11-2017

Contemporary Issues of Open Data in Information Systems Research: Considerations and Recommendations

Georg J.P. Link
University of Nebraska at Omaha, glink@unomaha.edu

Kevin Lumbard
University of Nebraska at Omaha, klumbard@gmav.unomaha.edu

Kieran Conboy
NUI Galway

Michael Feldman
University of Zurich

Joseph Feller
University College Cork

Communications of the Association for Information Systems

Accepted Manuscript

Contemporary Issues of Open Data in Information Systems Research: Considerations and Recommendations

Georg J.P. Link
University of Nebraska at Omaha
USA
glink@unomaha.edu

Kevin Lumbard
University of Nebraska at Omaha
USA

Kieran Conboy
Lero, NUI Galway
Ireland

Michael Feldman
University of Zurich
Switzerland

Joseph Feller
University College Cork
Ireland

Jordana George
Baylor University
TX, USA

Matt Germonprez
University of Nebraska at Omaha
USA

Sean Goggins
University of Missouri
USA

Debora Jeske
University College Cork
Ireland

Gaye Kiely
University College Cork
Ireland

Kristen Schuster
King's College London
United Kingdom

Matt Willis
Oxford Internet Institute
United Kingdom

Received date: 04/04/2017
Accepted date: 05/29/2017

Please cite this article as: Link, Georg J.P.; Lumbard, Kevin; Conboy, Kieran; Feldman, Michael; Feller, Joseph; George, Jordana; Germonprez, Matt; Goggins Sean; Jeske, Debora; Kiely, Gaye; Schuster, Kristen ; Willis, Matt : Contemporary Issues of Open Data in Information Systems Research: Considerations and Recommendations, *Communications of the Association for Information Systems* (forthcoming), In Press.

Published URL: http://aisel.aisnet.org/cais/vol41/iss1/25/

Communications of the Association for Information Systems

Research Article ISSN: 1529-3181

Contemporary Issues of Open Data in Information Systems Research: Considerations and Recommendations

Georg J.P. Link
University of Nebraska at Omaha, USA
glink@unomaha.edu

Kevin Lumbard
University of Nebraska at Omaha, USA

Kieran Conboy
Lero, NUI Galway, Ireland

Michael Feldman
University of Zurich, Switzerland

Joseph Feller
University College Cork, Ireland

Jordana George
Baylor University, TX, USA

Matt Germonprez
University of Nebraska at Omaha, USA

Sean Goggins
University of Missouri, USA

Debora Jeske
University College Cork, Ireland

Gaye Kiely
University College Cork, Ireland

Kristen Schuster
King's College London, United Kingdom

Matt Willis
Oxford Internet Institute, United Kingdom

Abstract:

Researchers, governments, and funding agencies are calling on research disciplines to embrace open data - data that is publicly accessible and usable beyond the original authors. The premise is that research efforts can draw and generate several benefits from open data, as such data might provide further insight, enabling the replication and extension of current knowledge in different contexts. These potential benefits, coupled with a global push towards open data policies, brings open data into the agenda of research disciplines – including Information Systems (IS). This paper responds to these developments as follows. We outline themes in the ongoing discussion around open data in the IS discipline. The themes fall into two clusters: (1) The motivation for open data includes themes of mandated sharing, benefits to the research process, extending the life of research data, and career impact; (2) The implementation of open data includes themes of governance, socio-technical system, standards, data quality, and ethical considerations. In this paper, we outline the findings from a pre-ICIS 2016 workshop on the topic of open data. The workshop discussion confirmed themes and identified issues that require attention in terms of the approaches that are currently utilized by IS researchers. The IS discipline offers a unique knowledge base, tools, and methods that can advance open data across disciplines. Based on our findings, we provide suggestions on how IS researchers can drive the open data conversation. Further, we provide advice for the adoption and establishment of procedures and guidelines for the archival, evaluation, and use of open data.

Keywords: Open Data, Open Research Data, Open Scientific Data, Open Data in Research, Data Sharing, Open Access to Data, Open Science

[Department statements, if appropriate, will be added by the editors. Teaching cases and panel reports will have a statement, which is also added by the editors.]

[Note: this page has no footnotes.]

This manuscript underwent [editorial/peer] review. It was received xx/xx/20xx and was with the authors for XX months for XX revisions. [firstname lastname] served as Associate Editor. **or** The Associate Editor chose to remain anonymous.]

1 Introduction

Around the globe, open data initiatives advocate for research data as a public good. Open data creates value because anyone can distil information and gain knowledge from such data to improve the well-being of society (OECD, 2007). Funding agencies around the world including in the U.S.A., the European Union, and China have started requiring that research data from publicly-funded research be made openly available. The National Science Foundation (NSF) defines open data as "publicly available data structured in a way to be fully accessible and usable" (NSF, 2016a).

The IS discipline is discussing open data as researchers recognize the value of promoting the publication of research data to validate research results and allowing secondary researchers to analyze data in novel ways to generate unpredictable knowledge (Ribes & Polk, 2014). Open data has been considered in other disciplines. One success story is the Human Genome Project and its GenBank, an open data collection of gene sequences (Benson et al., 2015). GenBank has proven to be a valuable resource in biological research, but no comparable infrastructure exists for open data in the IS discipline.

As such, the purpose of this paper is threefold. First, following an introduction to the discourse around open data internationally, we present a summary of the current discussion around open data in the IS discipline. Second, we present consolidated discussions from the pre-ICIS workshop "Issues in Shared and Collaborative Scientific Research" held in Dublin, Ireland, 2016. Third, we suggest a path forward by learning from partner disciplines to foster a larger discussion in the IS discipline, highlighting the skills IS can contribute and promote open data across disciplines.

2 Open Data in the Global Discourse

A number of developments in the last 20 years have contributed to the emergence of open data efforts, repositories, and policies. In addition to the advent of the Internet, the international open access movement emerged in 2002 (Budapest Declaration, 2002) and has shaped many practices and policies within IS and other disciplines. Open access was first applied to research publications (Budapest Declaration, 2002) and later expanded to include "original scientific research results, raw data and metadata, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material" (Berlin Declaration, 2003, p. 1). The term "open access" remains synonymous for open access to publications, which may include a number of open components (e.g., text, data, or graphics). In this paper, open data receives attention of its own.

Open data efforts have been ongoing in several disciplines. Some of these efforts predate the Internet era. Examples include (1) one of the oldest, international open data initiatives, World Data System, established in 1958 in the physics discipline (https://www.icsu-wds.org); (2) GenBank, which was created in 1982 for biological research (Benson et al., 2015); and (3) the Sloan Digital Sky Survey which has been collecting sky images since 2000, creating "the most detailed three-dimensional maps of the Universe ever made" (http://www.sdss.org).

In 2004, the Organisation for Economic Co-operation and Development (OECD) Committee for Scientific and Technological Policy made a global push for open data (OECD, 2004). The report reads:

> *Ministers recognised that fostering broader, open access to and wide use of research data will enhance the quality and productivity of science systems worldwide. They therefore adopted a Declaration on Access to Research Data from Public Funding, asking the OECD to take further steps towards proposing Principles and Guidelines on Access to Research Data from Public Funding, taking into account possible restrictions related to security, property rights and privacy. (OECD, 2004)*

The OECD followed up in 2007 with a recommendation called the *OECD Principles and Guidelines for Access to Research Data from Public Funding* (OECD, 2007). The goal of this recommendation is to "increase the return on public investments in scientific research" (OECD, 2007, p. 9) by improving "the efficiency and effectiveness of the global science system" (OECD, 2007, p. 13). The OECD identified many potential benefits of open data and recommends that national governments enact open data laws that are compatible with each other to foster more data sharing.

As with many innovations and data-driven efforts, public bodies and researchers became aware of the conditions that need to exist for open data to lead to meaningful insight. Unexpected barriers are only now being dismantled through a more collaborative effort across the different stakeholders involved such as the individuals and owners of devices that generated the data, the organizations holding the data, the professionals tasked with the analysis, and the general public. In other disciplines, barriers to data sharing are being systematically identified as a means to tackle them in the long term (e.g., van Panhuis et al., 2014).

However, while some barriers are taken down, new concerns about open data, information justice (Johnson, 2014), misinformation, data protection, and data abuse remain (Barry & Bannister, 2014). In order to gain the expected benefits from open data efforts and to increase the potential, several organizations in different countries have outlined new policies and expectations for open data.

While some of these outcomes are informal recommendations, several countries have established rules for research on open data to guide the access to, management and security of open data. In 2013, the United States' White House Office of Science and Technology Policy released the memorandum *Increasing Access to the Results of Federally Funded Research* (Holdron, 2013). Consequently, as one example, the National Science Foundation (NSF) responded with a public access plan, *Today's Data, Tomorrow's Discoveries*, that requires that all federally funded research projects, from 2016 onwards, provide a data management plan for disseminating and sharing research data (NSF, 2016b). Since 2015, the National Science Foundation of China (NSFC) requires all research papers resulting from projects funded by the NSFC be made openly available through the Open Repository of National Natural Science Foundation of China (http://or.nsfc.gov.cn/). In 2015, the EU started an *Open Research Data Pilot* (https://www.openaire.eu/h2020-oa-data-pilot) to experiment with policies for open data. The EU also adopted the FAIR Guiding Principle (European Commission, 2016) that was developed by researchers for scientific data management and stewardship; the principle focuses on enabling automation for finding, accessing, interoperating, and reusing (FAIR) data (Wilkinson et al., 2016).

3 Method

In response to the emerging global discourse, we believe it is incumbent on IS researchers to understand, participate, and shape the nature of open data. The following methods capture the discussion and identify themes relating to contemporary issues of open data in IS research.

3.1 Literature Review

We performed an extensive review of IS literature from January 2014 to November 2016. We considered the Association for Information Systems (AIS) as the home of the IS discipline and focused on the AIS eLibrary and the Senior Scholars' Basket of Journals which together include AIS conferences, affiliated conferences, AIS chapter proceedings, special interest group publications, and several journals. We conducted a pilot search and identified relevant search terms. The preliminary analysis of the AIS libraries indicated that the discussion of open data was sparse and relatively new. We expanded the search to resources from partner disciplinary bodies, including the Academy of Management (AoM) and the Association for Computing Machinery (ACM). Using NVivo 11 software, we ran a word frequency query and displayed the results in a word tree to identify commonly used word combinations in papers that discussed open data. The results of this query produced eight two-word combinations that commonly appeared together in papers discussing open data: open data, research data, data sharing, open publication, open access, data access, open research, and open science.

These eight terms were then used to identify relevant articles, starting with the literature referenced in the AIS libraries. The first and second author of this paper read the articles to confirm relevance and used an open coding method to identify themes. To establish preliminary themes and inter-coder agreement, both coders used NVivo 11 software, independently named themes for recurring topics in one paper, and resolved differences through discussion (Creswell, 2013). The coders used the preliminary themes on the remaining papers and discussed changes to the themes to encompass the nuances in the papers.

In the next step, we expanded our search to the AoM and ACM libraries to validate the themes, potentially identify additional themes, and discover seminal works. The complementary search used the same search terms and same time span. The first and second author applied the themes from the AIS search on the AoM and ACM results and looked for new themes.

3.2 Expert Workshop

To advance the discussion on open data in the IS discipline, an open invitation for a workshop was posted to the AISWorld mailing list. Interested researchers met on December 10, 2016 at a pre-ICIS workshop in Dublin, Ireland (see Appendix A, Figure 1). The workshop provided a platform for researchers from IS and related disciplines to present their research projects and ideas on how open data was shaping research topics and practices. Discussion groups were formed at five tables each with approximately four people (see Appendix A, Figure 2). Each presenter gave a 15-minute presentation and afterward the groups discussed various questions on the topic. Each group selected one question for the presenter to address. All questions not asked were collected in writing and provided to the presenter after the workshop. The first and second author took notes and identified the appearance of themes from the literature as well as new themes not found in the literature. A detailed report of individual presentations is available in Appendix B. After the presentations, the participants discussed issues of open data, including the development of the Open Community Data eXchange (OCDX) specification (see Appendix C), funding sources for open data, and the themes identified from the literature and from the workshop, which resulted in this paper.

4 Results: Open Data Themes in AIS

Using the literature review and workshop discussions, we identified two clusters of themes in the discussion on open data – motivation for open data and implementation of open data – each containing related themes. The identified themes includes those relevant to and represented in IS research, but also themes in other disciplines that have the potential to impact and shape IS research in this area. The findings of themes in the open data discussion are presented below.

The AIS libraries search yielded seventy-six results, of which seven discussed open data and sharing of research data: one journal article, one journal editorial, and five conference papers. We looked at partner disciplines housed by the Academy of Management (AoM) and the Association for Computing Machinery (ACM) to validated the themes found in the AIS literature. The discussion on open data in the AoM was comparable to AIS – with few results. AoM had five abstracts in proceedings; one was for a symposium (Bosco, Steel, & McDaniel, 2014) and one was for a panel discussion (Donia, Jimenez, & Shah, 2015), which indicates that the discussion is young and actively ongoing. Table 1 shows the papers found in each library.

In the context of our analysis, we noted trends of the divergence, convergence, and foci of the work currently available. Many themes present in AIS were also found in AoM abstracts, except for the themes of standards and data quality. The discussion on open data in ACM dates back further and includes numerous published papers compared to AIS or ACM. A top search result in ACM, Pasquetto et al. (2015), was a survey of publications on open data across multiple disciplines, which analyzed 10 years of highly cited publications and identified eight themes. The difference between the ACM themes and our AIS themes was primarily rooted in terminology. The discussion in the ACM is practice-oriented and empirically supported which is evident in the types of papers: Empirical studies exploring differences in data sharing practices between academic researchers and non-academic researchers (Pollock, 2016); preliminary results on how a research institution builds data expertise (Thompson, 2015); open data sharing infrastructure for academics (Cohen & Lo, 2014); citable companions to datasets (Robles et al., 2014); and cases regarding the practice of sharing and reusing data (Curty et al., 2016).

Following the identification of 19 articles (Table 1) across the different libraries, we identified two clusters of themes which focused on either the **motivation for open data** or the **implementation of open data**. The first cluster is a collection of four themes that motivate or discourage open data. This cluster includes topics such as mandated sharing, benefits to the research process, extending the life of research data, and career impact. The second cluster is a collection of five themes about issues concerning the realization of open data, including governance, socio-technical system, standards, data quality, and ethics. The last theme, ethics, emerged from the workshop discussions. Next we describe each theme as found in the literature and summarize concerns and issues that surfaced in the workshop discussions. For better readability, the workshop presentation summaries and resulting in-depth questions relating to open data motivation and implementation were moved to Appendix B.

Table 1. Open Data Papers Published Between 2014-2016

Library	Paper Titles
AIS	• Editorial—Big data, data science, and analytics: The opportunity and challenge for is research. (Agarwal & Dhar, 2014) • Content category selection towards a maturity matrix for ICT4D knowledge sharing platforms. (Biljon, Pottas, Lehong, & Platz, 2016) • Transparent data supply for open information production processes. (Laine, Lee, & Nieminen, 2015) • Flexibility relative to what? Change to research infrastructure. (Ribes & Polk, 2014) • Science through the "Golden Security Triangle": Information security and data journeys in data-intensive biomedicine. (Tempini, 2016) • Modes of governance in inter-organizational data collaborations. (van den Broek & Veenstra, 2015) • A commons perspective on genetic data governance: The case of BRCA data. (Vassilakopoulou, Skorve, & Aanestad, 2016)
AoM	• The "Big Science" revolution in management: Possibilities, technology, and applications. (Bosco et al., 2014) • Delay and secrecy: Does industry sponsorship jeopardize disclosure of academic research? (Czarnitzki, Grimpe, & Toole, 2014) • Research crowdsourcing, data sharing, and large-scale collaboration. (Donia et al., 2015) • Democratization or reflection: The paradox of databases' influences on knowledge production. (Paik & Binz-Scharf, 2014) • Open data in industrial R&D: Organizing open collaboration between firms and public science. (Perkmann & Schildt, 2014)
ACM	• Academic torrents: A community-maintained distributed repository. (Cohen & Lo, 2014) • Untangling data sharing and reuse in social sciences. (Curty et al., 2016) • Toward a conceptual framework for data sharing practices in social sciences: A profile approach. (Jeng, He, & Oh, 2016) • Exploring openness in data and science: What is "Open," to whom, when, and why? (Pasquetto, Sands, & Borgman, 2015) • Understanding scientific data sharing outside of the academy. (Pollock, 2016) • FLOSS 2013: A survey dataset about free software contributors: Challenges for curating, sharing, and combining. (Robles et al., 2014) • Building data expertise into research institutions: Preliminary results. (Thompson, 2015)

4.1 The Motivation for Open Data

The first theme in motivation for open data is **mandated sharing**. Several funding agencies have implemented policy changes such that research data be made publicly available. These policy changes encourage open data and are necessary complements to existing technological developments that enabled data sharing (Vassilakopoulou et al., 2016). Through the funding agencies' requirements, "policy of data ownership has been significantly reshaped, increasingly emphasizing sharing and reuse" (Ribes & Polk, 2014, p. 292). In a different context, research institutions choose to make open data a contingent requirement for choosing laboratories for collaboration and foster a culture of open data (Vassilakopoulou et al., 2016).

The second theme is **benefits to the research process**. The research process benefits from open data by allowing secondary researchers to replicate results (Ribes & Polk, 2014). Data sharing and transparency are encouraged since open data provides an opportunity for disproving or confirming research results (Agarwal & Dhar, 2014). The importance of open data in the research process is further evident in reported cases where medical data was no longer shared by a laboratory and researchers asked physicians to share the reported data (Vassilakopoulou et al., 2016).

The third theme is **extending the life of research data** beyond the collecting research project. Open data allows data users to apply new perspectives and use it for unexpected purposes that can uncover previously hidden details (Laine et al., 2015), for example by linking it to other data (Tempini, 2016). The importance of maintaining open data has secured funding for a research project where the value of the data would have diminished if the knowledgeable researchers had abandoned it (Ribes & Polk, 2014). Open data can be most valuable when it is used by people who already know the data or who can familiarize themselves with the data, extending its life (Ribes & Polk, 2014).

The fourth theme is **career impact**. Open data can benefit researchers by fostering collaboration allowing researchers to learn from one another, to pool resources, and to scale the outcome of their research (van

den Broek & Veenstra, 2015). Researchers can utilize open data when collecting data is too costly. While much discussion around open data highlights the benefits for researchers, the downsides or impediments of open data are also discussed. Not everyone is supportive of sharing their data openly out of fear to lose a competitive advantage (Vassilakopoulou et al., 2016), or as Ribes and Polk (2014) describes the impediments:

> *There are good reasons for scientists to be wary of sharing their data. Making data publicly available is often perceived as threatening the epistemic authority and fruitfulness of the scientist who must "give it up:" public data may be reanalyzed, possibly revealing flaws in the analytic method, and scientists may fear "being scooped" if other researchers are able to generate findings from their own data first. Furthermore, making data sharable is an arduous and unrewarding task: data are collected in ways that make them indecipherable to outsiders, organized in ways idiosyncratic to those who collected them, or stored in formats that are not easily transferred. Making data easily shared—interoperable—is laborious and expensive work.* (p. 297)

At the ICIS workshop, the motivations for open data resonated with participants and were prevelant in the presentations and discussions. Jordana George (Appendix B.1) introduced issues with setting up and running data repositories to meet the requirements of mandated sharing. She noted that when data repositories are designed, developers must be mindful of the usability and ease of use for the primary researchers, while also considering the need for proper documentation (e.g., authorship, contributors, and use of data complete with report references). Providing such features can have a career impact (a point noted by several workshop participants) as data citations can be strong motives for sharing of data. Debora Jeske (Appendix B.2) highlighted the motivational challenges for interdisciplinary research projects when different funding agencies and academic disciplines generate conflicting requirements which hinder collaboration efforts. Michael Feldman (Appendix B.3) pointed out that these challenges are heightened when a research project involves crowdsourcing and citizen science. Engaging experts and non-experts to help with analyzing open data is a novel approach that can benefit the research process and allow for unexpected findings beyond the purpose of what the data was originally collected for, but also creates new managing challenges.

4.2 The Implementation of Open Data

The first theme in the implementation of open data is **governance**. Open data governance is the result of a negotiation between stakeholders (van den Broek & Veenstra, 2015). Ribes and Polk (2014) provides an example where the negotiation over ownership resulted in an agreement that primary investigators were first authors on scientific findings but not on methodological findings. Data supply and access are also issues of governance. In some instances, researchers freely contribute their data. In other instances, contributing back to a database is a requirement for using open data. The goal of open data is to be accessible by all, but some research data contains sensitive information and requires protection to avoid ethical issues (Tempini, 2016).

The second theme is the enabling **socio-technical system** which collects and disseminates data. Open data provides the most value when it is "managed by well-defined and quality controlled information production processes" (Laine et al., 2015, p. 3). The need to maintain tools and curate data creates new roles for system administrators and data curators (Ribes & Polk, 2014). The socio-technical system can have an impact on how data is collected (Ribes & Polk, 2014) but allows researchers to judge the quality by providing provenance information (Laine et al., 2015). Tools need to be "accessible and their content useful to the target audience" (Biljon et al., 2016, p. 1) and even security should not increase complexity for using open data (Tempini, 2016).

The third theme is **standards** for representation and metadata which simplify data sharing (Ribes & Polk, 2014). Standards are needed in the creation of quality data, for example, to ensure that labels are used consistently or that time stamps refer to the same events (Laine et al., 2015). A lack of standards can discourage the contribution of open data to a database (Vassilakopoulou et al., 2016). Over time, standards in data representation or data collection might change which requires data transformation to ensure accessibility and comparability between historical and new data (Ribes & Polk, 2014). Metadata

standards capture provenance information, such as transformations, the context of data collection, or the original data creator (Laine et al., 2015).

The fourth theme is **data quality**. Data quality intersects with all other themes. The lack of data quality in a database discourages others from contributing their data (Vassilakopoulou et al., 2016). Data quality is improved with metadata which provides traceability through provenance information on the production process behind a dataset, who created the data for what purpose, how that data was created and manipulated, and what measurement errors or biases might exist (Laine et al., 2015).

At the ICIS workshop, the issues concerning the implementation of open data resonated with participants and were prevelant in the presentations and discussions. For example, the need for careful design of socio-technical systems underpinning data repositories to ensure data quality was discussed by Jordana George (Appendix A.1). However, some challenges identified at the workshop were related to general data management rather than open data per se. The first challenge identified was a greater degree of diversity in collaborating teams. Interdisciplinary research projects often have to meet certain standards for data representation and understanding (Appendix B.2). Good design of repositories will be key to facilitate and support these standards. Michael Feldman (Appendix B.3) pointed out that involving non-academics through crowdsourcing and citizen science yields further design issues. This means the design of socio-technical systems has to go hand-in-hand with system-specific training, considerations for the research process, and data quality management. The use of knowledgeable teams, researchers or citizens (non-experts) in open data efforts can generate its own set of opportunities as well as challenges. The work of one other workshop participant provided further insights in this respect. Gaye Kiely (Appendix B.4) focused on the management issues that distributed research teams face when interacting with open data. Coordination, knowledge sharing, sustainability, communication, and engendering team cohesion are key goals when managing a research team. As such, governance of a research project, including the negotiation between stakeholders but also the investment of those in the team, is dependent on how effectively such teams function.

A second challenge regards the need for effective coordination of research and the willingness to participate in such research. Matt Willis (Appendix B.5) investigated how collaboration occurs through documents, emails, and other products throughout digital collaboration systems. For researchers to share their collaboration documents, the research project needs to provide a benefit back to the researchers. This kind of data can be easily collected by means of automated tools that pose challenges for guaranteeing privacy, data quality, and analyzing the data.

Ethics emerged as the third challenge from the presentations and discussions. Ethics is particularly relevant to the properties of open data and was therefore added as a theme in the open data discussion. The emergence of ethics raised a number of issues including questions about how researchers can ensure the anonymity of research participants (Appendix B.1). These points also reflect our discussion around data quality and standards in data sharing practice. This includes a protocol for limiting the amount of identifiable information included in files on data repositories while keeping sufficient records to ensure transparency and trust in public bodies (O'Hara, 2012). Informed consent is a prerequisite in many social sciences where deception is used, but the rules and regulations vary across countries and other research disciplines, causing potential ethical concerns for interdisciplinary research and using open data (Vitak, Shilton, & Ashktorab, 2016). The issue of ethics in online data research is therefore quite pronounced. The workshop participants believe it to be a good practice to obtain informed consent from research participants for releasing anonymized datasets before sharing open data. This is especially critical when automatically collected data may contain personal information such as email addresses or the participants contributed to different datasets (e.g., via wearable devices), allowing for their triangulation and identification. This touches on the need to balance the stakeholders' interests, meet data quality concerns, while also extending the life of research data.

5 Discussion: A Roadmap Forward for Open Data in the IS Discipline

In this paper, we captured the contemporary issues of open data in IS research and provide recommendations to advance open data in the IS discipline. From the literature and the workshop, it is clear that open data is an important component of distributed scientific collaboration and the dissemination and appropriation of the knowledge created through these collaborations. Through our analysis of current discussions on open data in IS literature, we identified two main clusters of themes that centered on topics around motivation and implementation. Discussions at the workshop confirmed the themes identified in

the literature review and added a new theme to the discussion. This new theme (Ethics) was retained as it is particularly pertinent to open data properties and the preparation involved for data to be shared openly. Table 2 provides an overview of those themes.

Table 2. Current Themes in the Open Data Discussion in the IS Discipline

Themes		Short Description
Motivation	Mandated Sharing	Open data is mandated by funding agencies and partner organizations.
	Benefits to the Research Process	Open data advances sciences, allowing studies to be replicated and confirmed or disproved.
	Extending the Life of Research Data	Open data is available to research beyond the initial research project and makes other investigations more substantial.
	Career Impact	Open data has an impact on researchers' careers. A positive impact is that researchers benefit from open data, especially when funding for data collection is scarce. A negative impact is evident in the case of researchers with data who have little incentive to spend the effort required for sharing data.
Implementation	Governance	Open data governance is required and emerges from discussions of ownership, access and usage rights, responsibilities, maintenance of dataset, and other issues.
	Socio-technical System	Open data requires an infrastructure of tools, people, and processes, creating new institutions and requirements for research projects.
	Standards	Open data standards for describing, representing, labeling, and storing data creates value for data users while putting requirements on data creators.
	Data Quality	Open data quality is a requirement for quality research and mandates that data was generated, cleaned, and prepared using rigorous and documented methods and tools.
	Ethics	Open data has ethical concerns for research participants and researchers due to overlap in source materials, new tools, and new methods such as machine learning, triangulation, and de-anonymization.

5.1 Identification of Next Steps for IS: Tackling Sharing Motivation Barriers

While the case for open data is strong and most of the discussion in IS is positive, the benefits of open data will not be realizable unless stakeholders lend their support. Many institutions still need to be convinced to see value in publishing and sharing data and reward such activity. Further, the data sharing benefits for researchers' careers need to be clarified and promoted more widely (e.g., by standardizing and recognizing data citations). Researchers should also seek to educate commercial organizations about the benefits of opening data, especially when doing so could demonstrate environment or social impact (Sayogo et al., 2014). And finally, data sharing motives are often subject to restricted research time, limiting appraisal priorities (tenure/promotions), and often well-established but potentially inflexible resource conventions in research active institutions. In order to encourage more data sharing, several institutional practices and researcher-specific award, performance, and appraisal schemes would have to be updated to include and recognize open data-related efforts.

A number of the themes listed in Table 2 may explain drivers of data sharing and actual data sharing practices. For example, at present is it unclear to what extent extending the life of data will generate positive outcomes for researchers. Similarly, the principle of data governance and data quality has implications for the way research is practiced. Data often requires additional effort to prepare it to a standard of quality, detail, and transparency to make it accessible and usable by others. Expenditures for such activities are not standard provisos in most organizations.

Not sharing data is also influenced by the kind of performance criteria set by funders and organizations where researchers work. As a result, we can expect that many stakeholders will be slow to support and adopt their current procedures in recognition of open data efforts. In these cases, a more collaborative and wider effort of IS researchers may be more effective in encouraging change in their institutions (see more details in Section 5.2). The role of a Dean, as well as members of the executive, interview, or tenure review board may have to adapt. In an increasingly time-constrained environment with continually increasing internal and external pressures, the allocation of work, review of performance, and recruitment are all significantly challenging tasks in light of open data.

In many cases, institutions and funders (e.g., National Institutions of Health, see Piwowar, 2011) already have procedures in place for a number of legal, insurance, and risk management purposes. By engaging a collaborative effort with such groups IS researchers may be able to address and gain recognition for open

Table 3. Actionable Advice for Institutions and Researchers for Open Data Implementation

Motivation: / Implementation:	Mandated Sharing	Benefits to the Research Process	Extending the Life of Research Data	Career Impact
Governance	***Institutions:*** • Funders establish universal sharing requirements • Publishers support universal sharing requirements ***Researchers:*** • Be courteous about other researchers' demands and negotiate governance on equal terms	***Institutions:*** • Establish best practices and governance for open data in research ***Researchers:*** • Apply best practices for research project governance	***Institutions:*** • Establish standard licenses to simplify the use of data ***Researchers:*** • Release data under an open license	***Institutions:*** • Establish universal and harmonized recognition guidelines for researchers of all disciplines ***Researchers:*** • Give credit to data creators
Socio-technical System	***Institutions:*** • Support compliance through university libraries and repositories ***Researchers:*** • Become familiar with available socio-technical systems	***Institutions:*** • Host shared repositories for research-in-progress data ***Researchers:*** • Access and utilize open data	***Institutions:*** • Provide repositories for long term storage with easy discovery and access for all ***Researchers:*** • Access and utilize open data outside of discipline	***Institutions:*** • Create simple tools for sharing and referencing datasets ***Researchers:*** • Adopt open data tools for lower barrier to releasing and using open data
Standards	***Institutions:*** • Create common standards for both sharing and reporting to funders ***Researchers:*** • Have a data dissemination plan consistent with standards • Participate in standardization issues	***Institutions:*** • Establish standard procedures for replicating research results based on open data ***Researchers:*** • Utilize open data standards, allow for replication of findings	***Institutions:*** • Establish descriptions and documentation for data structures, provenance, tools, and methods ***Researchers:*** • Utilize open data standards to establish provenance	***Institutions:*** • Develop standards across disciplines and countries ***Researchers:*** • Reduce time for sharing data by using standard tools and standard data formats
Data Quality	***Institutions:*** • Provide institutional support for sanitizing data across disciplines ***Researchers:*** • Go beyond the minimum required data sharing to ensure highest data quality	***Institutions:*** • Reward replication of research studies that release open data ***Researchers:*** • Enable new studies by providing quality open data	***Institutions:*** • Hire data managers to maintain open data and curate it over time ***Researchers:*** • Let a data manger maintain dataset as resources permit	***Institutions:*** • Implement a peer-review process for publishing datasets ***Researchers:*** • Use and create quality open data for research
Ethics	***Institutions:*** • Establish data protection, privacy, and author rights as part of data sharing ***Researchers:*** • Engage in risk assessment prior to sharing data, even in the case of mandated sharing	***Institutions:*** • Establish rules on de-anonymizing • Add disclaimers about source of data to reduce privacy risks • Have mechanisms in place to identify and prevent potential ethical concerns before these arise ***Researchers:*** • Establish best practices and inform risk management policy	***Institutions:*** • Establish rules for open data re-use • Update and harmonize Institutional Review Board guidelines ***Researchers:*** • Establish common rules on ethics to lower barriers for using open data	***Institutions:*** • Update and harmonize code of ethics in all disciplines regarding open data • Require ethics training as a key requirement for cross-institutional collaborations ***Researchers:*** • Follow best practices to reduce personal risk • Follow ethics procedures to raise contributor confidence

data-specific issues, raising the awareness of open data research within their institution. Such activities may enable IS researchers to strategically connect their concerns to those of their institutions and convince them of the broader benefits. Institutional support for and recognition of IS research on open data from public institutions is more likely when IS researchers work with institutional leaders regarding the benefit of their work for the public, thus helping those publicly funded organizations to demonstrate commitment to, social impact on, and engagement with the wider community. Engagement with ethics and other professional bodies and the public at large may be the first steps to gain institutional approval and support for open data research. Table 3 outlines actionable advice for institutions and researchers that want to adopt and promote open data.

5.2 The Potential of IS to be a Leading Discipline in the Open Data Movement

IS has the opportunity to play a major role in how open data are collected, implemented, and analyzed, not just within IS but across disciplines. A number of researchers in IS (including several authors of this paper) are already actively engaged in the design and implementation of open data repositories, knowledge exchanges, and the design of guidelines for open data management. For IS researchers who are interested in working with open data, we would like to direct their attention to the existing discussions and resources that are available within IS but also other disciplines. Our list of references provides one such starting point. In addition, much can be learned from the debate on big data and IS (Abbasi, Sarker, & Chiang, 2016; Agarwal & Dhar, 2014). A number of books in this area (e.g., Borgman, 2015; Kitchin, 2014) may further support research and curriculum developments and stimulate the engagement with the new challenges that arise for open data both in the field, research centers and in the classroom. Debates with researchers in other fields can help identify concerns, standards (Kansa, Kansa, Burton, & Stankowski, 2010), and resources that are already available in IS and related disciplines, as demonstrated in the list of themes identified based on the literature review and discussions captured in the workshop.

Researchers can shape the conversation around open data and actively use new tools and analytics to make the most use of open data. IS researchers are well-positioned with the necessary skills and knowledge to employ open data effectively. Technological means to analyze open data are pioneered in many different fields outside IS – providing new opportunities for collaboration, testing, and potential starting points for optimization. By embracing interdisciplinary collaborations, IS researchers have the opportunity to apply their knowledge of systems, human-computer interaction, and analytics in other domains traditionally outside the IS field. For example, work is already being done on metadata in healthcare (Dugas et al., 2015) and other work has considered the problem of expected data integration problems experienced as disincentive to data sharing among government agencies (Peled, 2011). Simón and colleagues (2014) note that in line with open data, information professionals need to take on new roles to manage metadata, address questions of licensing, and implement new applications. As such, building joint data repositories focused on open data with healthcare providers, public administrations, and libraries may be distinct opportunities for collaborations.

To advance, the discussion of research opportunities (Abbasi et al., 2016; Agarwal & Dhar, 2014) needs to move towards the development of design and funding strategies to support research programs and inter-disciplinary multi-national research collaborations with the capability to connect and respond to the interests of the public, private, and business stakeholders. The themes identified in the paper can serve as starting points for specific research programs.

The new research programs and collaborations may be able to tackle a slew of projects and issues in this area. For example, the study of data accidents deserves more attention, as does the study of the outcomes of leaked data as a means to identify preventive measures. The question of data integration problems (Peled, 2011), accountability and ethics is one that will have to go hand in hand with open data research and the use of open data repositories. A systematic review of barriers as conducted by van Panhuis and colleagues (2014) may be important here. These authors identified technical, motivational, economic, political legal and ethical barriers to data sharing in public health. Barry and Bannister (2014) conducted a similar research project with Irish government officials. They identified concerns about risks (including abuse and fraud), cultural, and administrative barriers (including security) (Barry & Bannister, 2014). They shared the same concerns as van Panhuis and colleagues (2014) regarding the legal and economic barriers. A similar investigation may be helpful to dismantle these barriers in relation to IS work and the main concerns of important collaborators in open data research conducted by IS researchers.

Future research could examine the societal, institutional, and research-specific benefits, as well as drawbacks of open data sharing to assess the extent to which the scientific benefits expected by the

OECD (2007) and other proponents of open data have materialized and paid off the public investment. Weerakkody and colleagues (2017) called for work to evaluate the performance of open data repositories, acceptance, use, and access of government information by citizens (Yannoukakou & Araka, 2014). Picking up such lines of enquiry appears to be both timely and appropriate, given that many open data repositories need public commitment to operate and share data.

Further, it could prove important to identify best practices, tools, and knowledge of regulatory guidelines to ensure success of such open data research programs. Schulte and colleagues (2016) have considered aspects such as data-intensive transportation and grid systems, and outlined suggestions for open data practices, provisions, and administration in research. Further knowledge exchanges and collaboration within IS and across different disciplines may prove essential in this process. As became apparent in the workshop, rapid and often unregulated data accumulation can have unintended consequences for the quality and meaningfulness of data. Additionally, new analytical tools can generate new problems (e.g., de-anonymizing algorithms) by raising new privacy and security concerns among stakeholders (Wood, O'Brien, & Gasser, 2016). Direct outcomes of these developments require the need for more transparency and accountability in how data are handled and used by IS researchers. Indirect effects due to data sharing may also need to be considered more carefully. Moreover, the open data movement as fostered by public data sharing also raises new concerns about societal impact of what we learn and the contribution of our findings for the public good. Such work is also likely to evoke political, legal, and risk management issues. Dealing with these challenges may require professional bodies to take action by providing guidance or regulation for open data practices, potentially in cooperation with other professional bodies.

6 Conclusion

This report aims at facilitating a discussion of open data in the IS discipline. The goal is achieved in three steps. First, the report includes a summary of the recent discussion around open data in the IS discipline. Second, the report provides voices from a pre-ICIS 2016 workshop that enhances the discussion in open data. Detailed presentation and discussion summaries found in the appendix provide questions to be addressed in future open data discussion. Third, the report outlines a roadmap with practical suggestions for IS researchers who are interested in establishing open data in their discipline and want to continue the discussion. Insight from these activities were subsequently distilled into a number of actionable suggestions in the hope that this report will trigger a dynamic and productive conversation within the wider IS discipline. As we note, open data has the potential to benefit individual scientists, collaborative research groups, research institutions, and society as a whole. While many obstacles remain that thwart open data sharing and collaborations, our goal is to provide a starting point for a strategic discussion on how IS can approach, tackle, and become a leading discipline in the open data movement.

Acknowledgments

The authors thank the Open Collaboration Data Factories project for organizing the workshop and fostering a lively discussion during the workshop. We further thank Gregory R. Madey and Kevin Whelan from the University of Notre Dame, IN, USA for hosting the workshop at the O'Connell House in Dublin. We thank all participants of the workshop for contributing to the discussion, sharing ideas, and offering advice.

The workshop was supported by funds and organizational support from the NSF (SAVI Project: #1449188; #1449209), The TOTO Project, Lero – The Irish Software Research Centre, IAIS – Irish Chapter of the Association for Information Systems, the University of Notre Dame, the AIS SIGOPEN group, and the Open Collaboration Data Factories.

The first two authors equally share 70% of the work. The presenters and other authors are listed in alphabetical order.

ORCID

"ORCID's vision is a world where all who participate in research, scholarship, and innovation are uniquely identified and connected to their contributions across disciplines, borders, and time" (https://orcid.org/about/what-is-orcid/mission). Initiatives with publishers are establishing ORCID in the review and publication process. To date, researchers can voluntarily provide information. Once a year, ORCID releases their dataset as open data under the CC0 license (i.e. no copyright reserved). Since this paper is about the future path of open data, we show support for the initiative by providing our ORCID iDs:

Georg J.P. Link https://orcid.org/0000-0001-6769-7867

Kevin Lumbard https://orcid.org/0000-0001-9306-3040

Kieran Conboy https://orcid.org/0000-0001-8260-4075

Joseph Feller https://orcid.org/0000-0001-9335-4542

Jordana George https://orcid.org/0000-0003-3491-9448

Matt Germonprez https://orcid.org/0000-0003-2326-5901

Debora Jeske https://orcid.org/0000-0002-4779-9345

Gaye Kiely https://orcid.org/0000-0001-6407-8868

Matt Willis https://orcid.org/0000-0002-9120-7319

References

Abbasi, A., Sarker, S., & Chiang, R. (2016). Big data research in information systems: Toward an inclusive research agenda. *Journal of the Association for Information Systems*, *17*(2), i–xxxii.

Agarwal, R., & Dhar, V. (2014). Editorial—Big data, data science, and analytics: The opportunity and challenge for is research. *Information Systems Research*, *25*(3), 443–448. https://doi.org/10.1287/isre.2014.0546

Barry, E., & Bannister, F. (2014). Barriers to open data release: A view from the top. *Information Polity: The International Journal of Government & Democracy in the Information Age*, *19*(1/2), 129–152. https://doi.org/10.3233/IP-140327

Benson, D. A., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2015). GenBank. *Nucleic Acids Research*, *43*(D1), D30-35. https://doi.org/10.1093/nar/gku1216

Berlin Declaration. (2003, October 22). Berlin declaration on open access to knowledge in the sciences and humanities. Retrieved from http://openaccess.mpg.de/67605/berlin_declaration_engl.pdf

Biljon, J., Pottas, A., Lehong, S., & Platz, M. (2016). Content category selection towards a maturity matrix for ICT4D knowledge sharing platforms. *CONF-IRM 2016 Proceedings*.

Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world*. Cambridge, Massachusetts: The MIT Press.

Bosco, F. A., Steel, P., & McDaniel, M. A. (2014). The "Big Science" revolution in management: Possibilities, technology, and applications. *Academy of Management Proceedings*, *2014*(1). https://doi.org/10.5465/AMBPP.2014.16949symposium

Budapest Declaration. (2002, February 14). Budapest open access initiative. Retrieved from http://www.budapestopenaccessinitiative.org/read

Cohen, J. P., & Lo, H. Z. (2014). Academic torrents: A community-maintained distributed repository. In *Proceedings of the 2014 Annual Conference on Extreme Science and Engineering Discovery Environment* (p. 2:1–2:2). New York, NY, USA: ACM. https://doi.org/10.1145/2616498.2616528

Creswell, J. W. (2013). *Qualitative inquiry and research design: Choosing among five approaches* (3rd ed). Los Angeles: SAGE Publications.

Curty, R., Yoon, A., Jeng, W., & Qin, J. (2016). Untangling data sharing and reuse in social sciences. In *Proceedings of the 79th ASIS&T Annual Meeting: Creating Knowledge, Enhancing Lives Through Information & Technology*. Silver Springs, MD, USA: American Society for Information Science.

Czarnitzki, D., Grimpe, C., & Toole, A. (2014). Delay and secrecy: Does industry sponsorship jeopardize disclosure of academic research? *Academy of Management Proceedings*, *2014*(1). https://doi.org/10.5465/AMBPP.2014.16838abstract

Donia, M., Jimenez, A., & Shah, G. (2015). Research crowdsourcing, data sharing, and large-scale collaboration. *Academy of Management Proceedings*, *2015*(1). https://doi.org/10.5465/AMBPP.2015.18361symposium

Dugas, M., Jöckel, K. H., Friede, T., Gefeller, O., Kieser, M., Marschollek, M., … others. (2015). Memorandum "Open Metadata." *Methods of Information in Medicine*, *54*(4), 376–378.

European Commission. (2016). *H2020 Programme: Guidelines on FAIR Data Management in Horizon 2020* (Version 3.0). Retrieved from http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

Holdron, J. P. (2013, February 22). Increasing access to the results of federally funded scientific research. Executive Office of the President Office of Science and Technology Policy. Retrieved from https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

Jeng, W., He, D., & Oh, J. S. (2016). Toward a conceptual framework for data sharing practices in social sciences: A profile approach. In *Proceedings of the 79th ASIS&T Annual Meeting: Creating Knowledge, Enhancing Lives Through Information & Technology*. Silver Springs, MD, USA: American Society for Information Science.

Johnson, J. A. (2014). From open data to information justice. *Ethics and Information Technology*, *16*(4), 263–274. https://doi.org/10.1007/s10676-014-9351-8

Kansa, E. C., Kansa, S. W., Burton, M. M., & Stankowski, C. (2010). Googling the grey: Open data, web services, and semantics. *Archaeologies*, *6*(2), 301–326. https://doi.org/10.1007/s11759-010-9146-4

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures & their consequences*. Los Angeles, California: SAGE Publications.

Laine, S., Lee, C., & Nieminen, M. (2015). Transparent data supply for open information production processes. *ECIS 2015 Completed Research Papers*. https://doi.org/10.18151/7217404

NSF. (2016a). Open Data at NSF. Retrieved December 18, 2016, from https://www.nsf.gov/data/

NSF. (2016b). Public Access: Frequently Asked Questions: What is NSF's public access policy?. Retrieved from https://www.nsf.gov/pubs/2016/nsf16009/nsf16009.jsp#q1

OECD. (2004, January 30). Science, Technology and Innovation for the 21st Century. Meeting of the OECD Committee for Scientific and Technological Policy at Ministerial Level, 29-30 January 2004 - Final Communique. Retrieved from https://www.oecd.org/science/sci-tech/sciencetechnologyandinnovationforthe21stcenturymeetingoftheoecdcommitteeforscientificandtechnologicalpolicyatministeriallevel29-30january2004-finalcommunique.htm

OECD. (2007, April). OECD Principles and Guidelines for Access to Research Data from Public Funding. Retrieved from https://www.oecd.org/science/sci-tech/38500813.pdf

O'Hara, K. (2012). Transparency, open data and trust in government: Shaping the infosphere. In *Proceedings of the 4th Annual ACM Web Science Conference* (pp. 223–232). New York, NY, USA: ACM. https://doi.org/10.1145/2380718.2380747

Østerlund, C. (2008). The materiality of communicative practices. *Scandinavian Journal of Information Systems*, *20*(1), 4.

Paik, L., & Binz-Scharf, M. C. (2014). Democratization or reflection: The paradox of databases' influences on knowledge production. *Academy of Management Proceedings*, *2014*(1). https://doi.org/10.5465/AMBPP.2014.17513abstract

Pasquetto, I. V., Sands, A. E., & Borgman, C. L. (2015). Exploring openness in data and science: What is "Open," to whom, when, and why? In *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*. Silver Springs, MD, USA: American Society for Information Science.

Peled, A. (2011). When transparency and collaboration collide: The USA Open Data program. *Journal of the American Society for Information Science and Technology*, *62*(11), 2085–2094. https://doi.org/10.1002/asi.21622

Perkmann, M., & Schildt, H. (2014). Open data in industrial R&D: Organizing open collaboration between firms and public science. *Academy of Management Proceedings*, *2014*(1). https://doi.org/10.5465/AMBPP.2014.16350abstract

Piwowar, H. A. (2011). Who shares? Who doesn't? Factors associated with openly archiving raw research data. *PLOS ONE*, *6*(7): e18657. https://doi.org/10.1371/journal.pone.0018657

Pollock, D. (2016). Understanding scientific data sharing outside of the academy. In *Proceedings of the 79th ASIS&T Annual Meeting: Creating Knowledge, Enhancing Lives Through Information & Technology*. Silver Springs, MD, USA: American Society for Information Science.

Ribes, D., & Polk, J. (2014). Flexibility relative to what? Change to research infrastructure. *Journal of the Association for Information Systems*, *15*(5), 287–305.

Robles, G., Arjona Reina, L., Serebrenik, A., Vasilescu, B., & González-Barahona, J. M. (2014). FLOSS 2013: A survey dataset about free software contributors: Challenges for curating, sharing, and combining. In *Proceedings of the 11th Working Conference on Mining Software Repositories* (pp. 396–399). New York, NY, USA: ACM. https://doi.org/10.1145/2597073.2597129

Sayogo, D. S., Zhang, J., Pardo, T. A., Tayi, G. K., Hrdinova, J., Andersen, D. F., & Luna-Reyes, L. F. (2014). Going beyond open data: Challenges and motivations for smart disclosure in ethical consumption. *Journal of Theoretical and Applied Electronic Commerce Research; Curicó*, *9*(2), 1–16.

Schulte, F., Chunpir, H. I., & Voß, S. (2016). Open data evolution in information systems research: Considering cases of data-intensive transportation and grid systems. In *Design, User Experience, and Usability: Technological Contexts* (pp. 193–201). Springer, Cham. https://doi.org/10.1007/978-3-319-40406-6_18

Simón, L. F. R., Avilés, R. A., Botezan, I., Gastaminza, F. del V., & Serrano, S. C. (2014). Open data as universal service. New perspectives in the information profession. *Procedia - Social and Behavioral Sciences*, *147*, 126–132. https://doi.org/10.1016/j.sbspro.2014.07.128

Tempini, N. (2016). Science through the "Golden Security Triangle": Information security and data journeys in data-intensive biomedicine. *ICIS 2016 Proceedings*.

Thompson, C. A. (2015). Building data expertise into research institutions: Preliminary results. In *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*. Silver Springs, MD, USA: American Society for Information Science.

van den Broek, T., & Veenstra, A. F. van. (2015). Modes of governance in inter-organizational data collaborations. *ECIS 2015 Completed Research Papers*. https://doi.org/10.18151/7217509

van Panhuis, W. G., Paul, P., Emerson, C., Grefenstette, J., Wilder, R., Herbst, A. J., … Burke, D. S. (2014). A systematic review of barriers to data sharing in public health. *BMC Public Health; London*, *14*(1144). https://doi.org/10.1186/1471-2458-14-1144

Vassilakopoulou, P., Skorve, E., & Aanestad, M. (2016). A commons perspective on genetic data governance: The case of BRCA data. *ECIS 2016 Proceedings*.

Vitak, J., Shilton, K., & Ashktorab, Z. (2016). Beyond the Belmont Principles: Ethical challenges, practices, and beliefs in the online data research community. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (pp. 941–953). New York, NY, USA: ACM. https://doi.org/10.1145/2818048.2820078

Weerakkody, V., Irani, Z., Kapoor, K., Sivarajah, U., & Dwivedi, Y. K. (2017). Open data and its usability: An empirical view from the Citizen's perspective. *Information Systems Frontiers, 19*(2), 285-300. https://doi.org/10.1007/s10796-016-9679-1

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*. https://doi.org/10.1038/sdata.2016.18

Wood, A., O'Brien, D., & Gasser, U. (2016, September 26). Privacy and open data research briefing. Berkman Klein Center Research Publication No. 2016-16. Retrieved from https://ssrn.com/abstract=2842816

Yannoukakou, A., & Araka, I. (2014). Access to government information: Right to information and open government data synergy. *Procedia - Social and Behavioral Sciences*, *147*, 332–340. https://doi.org/10.1016/j.sbspro.2014.07.107

All URLs in this document were checked on April 2nd, 2017.

Appendix A: Pictures from the Workshop

Pictures are an essential part of any workshop report.

Figure 1. Group photo of most workshop participants in front of the O'Connell House (color online)

From left, front row: Sean Goggins, Ann Barcomb, Georg Link, Jordana George, Jeff Parsons; middle row: Souma Ray, Kristen Schuster, Kevin Lumbard, Gaye Kiely, Zeena Feldman; back row: Matt Germonprez, Joseph Feller, Kieran Conboy, Fergal Carton, Michael Feldman, Brian Fitzgerald, Matt Willis, Greg Madey.

Figure 2. Table discussion after each presentation for discussing and collecting questions (color online)

From left, table in front: Kevin Lumbard, Matt Willis, Greg Madey, Fergal Carton; Table in middle: Kieran Conboy, Zeena Feldman, Kristen Schuster, Michael Feldman; Left table in rear: Joseph Feller, Jeff Parsons, Souma Ray, Ann Barcomb; Right table in rear: Jordana George, Matt Germonprez, Sean Goggins.

Appendix B: Presentations at the Workshop

The account of the presentation topics has three goals: to summarize the content of the presentations at the workshop, to present questions asked during the discussions to generate thoughts and ideas for future research, and to identify the occurrence of themes in presentations and questions to provide a larger perspective on the discussion of open data. The order of summaries below mirrors the order of presentations at the workshop.

B.1 Jordana George: Open Data Sharing and Networking Platforms with an Eye on Data Marketplaces

Jordana George has a unique opportunity to gain insight into the creation of a data repository combined with social collaboration by a newly found benefit corporation, data.world (https://data.world/). A benefit corporation is a for-profit organization with a social mission. The objective of data.world is to build and establish a social, collaborative open data repository where people can find, use, store, and work together on myriad datasets. The data repository can be used by anyone to publish datasets and network with users. The site offers dataset descriptions, queries, search functionality, data visualization, personal profiles, and following of other users or datasets. George experiences the open data story firsthand through full access to this organization and its employees through biweekly field visits and access to archives of digital communications such as documents, emails, and chat from the inception of the company. George is interested in questions that arise from the development of a social data repository including: Why do people choose to share or not share data? What are the differences between required data sharing and volunteer data sharing? What are the challenges to sharing? How can the data be presented in a useful way to users? How can datasets be described and searched uniformly across the broad spectrum of use cases? What are the aspects to operationalizing open data? How about quality?

Table 4. Audience questions and identified themes for George's presentation

Audience Questions	• How can a data repository track the use of downloaded data by users and potentially make a claim about the social impact from the open data? • How can the trust in the data broker (repository provider) be maintained independently from the open data uploaded by third parties? • How can a researcher show the benefit of open data when acquiring funding? • How can open data related to human data be made safe (i.e. protect the weakest e.g. refugees)? • How can one protect research participants from de-anonymization in open data? • Can a data repository ensure that data is useful and used for social benefit? • How can one overcome the secretive tendencies of corporations and convince them that the benefits of open data outweigh the benefits of keeping data secret as a competitive advantage? • What standards for data and metadata should be used and how do they affect the willingness to share data?
Identified Themes	• Mandated Sharing • Benefits to the Research Process • Career Impact • Socio-technical Systems • Data Quality • Ethics* * A new theme emerged that was not found in the AIS literature: Because the organization under study is a benefit corporation, much discussion aligned with the new theme ethics especially as it relates to extending the life of data and ties into data governance

B.2 Debora Jeske: Collaborative Research that Crosses Traditional Boundaries of Disciplines

In her presentation, Debora Jeske focused on both opportunities and challenges in interdisciplinary collaboration involving, among others, open data efforts. Interdisciplinary collaborations carry many advantages, specifically those in terms of the use of analytical tools, theory and existing ethics procedures in some disciplines (ethical approval is becoming a standard feature for many funders and publishers). The decision to share data publicly, or use of open data across disciplines and countries, may raise

ethical concerns (and trigger legal, insurance, and risk management queries). This situation is often exacerbated when the ethical standards for collecting data vary across disciplines and countries, making it difficult to collaborate, to conduct research, and to use open data internationally (see Vitak et al., 2016). Further difficulties arise in terms of incompatible archival and publishing strategies (some disciplines and countries have journal lists, others use impact factors, and in some cases funders demand open access publications; see also Piwowar, 2011). Misunderstandings can arise due to different conceptualizations of data creators, users, purposeful and agreed (such as consent-given) use of data. Specific rules of funding agencies and non-disclosure agreements between organizations and researchers may also limit or even prevent the sharing of data (see also Sayogo et al., 2014). Overall, however, a consensus emerges across disciplines regarding the need for transparency, ethical deliberation and the need for caution when sharing results (e.g., Vitak et al., 2016).

Table 5. Audience questions and identified themes for Jeske's presentation

Audience Questions	• If one ignores the challenges to open data, what does inter-disciplinary collaboration enable? • Relating to who owns, manages, and uses data, can the use of open data be made visible to control and foster ethical use? • How can tensions between commercial use of data and public benefit of shared data be resolved? • How can one deal with biases in open datasets? • How can ethics be reflected in algorithms that automate much of today's lives?
Identified Themes	• Career Benefits or Impediments for Researchers • Socio-technical Systems • Standards • Data Quality • Ethics

B.3 Michael Feldman: Empowering Enthusiasts to Conduct Collaborative Analysis

Evidently, research through crowdsourcing and citizen science can provide a source of unexpected discoveries, leveraged research efforts, and skill development for interested participants. Moreover, Michael Feldman points out that citizen science phenomenon can be further extended by involving crowds not only in simple tasks such as classification, but by involvement in data analysis research. This expectation is timely due to the growing pool of open data that can be of high scientific interest. Unfortunately, good data scientists, who can make sense of data, are very rare and are not available to the amateur scientific community. Feldman explored the question whether non-experts can be involved in data analysis research and how to support non-experts in this endeavor. Feldman proposes to break data analytic related tasks into small coding tasks that can be completed by non-experts with very basic coding skills. In such scenario, participants are remotely supervised by an expert throughout the process and together achieve the desired outcome. A number of studies showed that data cleaning, a bottleneck activity of data analysis, is the most promising task to be outsourced where the crowd and experts produced a comparable quality. Future research aims at improving the platform used for collaborative data analysis, investigating what characteristics and requirements tools for collaborative data analysis must meet, and finding an economical break-even point where the crowd results outweigh the task of monitoring the crowd results.

Table 6. Audience questions and identified themes for Feldman's presentation

Audience Questions	• How can an expert be defined, especially considering the distinction between method and context issues? • What quality checkpoints could be built into the process, to not only check the quality at the end of the process? • How can not only collaborative data analysis be technically enabled but users empowered to it – how to build a learning process? • How can the diverse expertise of crowd workers be managed? • How can people be motivated to participate? How can people be encouraged to play with the data and improve their skills? • How can the narrow scientific focus of a research study be balanced with the broad social impact that citizen science aims for? • How does outsourcing parts of the data analysis affect research because data cleaning and other data operations have been a source of thought and ideas for scientists?

Identified Themes	• Benefits to the Research Process • Extending the Life of Research Data • Career Impact • Socio-technical System • Data Quality

B.4 Gaye Kiely: Practical Experience of the Barriers to Effective Collaborative Virtual Work

Gaye Kiely's focus is on the relationship of open data with distributed work. Drawing on her doctoral research on global distributed team coordination and experience of working as a software quality engineer in a distributed team, she examines the issues associated with remote collaboration. Distributed collaboration, especially in software development, has received much attention in previous research (particularly, the impact of geographical distance, temporal distance, cultural diversity, and team trust). While distributed software development teams have different goals to distributed scientific research teams, they share common challenges such as coordination, knowledge sharing, sustainability, and communication. Collaborative research teams need to be formed with consideration for the qualities and skills of the team members and specific open data challenges. The research team needs to be managed to ensure continuous operation and future sustainability. Issues of fatigue, isolation, and negative effects must be addressed, especially when collaborative research teams involve volunteers. Kiely observes that modern collaboration tools appear to bridge the gaps of geographical and temporal distances but do not fully address more "fuzzy" issues such as cultural diversity, and engendering (and maintaining) team cohesion. Kiely calls upon the research community to draw on existing research in global distributed team work, in order to identify solutions for collaborative, open data research teams with respect to coordination and sustainability.

Table 7. Audience questions and identified themes for Kiely's presentation

Audience Questions	• How can one study the end of a project when people walk away (think ghost town on the internet)? • How can momentum be maintained when tasks in a collaboration are divided into small tasks? • What is more easy to maintain, the momentum of a collaborative project or its long-term vision and goal? What other processes, e.g. peer production, face issues of motivation that can be learned from? • How can individual work be aggregated into a coherent end-product? • What is more important to a sustainable team, the collaboration processes or the combination of people? • How can the team selection findings from the well-researched human resources literature be implemented in online crowds and collaborative research projects?
Identified Themes	• Governance • Socio-technical System

B.5 Matt Willis: Distributed Scientific Teams and Conducting Collaborative Science

Matt Willis' interest is in improving the design of collaborative science by studying the collaborative structures distributed teams create through their work practices. The approach used in his research views documents as a data source to better understand distributed team dynamics and how these document structures support distributed scientific teams. With a focus on social scientists, his team is studying the socio-technical systems of small-scale collaborations. They theorize documents as anything written down that contains meaning for the group such as emails, whiteboards, notebooks, scraps of paper, drafts of manuscripts, documents from repositories, phone records, data from social media platforms, and trace data from collaboration platforms such as GitHub, Dropbox, figshare, and SharePoint. As socio-material, these documents are part of social experiences and are imbued with meanings due to being rooted in scientists' daily practices (Østerlund, 2008). This project is ongoing but initial findings suggest issues arise from misaligned practices around document version control and naming conventions. Additional conflicts to distributed collaborations arise from coordinating software choices and what platforms the collaboration is to use. For example, some group members may use Google Docs for writing and others use Microsoft Word. The differences in these software packages and platforms create various problems and

incompatibilities in collaborative groups. The research also found that lack of documentation creates problems with collaborative data analysis, particularly when qualitative data are concerned. Research teams using open data face the same issues and can benefit from considering naming conventions, dataset versioning rules, and documentation of tacit knowledge that cannot be easily understood without the use of a meeting especially when data are qualitative in nature or contain interpretive aspects. Willis then discussed ethical considerations for collecting social data and implications that led to not sharing the data. The prime ethical concern being how inextricable work and personal lives have become, making it impossible to look at work emails and documents without also understanding personal and private life.

Table 8. Audience questions and identified themes for Willis' presentation

Audience Questions	• How can informed consent be acquired without putting a burden on research participants? • How can the complete communication be collected into the research database? • How should non-work related communication be treated, e.g. emails with family? • Should the data collection be automatic or controlled by the research participant? • What value can researchers provide to the participants in return for sharing the data (e.g. metrics from data)? • How can best practices for collaborative research teams be distilled from the collected data?
Identified Themes	• Socio-technical System • Standards • Ethics

Appendix C: OCDX: A Metadata Initiative to Advance Open Data

As part of the workshop, participants were introduced to and then discussed the approaches being developed by the Open Community Data Exchange (OCDX), a metadata specification to describe research datasets (http://ocdx.io/). OCDX defines a bill of materials (OCDX document) that can be attached to computational social science research data. Kristen Schuster and Matt Germonprez presented and facilitated the discussion around the OCDX initiative. The OCDX document conveys information about the research data, such as when it was collected, where it was collected, where to acquire the dataset, what license it is published under, and who the original data creator is. OCDX documents can be automatically processed, searched, and stored, enabling researchers to make their datasets available and discoverable. The workshop discussion on OCDX resulted in a number of recommendations and concerns:

(1) Describing a dataset is a best practice amongst researchers for maintaining the usefulness of a dataset by documenting how, when, and where data was collected and how it was manipulated. The practice of maintaining a README file for each dataset could be standardized through the OCDX specification which can reduce barriers for releasing the dataset.

(2) Data licensing is an important consideration for sharing research. Creating a standard license list as a complementary product to the OCDX specification can simplify describing and choosing a license for open data.

(3) Providing a consistent way to reference a dataset (i.e. unique identifier such as DOI) provides a way to ensure standards and data quality (Sayogo et al., 2014). When a derivative dataset alters, combines, or enhances a dataset, it should reference the original datasets to provide a trail of origin (provenance information).

(4) Open data and their associated metadata need to be findable, accessible, interoperable, and reusable (FAIR) to satisfy the FAIR Guiding Principle (Wilkinson et al., 2016) adopted by the EU-Commission in 2016 (European Commission, 2016). This will increase the likelihood that open data is not just discoverable but used, while increasing the research and sharing benefits for funders and researchers.

About the Authors

Georg J.P. Link is a doctoral student at the University of Nebraska at Omaha. His current research interest is in open source communities. He holds a M.Sc. in Business Informatics from the Technische Universität Braunschweig, a Master of Business Administration with a concentration in Collaborative Science from the University of Nebraska at Omaha, and a B.A. in Business Administration from the WelfenAkademie, Braunschweig, Germany. His work appeared in the *Journal of Universal Computer Science*.

Kevin Lumbard is a doctoral student studying Information Technology at the University of Nebraska at Omaha. As a former web developer, he is experienced in developing and managing eCommerce applications and processes for small to medium sized companies. Kevin is interested in open collaboration communities, open source applications for small businesses, and project management in distributed environments. His current research interest includes the movement of individuals among organizations engaged in open source development, information sources used in open source design, and exposures of sharing and linking open data.

Kieran Conboy is a professor in information systems and leader of the Lero Irish software research centre at NUI Galway. He previously worked for Accenture Consulting and the University of New South Wales in Australia. He has worked with organisations such as Atlassian, Cisco Systems, Dell, Suncorp, and Ericsson, as well as many SMEs. Kieran has published over 180 articles in leading international journals and conferences including *Information Systems Research*, *European Journal of Information Systems*, *Journal of the Association for Information Systems,* and the *Communications of the Association for Information Systems*. A key focus of his current research is the examination of flow, agility, temporality and openness in organisations. He is on the board of the Irish Research Council, an editor of the *European Journal of Information Systems* and has chaired many international conferences in his field.

Michael Feldman is currently pursuing a doctoral studies in Computer Science at University of Zurich. He holds an M.Sc. in Industrial Engineering and Management from Ben Gurion University with a focus on Information Systems. In his recent research, Michael is studying the potential of crowdsourced collaborative data analysis. Specifically, he is fascinated with the idea of enabling diverse crowds with limited coding skills to participate in full, end-to-end data analysis projects. In his research, Michael draws from both organizational psychology as well as technical foundations.

Joseph Feller is Professor of Information Systems at the Cork University Business School, University College Cork, Ireland. His research focuses on how individuals, organizations, and societies can use IT to leverage collective intelligence, action, and resources. His work has appeared in journals such as *Information Systems Research*, *Journal of the Association of Information Systems*, *Journal of Strategic Information Systems*, *European Journal of Information Systems*, *Journal of Information Technology*, *Information and Organization*, and *Information Systems Journal*. He served as founding President of the Special Interest Group on Open Research and Practice (SIGOPEN) of the AIS from 2015-2016.

Jordana George is currently undertaking doctoral studies in Information Systems at Baylor University. She holds an MBA from Penn State and an MFA from the University of California at Davis. A former manager in client services, technical support, and general management at technology companies and educational institutions, she researches knowledge management, data management and open data, technology entrepreneurship, social entrepreneurship and benefit corporations, and social issues in technology and business.

Matt Germonprez is currently a faculty member at the University of Nebraska at Omaha. Prior to joining UNO, he was a faculty member at UW-Eau Claire, Case Western Reserve University, and a PhD student at the University of Colorado–Boulder. His research focuses on theory and method development and

investigation with particular focus on emerging and tailorable technologies. In particular, he explores how these new, user-centered technologies are designed and used in practice from the individual to the enterprise level. His work has been funded by the National Science Foundation and accepted in *MIS Quarterly*, *Journal of the Association for Information Systems*, *Information Systems Journal*, *Information & Organization*, and *Communications of the Association for Information Systems*.

Sean Goggins is an Associate Professor at the University of Missouri in the Computer Science department. He teaches, publishes, and conducts research on the uptake and use of information and communication technologies by small groups in medium to large scale sociotechnical systems; from Facebook, to online course systems. Sean conceptualizes "group informatics" as a methodological approach and ontology for making sense of the interactions between people in medium to large scale social computing environments. Sean spent 12 years as a software engineer and architect in industries ranging from medical devices to online publishing before pursuing his Ph.D. After four years at Drexel University in Philadelphia, he moved to Missouri in the fall of 2013, where he continued his work, and launched a new masters degree program in data science in the fall of 2016.

Debora Jeske is a lecturer in Work and Organisational Psychology at University College Cork, Ireland. Her research interests include psychology and technology at work, including e-HRM, ethics at work, information/ data management at work, Human Computer Interaction. Contact details: Dr. Debora Jeske, School of Applied Psychology, University College Cork, North Mall, Cork City, Ireland.

Gaye Kiely is a lecturer, researcher and co-director of BSc. Business Information Systems at University College Cork (UCC), Ireland. She holds a BSc in Business Information Systems (1998), an MSc (Research) on software lifecycles (2001), and a PhD on Global Virtual Team Coordination (2012). Her research interests have centered on global virtual software development teams, distributed collaboration, development methods and coordination. More recently, her research focus has turned to open data, collaborative work, and community sustainability.

Kristen Schuster is a lecturer in digital curation in the Department of Digital Humanities at King's College London. Her research and teaching focus on the role metadata plays in interdisciplinary data management projects.

Matt Willis is a researcher at the Oxford Internet Institute, University of Oxford. His current work looks at the use of automation technology in NHS England general practice services. Other work include socio-technical inquiry into personal health record use between patients and care teams. His research interests include patient-generated data, socio-technical systems in healthcare, computer supported cooperative work, social shaping of technology, and human-computer interaction in healthcare. He has been a researcher in academic, government, and private institutional settings including Sandia National Laboratories, the U.S. Department of Veterans Affairs, and several university affiliated research centres where he was a contributor to multiple grants from the National Science Foundation (NSF), National Institutes of Health (NIH), Defense Advanced Research Projects Agency (DARPA), and Intelligence Advanced Research Projects Activity (IARPA).