1.

The state of open data and open data research

François van Schalkwyk & Stefaan G Verhulst

Open government data, and the attendant excitement over its potential, emerged as an asset for social good just under a decade ago. It rose to prominence on the back of related trends and developments, including the rise of big data, the arrival of new analytical methods to derive insights and innovations from that data, and deteriorating trust in public institutions that are the custodians of large datasets related to the functioning of government and the allocation of public resources. In addition, the relative success of open source and open innovation provided new models on how to create public value. The Obama administration's move to increase access to government data (in particular, its launch of the data.gov site) also played a part in increasing the visibility and the legitimacy of open data.

Eight years after the launch of that site, open data has entered the mainstream of both policy and activism. Around the world, in both developed and developing countries, at the national and local levels, governments have created or are planning open data programmes and portals. Open data projects are playing an increasingly important role in economic and social development, spurring progress in areas as varied as healthcare, education, banking, agriculture, climate change and innovation. A growing list of private companies, whose businesses have hitherto depended on *private* data, are also coming to recognise the potential competitive and social benefits of opening up that data; and we are witnessing the emergence of social enterprises that rely on open data to provide tools and services for the public good.

So where do we stand now? And where do we go from here? This introductory chapter outlines some reflections on current developments in the field, and considers how they may affect the state of open data and open data research in the years to come. It describes a wide variety of trends – some positive, some more cautionary. If there is one overarching message, it is that for all the excitement

and hype, there is still much that we don't know about the contributions of open data to social and economic development.

The theoretical potential of open data has been established; but much work remains to be done, many challenges need to be overcome, and several gaps in our understanding must be breached if open data is, in fact, to help solve complex social problems and improve people's lives.

One of the purposes of this volume is, in fact, to begin that process of filling in the gaps in our knowledge. Each of the nine chapters published in this volume, in its own away, adds to our existing and steadily growing understanding of how open data works. Through these contributions, we see the importance of social dynamics – be they institutional or otherwise – across the value chain of open data. It is important to remember that each of these examples represents a specific instance, in a specific setting. But it is slowly, through individual examples like these, that our overall understanding of the real impact of open data will advance.

Current trends and their implications for open data

Rise of populism and regime change

Donald Trump's rise to power and, more generally, the emergence of nationalist strongmen with limited faith in democracy around the world, is likely to affect the perceived value proposition and use of open data. Two aspects of Trump-style governance will have a particular impact: a penchant for secretive deal-making, and the debasement of knowledge, facts and evidence both in governance and in public discourse.

These trends and others have already led some to highlight the value of open data as a force for accountability and transparency, and, more generally, as a tool for the 'resistance'. (This trend is evident, for instance, in increased interest in the storage and archiving of existing government data.) Paradoxically, however, we believe that this heightened interest may prove counter-productive to the spread of open data as it elevates only one value proposition (i.e. transparency) above other, potentially less controversial or difficult value propositions such as increased innovation and economic growth. Similarly, if open data comes to be equivalent in the public mind simply with archiving government data, then its potentially much greater value as a tool for real-time decision-making may be overlooked or ignored.

Transparency and accountability are of course valuable and crucial goals. However, many years of research and practice has repeatedly indicated that governments are more likely to create open data projects if they believe it will also spur economic growth, improve the efficiency of public service delivery and lead to innovation. It is therefore essential to keep highlighting these value propositions, making clear the full range of benefits that can potentially be conferred by open data – beyond making governments accountable.

The emerging narrative of the 'dark side' of data

Several popular books, including Cathy O'Neil's *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy,* have awakened some to the real and perceived threats posed by data. Primarily, these threats concern biases and various forms of inequality that may be inherent in and arise from a greater use of data and algorithms. While many of the concerns raised by these books are valid and important, there is also a great danger that these threats become the dominant trope in conversations and considerations of open data. Unfortunately, as a result of the increased negative connotations associated with data, the burden of proof for those who want to show its potential positive impact has become substantially higher than those who warn of data's risks. Most importantly, a narrative of 'destruction' (especially promoted by several progressive groups), while not exactly wrong, is simplistic and overlooks the many potential benefits of open data.

Partly as a result of this emerging 'destruction' narrative, data has become toxic among many non-government and other stakeholders. We are witnessing the rise of a burgeoning anti-data movement, one whose views are as simplistic and naive as those who have over-hyped and over-championed data. What's required is a far more nuanced and less polemical discussion about data. And, in order to make that discussion possible, we need policies, projects and research that are equally nuanced – that continue to increase access and use of data, yet that balance this against the need for more data responsibility and attention to the risks of data.

New data divides

None of the preceding discussion should be taken to indicate that we are minimising the risks. The challenges of using data are real, and among the most serious unintended consequences is the emergence of a new data divide that rides on, and in many ways exacerbates, the existing digital divide. The emergence of such a new divide is deeply ironic: after all, open data was intended as a tool for democratisation and empowerment. Yet, as with other assets, and as with technology in general, the understanding and the capacity to extract value from open data is not equally distributed. Those who may need data the most often don't realise the value data may have to improve their decision-making. Different skill-sets, and differential access to the tools required to store and analyse data, also mean that there is a very real risk that open data could reinforce existing inequalities and potentially create new ones.

What can we do to avoid such inequalities? Critically, all data stakeholders need to be as attuned to the *reality* of open data as the potential of open data. By this we mean that much greater attention needs to be paid to the actual, realisable possibilities of individuals and groups to access and extract meaning and insight

from data. Open data exists on a continuum of value: the final parts of the value chain, which involve extracting meaning, are as important as the earlier parts, which involve data collection and storage. It is not enough simply to make sure data is made 'open'. We need to ensure that people understand the questions data can answer and that they can use open data, either directly or indirectly.

The role of government is also key here, as it is government that holds the power to strike a balance between informational and human development; it is government that determines the corrective and redistributive policies required to create the conditions for balanced, inclusive development.

The 'magical thinking' of standards

As so often in the technology world, there is an emerging belief that open data as a field can only scale and become truly useful through a greater use of principles and standard-bearing bodies. For instance, the International Open Data Charter seeks to establish a set of standards, expectations and principles for how governments should publish their data. While standards and principles can of course be very useful to establish common expectations, it is also the case that they can hamper innovation and increase barriers to entry, especially among groups who may not have the requisite financial or institutional capacity to meet all requirements of a standard. This can be particularly problematic for countries from the developing world, or cities that want to make their data liquid yet lack the resources. Standards are generally set by early movers, which typically means more developed and resourceful countries; these standards can then set unrealistic or unfeasible expectations for 'late adopters'.

The concern is that, instead of scaling and promoting open data, standards and principles may ultimately hamper the exchange of data. Standards should not be seen as apolitical when their application is inevitably both political and varied across many social contexts. We need to remember that the ultimate goal is to improve people's lives by generating insights from data has been made accessible; not just compliance of principles and standards. In addition, a standard is only a standard, and only creates value, when it becomes widely accepted.

Understanding open data research

The preceding section outlines some key forces currently shaping the state of open data. But what is the state of open research – research that shapes our understanding of these trends and advances the field by providing new, empirically sound insights?

The first Open Data Research Symposium was held in Ottawa in May 2016. Selected papers from that Symposium were published in a special issue of the *Journal of Informatics* (JCI2016). The same journal published an earlier special issue in 2012 titled 'Community Informatics and Open Government Data' (JCI2012).

As far as we are aware, these are the only peer reviewed, edited volumes that focus exclusively on open data. Combined with this volume, *The Social Dynamics of Open Data* (SDOP), it may be instructive to explore what this small sample[1] of publications tells us about shifts in the open data research landscape (if anything). Of course, it is dangerous to talk of trends over a period of five years and across only three scholarly publications. To bolster those insights, we therefore also draw on a second sample of open government data research publications from the bibliographic index of the Clarivate Web of Science.[2, 3] While we acknowledge that the sample remains small – and, importantly, ignores all the research findings shared through other means, including the corpus of grey literature – such an analysis could nonetheless provide some insights into who is conducting research on open data, how they are writing up their research, and who is supporting that research.

How much research on open data is being published?

The sample of articles and chapters in the three publications focused exclusively on open data reveals little about the overall volume of research being published. The bibliometric data is more comprehensive but still excludes those journals (and books) not indexed in Clarivate's Web of Science as well as a vast body of grey literature. Google Scholar's indexing is more inclusive, but the data requires a level of checking and cleaning that is beyond the scope of this modest effort.[4] The data in the bibliometric sample of 216 pubclications do, however, show (1) a marked increase in the number of 'open data publications' from a modest 2 publications in 2008 to 86 publications in 2016, and (2) a rapid increase in the number of publication post-2010 (see Figure 1).

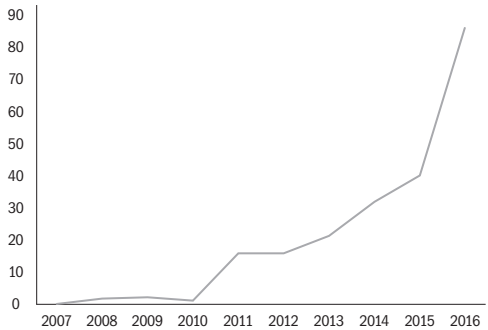
1 The sample of articles and chapters in the three publications focused exclusively on open data consisted of 22 chapters and papers in total: 6 in the 2012 special issue of the *Journal of Community Informatics*, 7 in the 2016 special issue of the same journal, and 9 in this publication. A total of 39 authors contributed to the chapters and papers in the sample.

2 The 'bibliometric sample' consisted of 216 journal articles, books and book chapters on open (government) data. The sample was generated by searching the Web of Science Core Collection for the 10-year period 2007 to 2016 using the search query "TI=('open data' OR 'open government data')" and limiting the search to the publication types 'article', 'book' and 'book chapter'. This returned 264 results. Results related to open science or open research data were removed to ensure a focus on open government data. The Journal of Community Informatics is not indexed by the Web of Science. Given that two of publications in the open data only sample were special issues of the *Journal of Community Informatics*, and that this volume has not yet been published, there is no overlap of publications between the two samples.

3 The collection, cleaning and analysis of the data relied on the primitive data skills of the lead author. The full dataset is available for verification and further analysis.

4 A search on Google Scholar using the same query and date range returned 17,100 results (search done on 14 September 2017).

Figure 1 Number of research publications on open government data indexed in the Web of Science 2007–2016 (n=216)



Where is research on open data being done?

The analysis[5] of the sample of open-data-only volumes shows that authors are mostly affiliated to universities (59%), followed by non-government organisations (30%) and research institutes (9%). Authors are most often and consistently affiliated to universities across all three publications (JCI2012 67%, JCI2016 43%, SDOP 56%).[6] Authors from non-government organisations, typically research-orientated, have emerged more recently (JCI2012 0%, JCI2016 36%, SDOP 44%), and those from research institutes (JCI2012 17%, JCI2016 14%, SDOP 0%), that is non-degree awarding private- or publicly-funded research organisations, have declined. Bibliometric data confirm that most researchers are based at universities (85%). However, only 1 corresponding author out of the 205 for which sufficient address data were available to make a determination as to their institutional affiliation, listed their affiliation as being a non-government organisation. In the case of research institutes, a proportion similar to that of the open data-specific publications was found at 8% (17). Other affiliations were also present in the bibliometric data: 3% (7) were from government and 2% (5) were from private corporations.

Who is conducting research on open data?

In terms of gender, 36% of all authors in the open data-specific sample were female. There were marked differences between the three publications with a sharp swing from predominantly female authors to predominantly male authors (JCI2012: Female 67%, Male 33%; JCI2016: Female 36%, Male 64%; SDOP:

5 The data was analysed using fractional counting in instances where a paper or chapter was multi-authored. For example, if there are three authors, each author is assigned a score of 0.33 and each author contributes fractionally to the variable being measured.

6 JCI2012 and JCI2016 refer to the issues of the *Journal of Community Informatics* published in 2012 and 2016 respectively. SDOP refers to this publication, *The Social Dynamics of Open Data*.

Female 17%, Male 83%). Determining the gender profiles for all authors using the bibliometric data was beyond the scope of this chapter as authors are not coded for gender in the Web of Science. However, using only corresponding authors and coding them for gender based on first names and a Google Search, 209 corresponding authors were identified as being either male or female. Of the 209 corresponding authors, 30% (70) were female.

No readily available data on the ages or career stages of authors were available for analysis.

Where are open data researchers from and how are they collaborating?

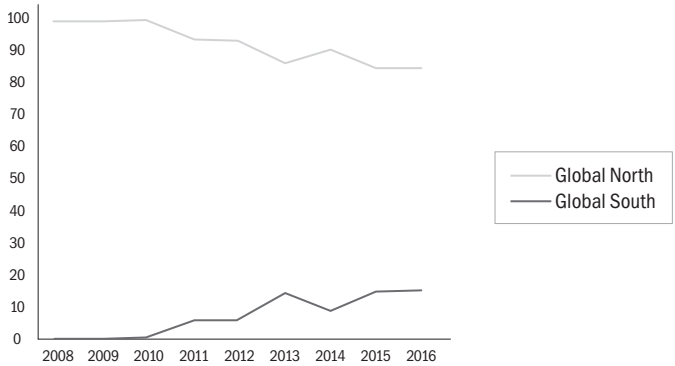
Most authors in the open data specific sample are from the Global North,[7] but only marginally so at 55% across all papers and chapters. However, closer analysis shows that representation of authors from the Global South was highest in the second special issue of the *Journal of Community Informatics* (71%). This is not surprising given that the focus of the special issue was on open data in developing countries. Authors from the Global South represent a much lower proportion in the other two publications (JCI2012 17%, SDOP 44%). This could be interpreted in two ways. First, that authors from the Global South are under-represented when topics aren't specifically focused on developing-country issues. Or, second, that there has been a positive shift from 17% to 44% in representation from the Global South when comparing the two publications that did not have a developing-country focus.

Bibliometric analysis shows that of 216 open data research publications, 88% (189) were published by authors in the Global North (using the corresponding author's address as an indicator of location). The trend data show that there is indeed an increase in the proportion of authors from the Global South, although the gap remains wide (see Figure 2).

What is more definitive, and worrying from a Global North–Gobal South collaborative point of view, is that for the 22 articles and chapters published in the publications focused exclusively on open data, there is not a single example of collaboration between authors of the Global North and the Global South. There is evidence of South–South collaboration in the case of two papers. In fact, collaboration in general is the exception. In the case of the first special issue of the *Journal of Community Informatics,* only 1 (16%) paper was co-authored, and in the *Social Dynamics of Open Data*, 3 (33%) papers were co-authored. The second special issue of the *Journal of Community Informatics* bucked the trend: all papers in that publication were co-authored. Bibliometric analysis of the larger sample of publications shows that the trend is for research publications on open data to be co-authored: 79% (170) publications were authored by two or more researchers, and the average number of authors per publication is 3.29.

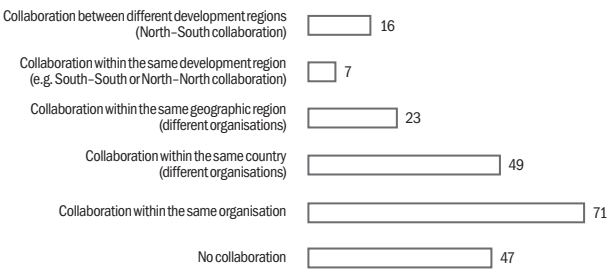
7 Countries were assigned to the Global North or Global South using the following map: https://www.mapsofworld.com/headlinesworld/miscellaneous/division-global-north-global-south

Figure 2 Authors of research publications on open government data from the Global North versus those from the Global South (%, n=216)



In terms of collaboration, the bibliometric data show that for the 213 publications for which author address data were available, 22% (47) of authors did not collaborate. Of those that did collaborate, 33% (71) did so with colleagues in the same organisation, 23% (49) collaborated with colleagues in the same country and 11% (23) collaborated across the region. Only 3% (7) collaborated between regions but within the same development region (e.g. collaboration between authors in the US and Europe), and marginally more (16, 8%) collaborated across development regions (i.e. North-South collaboration, e.g. between authors in Mexico and the US or between authors in Africa and Europe).

Figure 3 Collaboration between authors of publications on open government data indexed in the Web of Science 2007–2016 (n=213)



The stereotypical open data researcher

Sex:	**Male.**
Age:	**Unknown.**
Employment:	**University in the Global North.**
Behaviour:	**Most likely to co-author with colleagues at the same organisation.**

Accessibility of research on open data

All three open-data specific publications are open access. The picture from a universal access point of view is less positive in the case of journal articles, books and book chapters indexed in the Web of Science – only 19% (40) of those publications are published under an open access licence. This finding reflects a key paradox of academic research on open data: while the focus is on the value of open access on society, the authors still decide to publish their findings in a manner that is antithetical to the principles and values of open data.

Who is funding research on open data?

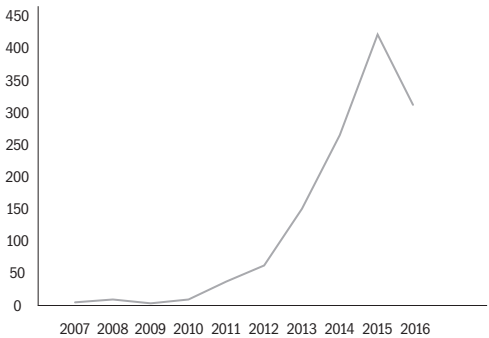
All but one paper in the second special issue of the *Journal of Community Informatics* acknowledge funding support from the IDRC. In the case of the first special issue of the *Journal of Community Informatics,* only 1 (16%) paper acknowledges a funder (The Asia Foundation), and in the *Social Dynamics of Open Data*, 3 (33%) papers acknowledge funding (from the National Commission for Scientific and Technological Research [Chile], Microsoft, IDRC and Avina Foundation). The bibliometric data show that 32% (69) publications included funding acknowledgements. No single funding agency stands out, with the possible exception of the European Union: 20% (14) acknowledge financial support from the EU in one form or another (e.g. European Commission or the European Commission's Seventh Framework Programme). If funding agencies are classified by their geographic focus area, the data show that most funding comes from national science councils and funding agencies (44, 64%). This finding could account for the high levels of intra-organisational and intra-national collaboration in conducting research on open data.

The 'impact' of research on open data

The impact of the new knowledge produced by open data researchers can either be measured within science (i.e. its contribution to further knowledge production) or on society (i.e. the change brought about in society attributable to new knowledge). Neither is easy to measure.

The impact on knowledge production is typically measured in the form of citations. The more frequently a research publication is cited by other researchers, the greater the scientific impact of the publication. Figure 4 shows the number of citations for the sample of publications indexed in the Web of Science. It shows a marked increase in the number of citations, which is to be expected as the number of publications on open data increased. On average, each paper in the sample is cited 5.88 times. Normalisation for scientific field would need to be done to provide an indication of whether citations to open government data journal articles are high or low. At this stage, given the small number of publications and the difficulty of ascribing open data research to a specific scientific field makes such analysis difficult.

Figure 4 Number of citations in the Web of Science (n=213)



Disciplinary perspectives on open data

Open data is inherently an inter-disciplinary topic that straddles a wide variety of areas of social inquiry and research. An analysis of the subject area assigned to the publications in the Web of Science sample shows that the most frequent category by subject is computer science (80, 27%), followed by information science and library science (69, 23%), after which there is a significant drop in the frequency of other subject categories (see Figure 6). The Top 10 subject categories account for 75% (221) of all the subject classifications; non-technical disciplines in the Top 10 such as geography, government and law, social sciences, communication, and public administration only account for a combined 16% (47). This suggests that publications on open data are mostly technical in terms of their content. Conversely, there appears to be a very limited social perspective that is brought to bear on open data by researchers. Further analysis would need to be done to determine whether there are any correlations between subject classification and region (e.g. 'Are researchers in the Global South more attuned to social

dynamics?'), or between subject classification and institutional affiliation (e.g. 'Are non-university research more interested in exploring the social dynamics of open data?').

Table 1 Top 10 subject categories for open data publications indexed in the Web of Science (n=294)

Geography	5
Public Administration	6
Communication	7
Social Sciences - Other Topics	7
Telecommunications	7
Government & Law	8
Business & Economics	14
Engineering	18
Information Science & Library Science	69
Computer Science	80

Note: The number of publications (294) exceeds the number of publications in the sample (216) because a single publication may be assigned to more than one subject category.

Conclusion

The findings and analysis above bring us back to the relevance of this collection of chapters on the social dynamics of open data.

First, the bibliometric data show that there is a relative dearth of scientific literature that focuses on the social dynamics that hinder, constrain, enable, promote or propel the supply, (re)use and impact of open data.

The current tendency in much of the research is, quite simply, to *measure what is measurable*. In practice, this usually means focusing on volume or supply: measuring the amount of data, or number of datasets, that are being released or accessed. Such an approach fails to take into account the full range of factors – social dynamics – that determine the impact of open data, and overlooks the multiple axes along which open data operates. It also doesn't make us smarter about users and non-users. In addition, the problem with this approach is that, over time, researchers and policy-makers tend to start valuing what is measureable and simply what gets measured. This collection's focus on social dynamics goes some way in remedying this asymmetry.

Second, one might argue that the data on the classification of publications by subject is far from reliable as it is often difficult to categorise a topic such as open data, and that classifiers may default to more technical subject areas because of the perceived technicality inferred by 'data' at the expense of the

social complexities inferred by openness. This may be true. And if it is, then combining several research publications on open data into a single volume brings to the attention of researchers, policy-makers and other stakeholders in a clear and unambiguous manner the social dimension of open data. In other words, the chapters in this collection will hopefully escape the fate of research papers that may well share a focus on social dynamics but are unfortunately buried by technically-biased classifications.

Third, while the social perspective appears to be under-represented, the citation data show that publications that deal with social dynamics are some of the most highly cited in the scientific community. This could be interpreted as a proxy indicator of the need for more empirical evidence on the social dynamics at work in contexts in which open data initiatives are conceived and implemented.

Open data is a networked movement and power in networks is corralled by the first-movers and consolidated by those with historically-endowed privilege. The networked nature of the global open data 'movement' is highly relevant in relation to unlocking the development benefits of open data. The network power of global movements, including the open data movement, will, as a structural feature of networks, continue to exclude by determining the rules of inclusion. Under such conditions, social development will fail. And it is the responsibility of research to provide the evidence that exposes the unintended exclusionary outcomes of open data, while simultaneously deploying theory as a tool to explain the observed outcomes in order to recalibrate open data initiatives such that they live up to their potential for creating a more equitable and just society.

Finally, open data has shown robust growth over the last decade, and its potential is now indisputable. But recent years have also shown a few tears in the seams. In particular, the current ideological or faith-based approach to open data, guided primarily by well-intentioned but under-informed enthusiasm, is starting to show its limits. Without more evidence and fact-based analysis, the case for open data – for data 'owners' to release it and for users to access it – may weaken, especially as the case of the potential harm starts to overshadow all debate. We need to develop a more rigorous and fine-combed analysis not only of why open data is valuable, but how it is valuable, and under what specific conditions.

The objective of the Open Data Research Symposium and the subsequent collection of chapters published here is to build such a stronger evidence base. This base is essential to understanding what open data's impacts have been to date, and how positive impacts can be enabled and amplified. We hope this collection provides a foundation for further and deeper research, and especially more evidence-based practice, and hope you will join us in building a community of open data researchers moving forward.