Identifying and Mitigating Risks to the Quality of Open Data in the Post-truth Era

Adrienne Colborne, Michael Smit
School of Information Management
Dalhousie University
Halifax, Canada
Email: Adrienne.Colborne@dal.ca, Mike.Smit@dal.ca

***Abstract*—Big Data analysis often relies on open data, integrating it with large private data sets, using it as ground truth information, or providing it as part of the input to large simulations. Data can be released openly by governments to achieve various objectives: transparency, informing citizen engagement, or supporting private enterprise, to name a few. To the latter objective, Big Data analytics algorithms rely on high-quality, timely access to various data sources, including open data. Examples include retail analytics drawing on open demographic data and weather forecast systems drawing on open weather and climate data. In this paper, we describe the rise of post-truth in society, and the risks this poses to the quality, integrity, and authenticity of open data. We also discuss approaches to identifying, assessing, and mitigating these risks, and suggest future steps to manage this data quality concern.**

***Keywords*-open data; post-truth; fake news; risk identification; risk mitigation; data quality assurance**

I. Introduction

The volume and variety of data currently available to individuals is unprecedented in our history [1]. Improvements in storage and distribution technology along with mobile technology have put vast amounts of data at our fingertips. One societal impact of the reduction of barriers to generating, publishing, and accessing data is the common expectation that all data be open and available [2], particularly data from communities, organizations, and societies to which one belongs. When publicly released by governments, as is increasingly common, this data is called *open data*. To realize the potential of this data, it must be used effectively, which includes analyzing and visualizing raw data to produce information and increase knowledge. The quality of open data is often assumed by end-users, who assume it is an accurate lens into the inner functioning of government.

Typically, open data is defined as any data collected, acquired, or curated by government agencies, and then released[1].Open data is believed to have positive impacts that include encouraging an informed populace, supporting government transparency, and enabling value-added services [3]. Publicly available information about governments, political processes, and elected officials is generally considered an important part of a democracy, particularly when it leads to citizens that are equitably informed [4], [5].

[1]This is less restrictive than other definitions, which require that data be released in a non-proprietary, well-structured digital format.

Open data has come about slowly: early reporters on discussions in the House of Commons were jailed for contempt of Parliament [6], yet by the early 1900's (1880 in Canada and 1909 in the UK), government-funded in-house teams assembled a semi-verbatim record published in the *Hansard* [6]. Later in the 20th century, governments passed legislation requiring the release of certain government documents on request (the Access to Information Act [7], in Canada). Today, the expectation is the default release of datasets designed to be well-formatted and easily usable by machines[2]. In June 2013, the G8 member countries signed an Open Data Charter, committing to build on existing open data initiatives.

Developers are making use of open government data to support transparency, and to drive innovative applications [8], including applications driven by big data analytics and machine learning [9]. These applications require timely access to high-quality information. For machine learning in particular, small errors in data used as input to training stages can be magnified by orders of magnitude as the trained algorithm is scaled to process Big Data.

As governments worldwide push forward open data plans, and developers, analysts, and entrepreneurs continue to rely on extant open data sets, the way some governments view data, and the way some people view governments, is undergoing an important shift. It is important to have a cross-discipline conversation about how our acknowledged "post-truth society" poses risks to the usefulness, growth, and ongoing maintenance of open data.

To begin this conversation, we identified and categorized a set of fundamental risks posted by post-truth governments, along with assessments of how easy each is to detect. We also identified a variety of mechanisms to mitigate these risks, and assessments of how easy each mitigation strategy is. We conclude the most concerning risk, due to its difficulty to detect and mitigate, is increasing the amount of low quality data, specifically data that is subtly flawed or biased in a particular policy direction.

The remainder of the paper is organized as follows. Section II provides additional background and literature,

[2]See for example federal (http://open.canada.ca/en), provincial (e.g. https://data.novascotia.ca/), and municipal (e.g. https://www.halifaxopendata.ca/) sites.

including a more detailed look at the concept of a post-truth society. Section III introduces the risks and assesses their difficulty to detect as well as how difficult it would be to launch each risk. Section IV discusses mitigation strategies. We offer a discussion and conclusion in Section V.

II. Background and Related Work

The concept of a 'post-truth' society was coined in the 1990s [10]. Oxford dictionary, in naming 'post-truth' the word of the year in 2016, defines it as "relating to or denoting circumstances in which objective facts are less influential in shaping public opinion than appeals to emotion and personal belief" [11]. There is concern in the scientific community that Donald Trump's election establishes that we have now definitely entered a post-truth era in politics [12]. The post-truth world poses several challenges to the scientific community and any who prefer "decisions guided by evidence" [13]. Without a perceived need for this guidance, it's unclear what role open data will continue to play in democratic society.

While a certain amount of untruth has always been expected in politics, it is new for newspapers to maintain databases of all false claims the elected president of the United States of America has made [14]. The current political climate is tense and highly polarized [15], which is a threatening environment for truth. Two commonly-cited examples are Brexit in the UK and the election of Donald Trump in the USA, two examples where demonstrably false claims were not considered disqualifying, and where the democratic decisions took place in highly partisan environments, and witnesses pandering to extremist groups, particularly right-wing militant groups like neo-Nazis and the KKK [15]. The presentation of data and information may be influenced by political belief, consciously or unconsciously, and in a post-truth world, it's not clear that most citizens would raise an objection.

It should be noted that politicians and the civil service are different entities, and that often open data is managed by dedicated, honest public servants who are committed to providing data of the highest quality. Yet in a democratic society, we elect leaders and expect them to provide oversight of government agencies, including the departments responsible for open data. There is, generally speaking, an opportunity for the views of politicians to be made clear to those providing open data.

A. Open Data and Trust

Opening up government data is meant to make government data a public good and let people engage with the data in new ways [8]. The increased transparency that results can increase warranted trust while reducing the levels of unwarranted trust [16]. Yet open data is not a universal good: one oft-discussed concern is 'openwashing': releasing large volumes of mainly worthless data to create an illusion of openness, or releasing data in an obscure manner [17]. There is also still work to be done on standards of open data, particularly for quality assessment, and a need to improve the management of the data [17].

Public trust is vital to good governance. It "promotes public participation and collaboration" [18]. One might thus assume that a lack of trust is a critical threat to open data: a state that reduces the likelihood of citizens using open data, and that requires constant vigilance. Yet the higher levels of trust promoted by open data are potentially a greater threat, if that trust is unwarranted. A government that merits trust can threaten open data through inattention, leading to insufficient quality control, preservation, and management of data. An untrustworthy government can do that and more to abuse the misplaced trust of its citizens.

B. Trust as a Dimension of Data Quality

Trust has long been accepted as an important dimension of data quality. The abstract nature of trust leads to a variety of terms reflecting narrower, more measurable attributes, like trustworthiness, veracity, reliability, or dependability [19], and there is some variation in establishing the criteria for when this quality dimension is achieved. The general sense is that "the data can be counted on to convey the right information" [19].

Trust is often suggested as a component of the fourth 'v' of Big Data: volume, variety, velocity, and *veracity*, though it is not one of the original three [20]. Scannapieco [21] goes further, suggesting that a state of trust only exists between organizations when each believes the other will provide high quality data. Wang and Strong [22] describe an enhanced understanding of "accuracy" that includes reliability and certification as components that enhance trustworthiness.

There is also an indication that there is a two-way relationship: that without trust in an organization, we don't consider their data to be of high-quality; but also that if a person is able to identify that data is of low quality, they are less likely to trust the providing institution [23][24].

III. Risks to Open Data

In a post-truth era, principles like objectivity and proof are diminished. A government operating on post-truth principles has to deal with open data somehow, yet open data is "a threat to their power" [25]. If we are to be effective stewards of government open data, we must identify and understand the risks post-truth poses to open data. While perhaps the most obvious step is to close down open data portals, this is a highly overt step, and potentially a self-defeating one given the role open data can play in increasing trust in government. We spend more time on subtle approaches to influencing open data, which can have far-reaching effects as open datasets are leveraged in machine learning and big data analytics.

Table I provides an overview of the top-level categories of identified risk, including a brief description. For each risk, we provide a coarse assessment of four factors: how difficult it is to threaten open data to take advantage of that risk; how likely it is that a post-truth government would exploit that risk; the expected level of harm that would result from that risk if exploited; and how easy it would be for an outsider to detect that this risk has been exploited. Each risk is described in more detail in the following sections.

While each risk is described individually, it should be noted that these risks are all present at the same time, and multiple can be exploited at the same time.

A. Closing down open data

The most obvious risk to open government data is that the government will shut down open data portals. This does not have to be malicious; there could be funding challenges, or the usefulness of the data being open is contested. However, it is possible that the government might do this for partisan reasons, or to restrict public access to certain data types. It would be very easy for this to happen: all the government would have to do is to close off access to government data, shut down open data portals, and stop funding open data initiatives. This would take very little time, and it would mean that some or perhaps all government data is no longer accessible. If done at the level of an entire data portal, this would be an overt move – the public would know immediately, especially in this era of social media – but it has already begun in some places.

A more selective approach, where certain data sets are removed, might be more difficult to notice. For example, after a major hurricane devastated Puerto Rico, a political debate ensued about whether the US federal government was doing enough to support this US territory. The *Washington Post* reported that data about citizens without water and without electricity had been removed from the Federal Emergency Management Agency (FEMA) website [26]. FEMA responded that the data were available elsewhere, but in response to complaints they would resume their practice of replicating the data on their own website [26]. This case could be an innocent error, or could indicate that a government agency was willing to stop sharing open data to support a partisan position. Either way, it demonstrates that even a single data set being removed is an overt act when that data set is highly politicized. In contrast, the removal of climate change data from certain government websites passed with less notice, and has not been returned.

A more passive approach can be to simply let entropy, and the basic passage of time, minimize the impact and growth of open data. This action is difficult to attribute to malice and can simply reflect differing priorities. Few have noticed that the White House Office of Science and Technology Policy website once featured a significant section dedicated to open data, yet now no longer mentions open data at all.

Each of these strategies can be executed easily – the third example requires literally doing nothing. We rate the likelihood as 'Very High', in part because there is some evidence this is already happening in the US. The potential harm is high, as some of this data is uniquely only available from the government, and this strategy has the potential to reduce access to that data. This would reduce the effectiveness of many forms of Big Data analytics It is easy to detect when data that was once open is now closed.

B. Introducing low quality data

This risk takes advantage of the general levels of trust afforded to open data to change the character of the data that's available. For example, funding information gathering or scientific study in government agencies that is based on low-quality science that toes the party line, often by redirecting funds from actual evidence-based inquiry, or cuts to both generating and storing data [27]. A government can indirectly influence the content available.

The difficulty of this option is not the actual act, which is a relatively straightforward undertaking. The difficulty arises in the time it takes for this approach to have an impact on the open data portal, and in the challenge of promoting the research as legitimate. Depending on the discipline, it may be very easy to fool the layperson into calling the research sound, or it may be more trouble than it's worth. The same goes for the chances of detecting these poor data; it may be close enough to legitimate data that only field experts would be able to catch the discrepancies.

In the best case, the harm from this action is reduced value and reduced trust in open data. In the worst case, open data can be used to support and convince people of the quality of the data, exploiting unwarranted trust in the quality of open data. Both of these outcomes have the potential to cause substantial harm, as the loss of confidence in open data is likely to be transferred to Big Data.

C. Tweaking open data

This risk also takes advantage of trust in open data, but in a more direct form: simply altering the data already present in various open data portals. Particularly with open research data and open observation data, updating datasets due to new information is common: instrument recalibration, quality control, error correction, etc. In some cases, tweaking data can can produce a slightly different picture and a different conclusion. We know that small errors or biases in training data can have a substantial impact on machine learning or analytics algorithms, so the potential harm here is high.

While one might be able to detect this by looking for unexpected data sets that support a particular perspective, but in general could be very difficult to detect. Domain experts would be necessary to detect all but the most blatant changes, and there may be nothing to arouse suspicion. Happily, in order to not raise suspicions, the changes would

Table I
IDENTIFIED RISKS AND A COARSE ASSESSMENT AS TO HOW DIFFICULT EACH IS TO EXECUTE, HOW LIKELY IT IS TO BE DEPLOYED, THE MAGNITUDE OF THE EXPECTED HARM, AND THE EASE OF DETECTION.

Risk	Difficulty	Likelihood	Harm	Ease of Detection
Closing down open data: either by removing the data entirely or by defunding open data	Easy	Very High	High	Easy
Diluting the quality of open data by introducing data with bias (e.g. from poor science)	Hard to do quickly	Varies by Discipline	Very High	Varies by Discipline
Tweaking data: making small changes (e.g. 'corrections' and 'recalibrations')	Easy	Medium	Magnified by Big Data	Medium
Reduce the volume of quality open data by reducing spending on information gathering and research projects	Easy	Likely	High	Easy
Making data difficult to locate: obscuration and obfuscation	Easy	Medium	Low	Easy/Hard

have to be made with a certain level of expertise, which might be more trouble than it's worth and makes this option less likely than the others.

D. Reducing high quality data

Much of the high quality data in open data portals is due to government-funded information gathering or research programs, from census data to genetics data. There are various approaches to keeping this data out of open data portals, ranging from no longer approving the release of the data to not staffing key positions to no longer funding the programs at all. One can also reduce the focus on, and funding for, and entire discipline [27].

We rate it 'easy' to exploit this risk, because it again does not take a great deal of effort. While some mechanisms of exploiting this risk would require a government to give up its own high-quality data, we believe it is plausible that a post-truth government would make this "sacrifice". It is thus quite likely.

The potential damage from this risk is quite high, as this would leave a major evidence gap in government decision-making. This may be difficult to detect if it doesn't happen quickly, or may be mistaken as a symptom of another agenda (general attacks on science, for interest).

E. Making data difficult to locate

This risk is another that may occur completely naturally, and thus can be difficult to prove. Data is produced at a prodigious rate, and government data is no exception. With this much data, it is possible that good data could be obscured by bad or useless data accidentally. Deliberate obscuring of data could be accomplished by hiding datasets in different parts of a website, placing datasets on different websites entirely, or by choosing not to reveal its existence. In the FEMA case discussed above, their spokesperson protested that the data was still on the FEMA website, it was just difficult to find. As this phenomenon is happening naturally, it could be fairly easy to hide any deliberate obscuring: citizens almost expect their government website to be difficult to use. The data still exist and are not tampered with, yet data is difficult to discover and access, so we rate the overall harm as 'low'.

IV. MITIGATING RISKS TO OPEN DATA

Our assumption is that while we may be describing a post-truth era, there still exists a set of people who are concerned about the quality of open data, who wish to see it used effectively, and who in general recognize the value of truth and objective fact. In support of this group of people, we suggest avenues to explore for mitigating these risks: reducing the likelihood of occurrence, decreasing the magnitude of harm, or increasing capacity to detect them. This is an initial starting point, and while not quite a research roadmap, does suggest avenues for further inquiry.

Table II provides an overview, for each identified risk, of mitigation strategies and the ease of implementing that strategy. Mitigation strategies are described in more detail in the following sections.

A. Closing down open data

It is difficult to mitigate the likelihood of this happening. The government closing open data could happen with little warning. This is not an issue of public consultation, and waking up to missing open data sets, or even portals, is a real risk. However, it is possible to reduce the magnitude of harm that would result. For example, data rescue and data replication take currently open data and ensure that it will remain open in some capacity, if not in a government portal. This is a reasonably straightforward solution, and is worthwhile to explore even if there is no imminent risk. For example, there were efforts to archive US climate change data in Canada, led by University of Toronto professors [28].

If the data has not been rescued before it is closed, it is possible to raise protests about this in the public sphere. A case can be made that the public deserves access to data. This is not guaranteed, however, and it might take a great

Table II
SUMMARY OF MITIGATION STRATEGIES FOR OPEN DATA RISKS, AND THE EASE OF EMPLOYING THE STRATEGY.

Risk	Mitigation	Ease of Mitigation
Closing down open data	rescue and replication, protesting	Easy to medium
Introducing low quality data	Watchdogs, data literacy, scrutiny from domain experts, external data quality ranking	Hard
Tweaking open data	Watchdogs, rescue and replication, versioning, fingerprints	Hard
Reducing high quality data	Global community	Hard
Making data difficult to locate	Big Data analytics, scrutiny, crowdsourcing, Alternate portals	Medium

deal of effort only to find out that the data will not be reopened.

B. Introducing low quality data

There are a few different solutions for this particular risk. Designating watchdogs for government data, independent organizations made up of domain experts who can evaluate data properly and as objectively as possible, is a good start. There could also be an outside ranking of data quality, so any government datasets which end up in other repositories may be rated according to its *actual* quality. These rankings would be most effective if they were international, or at the least impartial and covering multiple disciplines. Finally, encouraging data literacy [29] among the public would allow the public itself to be a watchdog. A data literate population could critically evaluate data they receive, and this would both help detect and help reduce the harm of this risk.

Unfortunately, none of these solutions are easy. There are already some watchdog organizations such as the Center for Scientific Integrity, a non-profit that is the parent organization of Retraction Watch [30], but with the volume of open government data being released, it will be difficult to monitor everything. The ranking idea could have many other applications than open government data, but there is no official ranking system as of yet and it could take years before one is developed and approved. It would also be excellent if the public was more data literate, but this would also be a long-term educational effort. It will also be difficult to establish these solutions across every possible discipline, particularly in terms of public data literacy.

C. Tweaking open data

As stated above, this option is a great deal of effort for government-funded research. However, if it does happen, it would be more difficult for even domain experts to detect subtle changes, meaning that the watchdogs would have to be vigilant. Data rescue and replication would also be of use here in order to see different versions of datasets, but the challenges of that have been explained above.

Other approaches include examining versioning, including automated analysis of checksums to see if a file has changed, so that datasets in any repository can show when revisions have been made. Dataverse already has versioning options in their repositories, and anyone accessing their data may view a file's versioning history (TK). If this was held as a standard for open government data, it would be an effective method of mitigating this risk, because it would be easier to see when new versions were created, and it would be easier to question why that data was changed. Similarly, Dataverse, a common data repository, advocates for (and supports) the Universal Numerical Fingerprint [31], which operates at the semantic level of the dataset to calculate a type of "fingerprint" for the dataset.

D. Reducing high quality data

In the case of scientific research programs, there is a mitigation strategy: seeking funding elsewhere. If researchers relocate to a different country, they may be able to continue their research without their government's support. This may also put pressure on the offending government to reconsider their decisions, especially if there is a mass exodus of prominent researchers. However, this solution is not sustainable. It may be difficult for many researchers to move, nor is it wise to encourage mitigation strategies that result in a brain drain.

In the case of information gathering programs, and as an alternative to the above mitigation strategy for research programs, the global community can assume the burden of collecting the information in question. While census data (for example) would be difficult to replace externally, some forms of observation (satellite and other remote sensing) can be supported in other countries.

E. Making data difficult to locate

This particular problem is one that all of society is facing, post-truth or not. Developments in big data analytics are the most obvious steps to limit the impacts of this risk, improving our ability to sift through vast amounts of information to find the important pieces. Watchdog scrutiny may be able to help with watching how data is released, noting any patterns and addressing any concerns. Finally, crowdsourcing can be an unofficial watchdog, using numbers of people rather than expertise to monitor the ebb and flow of open government data.

In case of data that is online but simply difficult to find due to the structure of the website, replication of open data sets on "shadow" websites is permitted by most open data licenses. Third parties could provide portals that are easier to use, or easier to search, or that aggregate open data from multiple disparate locations.

V. Conclusion

The arrival of the post-truth era should cause all of us who work in the Big Data space to critically evaluate the data we rely on, including open data (big and small). We have identified and described various risks, and strategies for exploiting these risks, that we might encounter when interacting with government in a post-truth era. Most risks were easy to exploit, and had the potential to cause substantial harm. Fortunately, for each there are mitigation strategies, at varying levels of difficulty and complexity.

An important next step is to undertake development and detailed assessment of these mitigation strategies, including the further development of strategies like dataset quality assessment, dataset fingerprinting, and data rescue tools and processes. Simultaneously, we hope for a robust and thoughtful discussion about the risks we've identified, how we categorized multiple variants of a risk, and which risks we may not have identified at all.

Open data is released by governments, and it's easy to think about that as being about politicians and civil servants. We must remember that open data is a public asset – *our* public asset. It's important that we be thoughtful and vigilant in protecting this asset for the benefit of society.

References

[1] P. C. Zikopoulos, C. Eaton, D. deRoos, T. Deutsch, and G. Lapis, *Understanding Big Data*. McGraw-Hill, 2012.

[2] C. Shirky, *Here Comes Everybody: The Power of Organizing Without Organizations*. Penguin Press, 2008.

[3] T. Davies, "Open data, democracy and public sector reform: A look at open government data use from data.gov.uk." Edited version of Masters dissertation available from http://practicalparticipation.co.uk/odi/report/, University of Oxford, August 2010.

[4] M. X. Delli Carpini, "In search of the informed citizen: What Americans know about politics and why it matters," *The Communication Review*, vol. 4, no. 1, pp. 129–164, 2000.

[5] A. Blandford, D. Taylor, and M. Smit, "Examining the role of information in the civic engagement of youth," in *Proceedings of the Annual Meeting of the American Society for Information Science and Technology*, 2015, pp. 1–9.

[6] J. Ward, *The Hansard Chronicles: A Celebration of the First Hundred Years of Hansard in Canada's Parliament*. Toronto, Canada: Deneau and Greenberg, 1980.

[7] Access to Information Act, Revised Statutes of Canada 1985, c. A-1, 1985.

[8] T. Jetzek, M. Avital, and N. Bjorn-Andersen, "Data-Driven Innovation through Open Government Data," *Journal of Theoretical and Applied Electronic Commerce Research; Curicó*, vol. 9, no. 2, pp. 100–120, May 2014. [Online]. Available: http://search.proquest.com.ezproxy.library.dal.ca/docview/1535033791/abstract/489982A48C3A4A3FPQ/96

[9] R. Kitchin, *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage, 2014.

[10] R. Kreitner, "Post-truth and its consequences: What a 25-year-old essay tells us about the current moment," *The Nation*, 2016. [Online]. Available: http://www.thenation.com/article/post-truth-and-its-consequences-what-a-25-year-old-essay-tells-us-about-the-current-moment/

[11] "Word of the Year 2016 is... | Oxford Dictionaries." [Online]. Available: https://en.oxforddictionaries.com/word-of-the-year/word-of-the-year-2016

[12] J. L. Vernon, "Science in the Post-Truth Era," *American Scientist*, vol. 105, no. 1, pp. 2–2, Feb. 2017. [Online]. Available: http://ezproxy.library.dal.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eih&AN=120308231&site=ehost-live

[13] I. L. Boyd, "Take the long view," *Nature News*, vol. 540, no. 7634, p. 520, Dec. 2016. [Online]. Available: http://www.nature.com/news/take-the-long-view-1.21189

[14] *Daniel Dale's Trump Fact Checks*, October 5, 2017. Toronto Star. [Online]. Available: https://www.thestar.com/news/donald-trump-fact-check.html

[15] M. Gross, *The dangers of a post-truth world*. Elsevier, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0960982216315159

[16] K. O'Hara, "Transparency, open data and trust in government: shaping the infosphere," in *Proceedings of the 4th annual ACM web science conference*. ACM, 2012, pp. 223–232. [Online]. Available: http://dl.acm.org/citation.cfm?id=2380747

[17] F. Gonzalez-Zapata and R. Heeks, "The multiple meanings of open government data: Understanding different stakeholders and their perspectives," *Government Information Quarterly*, vol. 32, no. 4, pp. 441–452, Oct. 2015. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0740624X1530006X

[18] S. A. Fadairo, R. Williams, and E. Maggio, "Accountability, Transparency and Citizen Engagement in Government Financial Reporting," *The Journal of Government Financial Management; Alexandria*, vol. 64, no. 1, pp. 40–45, 2015. [Online]. Available: http://search.proquest.com.ezproxy.library.dal.ca/docview/1711620148/abstract/489982A48C3A4A3FPQ/39

[19] Y. Wand and R. Y. Wang, "Anchoring data quality dimensions in ontological foundations," *Commun. ACM*, vol. 39, no. 11, pp. 86–95, Nov. 1996. [Online]. Available: http://doi.acm.org/10.1145/240455.240479

[20] D. Laney, "3d data management: Controlling data volume, velocity, and variety," in *Application Delivery Strategies*. META Group Inc. (now Gartner), 2001, vol. 949.

[21] M. Scannapieco, A. Virgillito, C. Marchetti, M. Mecella, and R. Baldoni, "The daquincis architecture: a platform for exchanging and improving data quality in cooperative information systems," *Information systems*, vol. 29, no. 7, pp. 551–582, 2004.

[22] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," *Journal of management information systems*, vol. 12, no. 4, pp. 5–33, 1996.

[23] A. I. Nicolaou and D. H. McKnight, "Perceived information quality in data exchanges: Effects on risk, trust, and intention to use," *Information systems research*, vol. 17, no. 4, pp. 332–351, 2006.

[24] T. C. Redman, "The impact of poor data quality on the typical enterprise," *Communications of the ACM*, vol. 41, no. 2, pp. 79–82, 1998.

[25] J. Gurin, "Open Governments, Open Data: A New Lever for Transparency, Citizen Engagement, and Economic Growth," *SAIS Review of International Affairs*, vol. 34, no. 1, pp. 71–82, Jun. 2014. [Online]. Available: https://muse.jhu.edu/article/547662

[26] J. Johnson, "FEMA removes — then restores — statistics about drinking water access and electricity in Puerto Rico from website," *Washington Post*, Oct. 2017. [Online]. Available: https://www.washingtonpost.com/news/post-politics/wp/2017/10/05/fema-removes-statistics-about-drinking-water-access-and-electricity-in-puerto-rico-from-website/

[27] J. A. Teixeira Da Silva and J. Dobránszki, "Potential Dangers with Open Access Data Files in the Expanding Open Data Movement," *Publishing Research Quarterly; New York*, vol. 31, no. 4, pp. 298–305, Dec. 2015. [Online]. Available: http://search.proquest.com.ezproxy.library.dal.ca/docview/1731483899/abstract/489982A48C3A4A3FPQ/45

[28] *Toronto 'guerrilla' archivists to help preserve US climate data*, December 15, 2016. BBC. [Online]. Available: http://www.bbc.com/news/world-us-canada-38324045

[29] C. Ridsdale, J. Rothwell, M. Smit, H. A. Hassan, M. Bliemel, D. Irvine, D. Kelly, S. Matwin, and B. Wuetherick, "Strategies and best practices for data literacy education: Knowledge synthesis report," Dalhousie University, Tech. Rep., 2015. [Online]. Available: http://hdl.handle.net/10222/64578

[30] "The Center For Scientific Integrity," Retraction Watch. [Online]. Available: http://retractionwatch.com/the-center-for-scientific-integrity/

[31] D.-L. Magazine, "The dataverse network®: an open-source application for sharing, discovering and preserving data," *D-lib Magazine*, vol. 17, no. 1/2, 2011.