In order to fully realise the value of open data researchers must first address the quality of the datasets

blogs.lse.ac.uk/impactofsocialsciences/2017/04/26/in-order-to-fully-realise-the-value-of-open-data-researchers-must-first-address-the-quality-of-the-datasets/

4/26/2017

There has been a phenomenal increase in the availability of data over the last decade. Open data is provided as a means of empowering users with information and in the hope of sparking innovation and increased efficiency in governments and businesses. However, in spite of the many success stories based on the open data paradigm, concerns remain over the quality of such datasets. ***Marta Indulska*** *and* ***Shazia Sadiq*** *argue that in order to facilitate more effective and efficient realisation of value from open data, research must reach a shared consensus on the definition of data quality dimensions, provide methods and guidelines for assessing the potential usefulness of open datasets using exploratory tools and techniques, and develop rigorous theoretical underpinnings on effective use of open data.*

The Open Data movement has seen a phenomenal increase in the availability of data since then President Obama signed the Memorandum on Transparency and Open Government almost a decade ago, giving birth to Data.gov – a government portal of freely available data. Governments across the world have jumped on board and created similar Open Data portals. Our review of open data availability indicates that the Australian government open data portal has seen extraordinary growth between 2013 and 2017 in terms of the number of available datasets (from 573 in December 2013 to over 49,000 in April 2017). This is a similar sized portal to the UK's Open Data portal (over 42,000 datasets), yet still relatively small given USA's portal has over 192,000 datasets, and Canada's has over 118,000. These portals represent only a subset of the available open data, with many organisations also offering up some of their data openly.

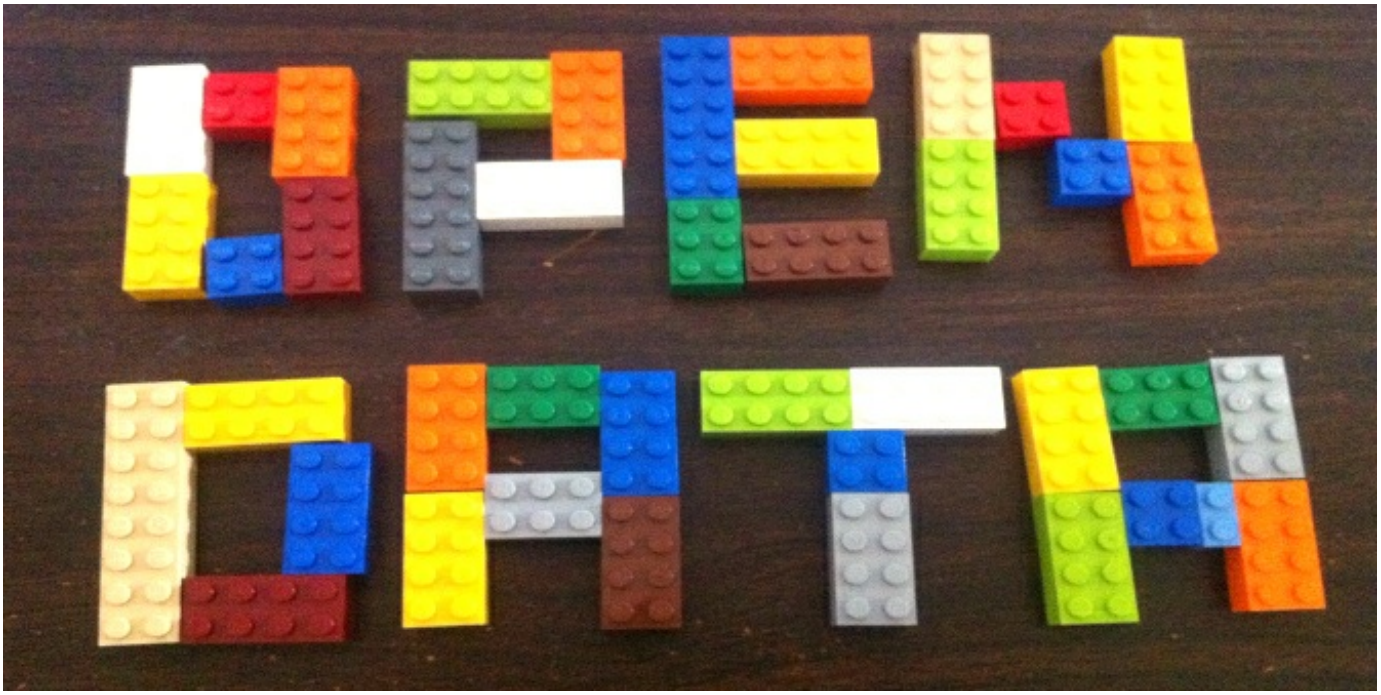

Image credit: open data (lego) by justgrimes. This work is licensed under aCC BY-SA 2.0 license.

So what is open data and what is driving its growth? Put simply, open data is data – organisational, government or individual, for example – that is made publicly available for anyone's use or reuse. It is provided in the interest of empowering users with information and in the hope of sparking innovation and increased efficiency. Indeed, there are many success stories that stem from open data. For example, releasing transport data in London made Citymapper possible, now a global company that continues to create jobs. Similarly, Landsat is reported to save the US government, NGOs and the private sector over US$350 million per year by conservative estimates. Such individual success stories aggregate to big impact of open data. An earlier study indicates that by 2020, open data will be responsible for a €1.7 billion reduction in public administration costs in the EU. Further, it is estimated that "effective use" of open data could reduce power consumption by 16%, as well as reducing the time we all spend in traffic jams.

The unprecedented opportunities that open data presents for governments and businesses have resulted in events (hackathons – such as GovHack or Open Data Day) designed to mobilise public interest in open data and spark data-driven innovation. While there have been many successes, the realisation of value from open data is constrained by our lack of knowledge about its data quality and is dependent on "effective use", which implies an understanding of not only quality but the schema and generative process of the data. Yet, metadata – a description of the available data – is often missing, incomplete, or difficult to access on open data portals, making data quality assessment a resource intensive, and therefore expensive, exercise. Our recent analysis of open datasets indicates that many datasets have issues that can be roadblocks to value generation. For example, in public transport data, if the data consistency of bus stop names is low, this may have serious implications for use of the data for timetabling and traffic monitoring. The old adage of "garbage in, garbage out" applies to open data, yet data consumers may not be aware of data quality problems or small nuances between how the data was initially collected versus its newly planned use. This presents a serious challenge that not only limits value realisation but may also result in serious negative consequences if poor data is used as a basis for decision making.

To facilitate more effective and efficient realisation of value from open data, in our recent publication we argue that research needs to proceed in three directions.

First, given the diverse community of creators, custodians, curators and consumers of data, we need a shared consensus on the definition of data quality dimensions to facilitate a more coordinated effort to increasing data quality within the open data community. Currently, there is significant variation in the interpretation of data quality dimensions, for example accuracy can mean accuracy with respect to reality or accuracy to a unit of measurement. A "one world" view of open data quality would assist with benchmarking data available to consumers and would also allow open data curators to more effectively manage and communicate data quality of their datasets.

Second, we see the need for bottom-up data exploration approaches to empower consumers to quickly assess the usefulness of a dataset for their specific purpose. While data exploration has been researched for over the last decade, there is still a lack of methods and guidelines for assessing the potential usefulness of open datasets using exploratory tools and techniques.

Third, we see a critical need for studying the contexts and factors of successful open data use. Currently there are no rigorously developed theoretical underpinnings on effective use of open data. Such studies will create much needed guidance for open data use initiatives, inform policy on data release, and accelerate the time to value generation from open data.

This blog post is based on the authors' article, "Open data: Quality over quantity", published in the International Journal of Information Management (DOI: 10.1016/j.ijinfomgt.2017.01.003).

Note: This article gives the views of the authors, and not the position of the LSE Impact Blog, nor of the London School of Economics. Please review our comments policy if you have any concerns on posting a comment below.

About the authors

Marta Indulska *is Associate Professor and Leader of the Business Information Systems discipline at the UQ Business School, University of Queensland, Australia. Her main research interests include conceptual modelling, business process management and open innovation.*

Shazia Sadiq *is a Professor in the Data and Knowledge Engineering (DKE) research group at the School of Information Technology and Electrical Engineering at the University of Queensland, Australia. Her main research interests are innovative solutions for business information systems that span several areas including business process management, governance, risk and compliance, and information quality and use.*