International Journal of Information Management 37 (2017) 150–154

Contents lists available at ScienceDirect

International Journal of Information Management

journal homepage: www.elsevier.com/locate/ijinfomgt

Research Note

Open data: Quality over quantity

Shazia Sadiq [a], Marta Indulska [b,*]

[a] School of Information Technology and Electrical Engineering, The University of Queensland, St Lucia, QLD 4072, Australia
[b] UQ Business School, The University of Queensland, St Lucia, QLD 4072, Australia

ARTICLE INFO

Article history:
Received 22 December 2016
Accepted 8 January 2017

Keywords:
Open data
Data quality

ABSTRACT

Open data aims to unlock the innovation potential of businesses, governments, and entrepreneurs, yet it also harbours significant challenges for its effective use. While numerous innovation successes exist that are based on the open data paradigm, there is uncertainty over the data quality of such datasets. This data quality uncertainty is a threat to the value that can be generated from such data. Data quality has been studied extensively over many decades and many approaches to data quality management have been proposed. However, these approaches are typically based on datasets internal to organizations, with known metadata, and domain knowledge of the data semantics. Open data, on the other hand, are often unfamiliar to the user and may lack metadata. The aim of this research note is to outline the challenges in dealing with data quality of open datasets, and to set an agenda for future research to address this risk to deriving value from open data investments.

1. Introduction

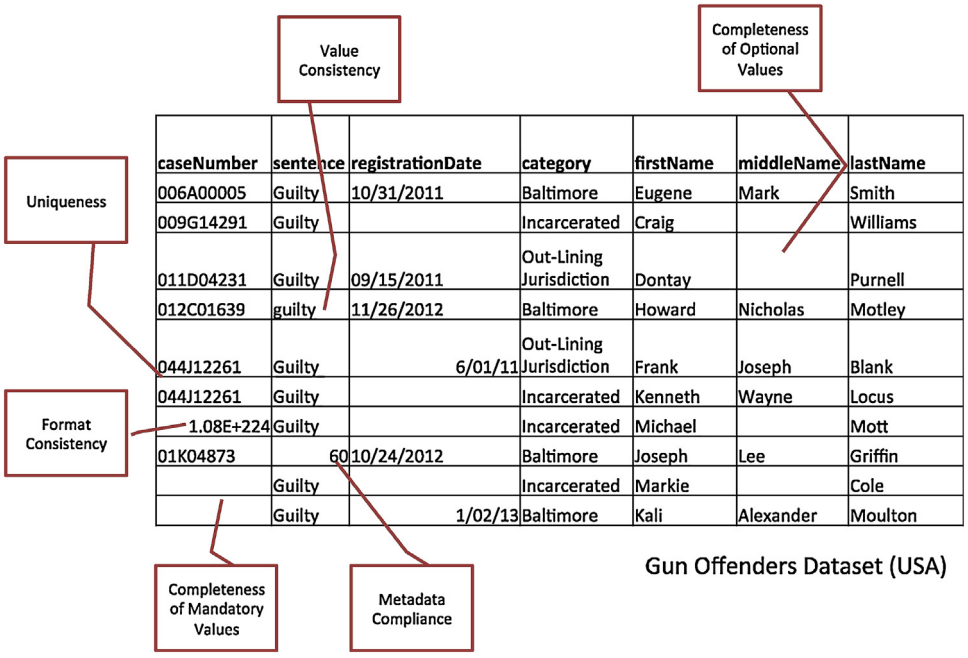
Open data is data made freely available by governments, organizations, researchers, among others, for use by anyone without copyright restrictions. The growth of the open data movement has been a very significant one. Our review of open data availability indicates that the Australian government open data portal has grown by over 900% between 2013 and 2015 in terms of the number of available datasets (from 573 in December 2013 to 5767 in December 2015). Similarly, other governments have open data initiatives, with New Zealand having over 3800, UK over 23,000, USA over 194,000, and Canada over 240,000 datasets available on their open data government portals (see respective data.gov). These statistics do not include the large numbers of organizational and other datasets made available by non-government sources e.g. GeoNames, Wikidata, DBPedia, to name a few.

The proliferation of publicly available datasets (Duus & Cooray, 2016) and emergence of data markets (Elbaz, 2012) presents an unprecedented opportunity to governments, business and entrepreneurs to harness the power of data for economic, social and scientific gains. Open data is envisaged to form the basis of innovation and there are indications that data-driven innovation is estimated to have added $67 billion in new value to the Australian economy, that is, 4.4 percent of Australian GDP (PWC, 2014). A number of competitions (hackathons), such as the Australian GovHack or the annual Open Data Day in the USA, have been introduced to mobilize public interest and generate a culture of innovation through data towards economic and societal gains.

While open data competitions have given rise to some success stories in terms of start-ups and apps, well documented in published case studies on government open data portals e.g. (Queensland Government, 2016), there is also some evidence that the time-to-value from these datasets remains prohibitively long primarily due to lack of knowledge on the quality characteristics of the data and resulting effort of making the data ready for use (Belkin & Patil, 2016). At the same time, the metadata, as well as the underlying data quality for these datasets, is known to be deficient. For example, many open datasets have duplicate, inconsistent, and missing data and generally lack easily accessible schema descriptions, e.g. the MusicBranz.org open dataset consists of 324 schema-less CSV files with a data volume of 35.1GB. An analysis of open datasets (Zhang, Jayawardene, Indulska, Sadiq, & Zhou, 2014) indicates that many such problems exist in open data. For example, in public transport data, the data consistency of bus stop names is low, which may have serious implications for use of the data that requires grouping or search on bus stop names, such as timetabling and traffic monitoring. Similarly, see Fig. 1 as an example, where several data quality problems can be identified in the USA Gun Offenders Database.

* Corresponding author.
E-mail addresses: shazia@itee.uq.edu.au (S. Sadiq), m.indulska@business.uq.edu.au (M. Indulska).

http://dx.doi.org/10.1016/j.ijinfomgt.2017.01.003
0268-4012/



caseNumber	sentence	registrationDate	category	firstName	middleName	lastName
006A00005	Guilty	10/31/2011	Baltimore	Eugene	Mark	Smith
009G14291	Guilty		Incarcerated	Craig		Williams
011D04231	Guilty	09/15/2011	Out-Lining Jurisdiction	Dontay		Purnell
012C01639	guilty	11/26/2012	Baltimore	Howard	Nicholas	Motley
044J12261	Guilty	6/01/11	Out-Lining Jurisdiction	Frank	Joseph	Blank
044J12261	Guilty		Incarcerated	Kenneth	Wayne	Locus
1.08E+224	Guilty		Incarcerated	Michael		Mott
01K04873	60	10/24/2012	Baltimore	Joseph	Lee	Griffin
	Guilty		Incarcerated	Markie		Cole
	Guilty	1/02/13	Baltimore	Kali	Alexander	Moulton

Fig. 1. Data quality problems identified in an open dataset.

We note that the value of the dataset is inescapably tied to the underlying quality of the data (Johnston & Carrico, 1988; O'Reilly, 1982). Although for open data, value and quality may be correlated, they are conceptually different (Abiteboul et al., 2015). For example, a complete and accurate list of names of all countries in Asia may not have much value. Whereas incomplete and noisy GPS data from public transport vehicles may have a high perceived value for transport engineers and urban planners. When dealing with such large and unknown datasets, a user might experience long query processing times, only to realize that the results obtained are of poor quality. Alternatively, the user may not realise the data is of inadequate quality, thus affecting any subsequent decisions made, based on the query result (Yeganeh, Sadiq, & Sharaf, 2014).

Despite such issues, there is an increasing tendency to gather significant volume of external and internal data into so-called data lakes (Stamford, 2014), which are typically described as enterprise data management platforms for storing, curating and analysing data that comes from a number of disparate sources, including open data sources. Although there is a heightened interest in the big data phenomenon, lessons learnt from years of research in information system use have shown that the assumption of 'more use is better' is clearly not the case (Seddon, 1997). As the growth in the amount of open datasets and sources continues at an exponential rate, it is leaving data consumers with a massive footprint of unexplored, unfamiliar datasets that may or may not generate valuable insights for them. Organizations are thus starting to face the 'dark data' (Tittel, 2014) syndrome where a large proportion of their information assets are under-utilized. With,out scientifically credible knowledge that provides the ability to efficiently evaluate the underlying quality characteristics of the data, there is a significant risk of organizations and governments accumulating large volumes of low value density data (Curry, 2010), falling into analytical traps (Silver, 2012) and/or investing in low ROI data products (Belkin & Patil, 2016).

On February 8th, 2015, a group of global thought leaders from the database research community outlined some grand challenges in getting value from big data (Abiteboul et al., 2015). The key message was the need to develop the capacity to *'understand how the quality of data affects the quality of the insight we derive from it'*. Given that the social, economic and scientific benefits that justify the global investments into open datasets are still in their infancy there is a need for in-depth understanding of the 'quality-to-use' dynamics of open data.

In this paper we first outline the state-of-art in evaluating data quality and highlight challenges in applying these techniques for evaluating the quality of datasets that exhibit characteristics typical in the open data space. We then reflect on how these challenges undermine the ability to generate value from open data use and present an agenda for future research to enable requisite understanding of the 'quality-to-use' dynamics of open data.

2. Evaluating data quality

Data quality has been studied widely by both the research and practitioner community (Sadiq, 2013). Data quality dimensions (Jayawardene, Sadiq, & Indulska, 2013) such as accuracy, completeness, consistency, are a fundamental notion in the definition and measurement of data quality. Evaluating the quality of a data set is a fundamental task in most, if not all, data quality management projects (Batini, Cappiello, Francalanci, & Maurino, 2009). Quality of data is typically evaluated against a certain stated requirement (English, 2009; ISO, 2011; Loshin, 2001). The last 20 years of data quality research (Sadiq, Yeganeh, & Indulska, 2011) have been based on this fundamental principle of fitness for use (Juran et al., 1974). Thus existing methodologies for data quality management are inevitably top-down (McGilvray, 2008; Redman & Blanton, 1997), wherein data quality requirements are determined in a top-down manner following well-understood usage requirements and enforced using good data governance practices.

Batini et al. (2009), provide a comprehensive analysis of existing approaches for data quality assessment and requirements identification, indicating that such approaches typically include three core aspects, viz. data and process analysis, data quality requirements analysis, and, data quality analysis. Data and process analysis includes examination of data schemas, performing interviews, and meetings with data users to reach a complete understanding of