In *Proceedings of the 12th international conference on World Wide Web*, WWW '03, pages 197–206, New York, NY, USA, 2003. ACM. ISBN 1-58113-680-3. doi: 10.1145/775152.775180. URL http://doi.acm.org/10.1145/775152.775180.

[2] S. Abiteboul, B. Peter, and D. Suciu. *Data on the Web: From Relations to Semistructured Data and Xml*. Data Management Systems Series. Morgan Kaufmann Publishers, 2000. ISBN 9781558606227.

[3] C. Anderson. *The Long Tail: Why the Future of Business is Selling Less of More*. Hyperion Books. Hyperion, 2008. ISBN 9781401309664.

[4] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. Dbpedia: a nucleus for a web of open data. In *Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference*, ISWC'07/ASWC'07, pages 722–735, Berlin, Heidelberg, 2007. Springer-Verlag. ISBN 3-540-76297-3, 978-3-540-76297-3. URL http://dl.acm.org/citation.cfm?id=1785162.1785216.

[5] BBC. The age of big data. http://www.bbc.co.uk/programmes/b01rt4c7, BBC Horizon, 4 April 2013. Accessed: 4 May 2013.

[6] D. Beer. Power through the algorithm? participatory web cultures and the technological unconscious. *New Media & Society*, 11(6):985–1002, 2009. doi: 10.1177/1461444809336551.

[7] T. Berners-Lee and N. Shadbolt. There's gold to be mined from all our data. http://www.thetimes.co.uk/tto/opinion/columnists/article3272618.ece, The Times, London, 31 December 2011. Accessed: 2 April 2013.

[8] P. Bingham, N. Q. Verlander, M. J. Cheal, J. Snow, and W. Farr. John Snow, William Farr and the 1849 outbreak of cholera that affected London: a reworking of the data highlights the importance of the water supply. *Public Health*, 118(6):387–394, Sep 2004.

[9] C. Bizer, T. Heath, D. Ayers, and Y. Raimond. Interlinking open data on the web. In *Demonstrations Track, 4th European Semantic Web Conference, Innsbruck, Austria*, 2007.

[10] C. Bizer, T. Heath, K. Idehen, and T. Berners-Lee. Linked data on the web (ldow2008). In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 1265–1266, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-085-2. doi: 10.1145/1367497.1367760. URL http://doi.acm.org/10.1145/1367497.1367760.

[11] C. Bizer, P. Boncz, M. L. Brodie, and O. Erling. The meaningful use of big data: four perspectives – four challenges. *SIGMOD Rec.*, 40(4):56–60, Jan. 2012. ISSN 0163-5808. doi: 10.1145/2094114.2094129.

[12] C. L. Borgman. Data, disciplines, and scholarly publishing. *Learned Publishing*, 21(1):29–38, 2008. doi: 10.1087/095315108X254476. URL http://www.ingentaconnect.com/content/alpsp/lp/2008/00000021/00000001/art00005.

[13] S. Boseley. Children's heart surgery at leeds general set to resume next week. http://www.guardian.co.uk/society/2013/apr/05/childrens-heart-surgery-leeds-general, The Guardian, 5 April 2013. Accessed: 11 April 2013.

[14] H. Brody, M. R. Rip, P. Vinten-Johansen, N. Paneth, S. Rachman, and J. Snow. Map-making and myth-making in Broad Street: the London cholera epidemic, 1854. *Lancet (London, England)*, 356(9223): 64–68, Jul 2000.

[15] I. B. Cohen. Florence nightingale. *Scientific American*, 250(3):128–137, 1984.

[16] K. Cukier. Data, data everywhere. http://www.economist.com/node/15557443, The Economist, 25 February 2010. Accessed: 4 May 2013.

[17] A. Doan, J. Madhavan, P. Domingos, and A. Halevy. Ontology matching: A machine learning approach. In *Handbook on Ontologies in Information Systems*, pages 397–416. Springer, 2003.

[18] G. Eysenbach. Medicine 2.0: social networking, collaboration, participation, apomediation, and openness. *J. Med. Internet Res.*, 10(3):e22, 2008.

[19] H. Glaser, I. Millard, and A. Jaffri. Rkbexplorer.com:a knowledge driven infrastructure for linked data providers. In *European Semantic Web Conference*, volume 5021/2, pages 797–801. Springer, June 2008.

URL http://eprints.soton.ac.uk/265152/. Event Dates: 1-5 June 2008.

[20] A. Graf. Rdfa vs. microformats. *Digital Enterprise Research Institute, Innsbruck*, 2007.

[21] S. D. Graham. Software-sorted geographies. *Progress in Human Geography*, 29(5):562–580, 2005. doi: 10.1191/0309132505ph568oa. URL http://phg.sagepub.com/content/29/5/562.abstract.

[22] A. Halevy, O. Etzioni, A. Doan, Z. Ives, J. Madhavan, L. McDowell, and I. Tatarinov. Crossing the structure chasm. In *Proceedings of the First Biennial Conference on Innovative Data Systems Research (CIDR)*, 2003.

[23] S. Halford, C. Pope, and M. Weal. Digital futures? sociological challenges and opportunities in the emergent semantic web. *Sociology*, 47(1):173–189, 2013. doi: 10.1177/0038038512453798. URL http://soc.sagepub.com/content/47/1/173.abstract.

[24] H. Halpin, P. Hayes, J. McCusker, D. McGuinness, and H. Thompson. When owl:sameas isn't the same: An analysis of identity in linked data. In P. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J. Pan, I. Horrocks, and B. Glimm, editors, *The Semantic Web – ISWC 2010*, volume 6496 of *Lecture Notes in Computer Science*, pages 305–320. Springer Berlin Heidelberg, 2010. ISBN 978-3-642-17745-3. doi: 10.1007/978-3-642-17746-0_20.

[25] E. Hargittai. Second-level digital divide: Differences in people's online skills. *First Monday*, 7(4), 2002. ISSN 13960466. URL http://firstmonday.org/ojs/index.php/fm/article/view/942.

[26] S. Haustein and J. Pleumann. Is participation in the semantic web too difficult? In *Proceedings of the First International Semantic Web Conference on The Semantic Web*, ISWC '02, pages 448–453, London, UK, UK, 2002. Springer-Verlag. ISBN 3-540-43760-6. URL http://dl.acm.org/citation.cfm?id=646996.711277.

[27] J. Hendler, N. Shadbolt, W. Hall, T. Berners-Lee, and D. Weitzner. Web science: an interdisciplinary approach to understanding the web. *Communications of the ACM*, 51(7):60–69, July 2008. ISSN 0001-0782. doi: 10.1145/1364782.1364798. URL http://doi.acm.org/10.1145/1364782.1364798.

[28] I. D. Hill. Statistical society of london–royal statistical society: The first 100 years: 1834-1934. *Journal of the Royal Statistical Society. Series A (General)*, 147(2): pp. 130–139, 1984. ISSN 00359238. URL http://www.jstor.org/stable/2981670.

[29] N. Huijboom and T. Van den Broek. Open data: an international comparison of strategies. *European journal of ePractice*, 12(1):1–13, 2011.

[30] D. F. Huynh, D. R. Karger, and R. C. Miller. Exhibit: lightweight structured data publishing. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 737–746, New York, NY, USA, 2007. ACM, ACM. ISBN 978-1-59593-654-7. doi: 10.1145/1242572.1242672. URL http://doi.acm.org/10.1145/1242572.1242672.

[31] R. Khare. Microformats: the next (small) thing on the semantic web? *Internet Computing, IEEE*, 10(1): 68–75, 2006. ISSN 1089-7801. doi: 10.1109/MIC.2006.13.

[32] E. Magnello. Eminent victorians and early statistical societies. *Significance*, 6(2):86–88, 2009. ISSN 1740-9713. doi: 10.1111/j.1740-9713.2009.00357.x. URL http://dx.doi.org/10.1111/j.1740-9713.2009.00357.x.

[33] L. McDonald. Florence nightingale and the early origins of evidence-based nursing. *Evidence Based Nursing*, 4(3):68–69, 2001. doi: 10.1136/ebn.4.3.68. URL http://ebn.bmj.com/content/4/3/68.short.

[34] S. Newsom. Pioneers in infection control: John snow, henry whitehead, the broad street pump, and the beginnings of geographical epidemiology. *Journal of Hospital Infection*, 64(3):210 – 216, 2006. ISSN 0195-6701. doi: 10.1016/j.jhin.2006.05.020. URL http://www.sciencedirect.com/science/article/pii/S0195670106002830.

[35] B. Obama. Transparency and open government. http://www.whitehouse.gov/the_press_office/TransparencyandOpenGovernment, The White House, January 2009. Accessed: 18 April 2013.

[36] B. Peters and E. Stringham. No booze? you may lose: Why drinkers earn more money than nondrinkers. *Journal of Labor Research*, 27(3):411–421, 2006. ISSN 0195-3613. doi: 10.1007/s12122-006-1031-y.

[37] O. J. Reichman, M. B. Jones, and M. P. Schildhauer. Challenges and opportunities of open data in ecology. *Science*, 331(6018):703–705, 2011. doi: 10.1126/science.1197962. URL http://www.sciencemag.org/content/331/6018/703.abstract.

[38] M. Savage. *Identities and Social Change in Britain Since 1940: The Politics of Method*. OUP Oxford, 2010. ISBN 9780191582936.

[39] R. Smolan. How 'big data' is changing lives. http://www.bbc.co.uk/news/technology-21535739, BBC News - What If?, 26 February 2013. Accessed: 4 May 2013.

[40] J. Snow. *On the mode of communication of cholera*. John Churchill, 1855.

[41] R. Steel, J. Torrie, and D. Dickey. *Principles and procedures of statistics: a biometrical approach*. McGraw-Hill series in probability and statistics. McGraw-Hill, 1997. ISBN 9780070610286.

[42] E. Thwaites. Uk leads demand for open data as odi opens for business. http://www.theodi.org/news/uk-leads-demand-open-data-odi-opens-business, The ODI, 4 December 2012. Accessed: 11 April 2013.

[43] M. Van Kleek, D. Alexander Smith, H. S. Packer, J. Skinner, and N. R. Shadbolt. Carpé data: Supporting serendipitous data integration in personal information management. *ACM SIGCHI Conference on Human Factors in Computing Systems (CHI2013)*, 2013. URL http://eprints.soton.ac.uk/347760/.

[44] S. Vie. Digital divide 2.0: Generation m and online social networking sites in the composition classroom. *Computers and Composition*, 25(1):9 – 23, 2008. ISSN 8755-4615. doi: 10.1016/j.compcom.2007.09.004. URL http://www.sciencedirect.com/science/article/pii/S8755461507000989. Media Convergence.

[45] E. R. Weitzman and I. Kawachi. Giving means receiving: the protective effect of social capital on binge drinking on college campuses. *Am J Public Health*, 90(12):1936–1939, Dec 2000.