IJDC | *Peer-Reviewed Paper*

The State of Data Guidance in Journal Policies: A Case Study in Oncology

§

Deborah H. Charbonneau
School of Library and Information Science
Wayne State University

Joan E. Beaudoin
School of Library and Information Science
Wayne State University

Abstract

This article reports the results of a study examining the state of data guidance provided to authors by 50 oncology journals. The purpose of the study was the identification of data practices addressed in the journals' policies. While a number of studies have examined data sharing practices among researchers, little is known about how journals address data sharing. Thus, what was discovered through this study has practical implications for journal publishers, editors, and researchers. The findings indicate that journal publishers should provide more meaningful and comprehensive data guidance to prospective authors. More specifically, journal policies requiring data sharing should direct researchers to relevant data repositories, and offer better metadata consultation to strengthen existing journal policies. By providing adequate guidance for authors, and helping investigators to meet data sharing mandates, scholarly journal publishers can play a vital role in advancing access to research data.

Received 27 May 2015 | *Accepted* 25 November 2015

Correspondence should be addressed to Deborah H. Charbonneau, Ph.D., Assistant Professor, School of Library and Information Science, 106 Kresge Library, Wayne State University, Detroit, MI, USA 48202. Email: dcharbon@wayne.edu

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

The research landscape is changing dramatically with many disciplines facing "complex problems requiring new and creative uses of diverse data" (Parsons and Berman, 2013). The ability of researches to address these complex problems rests on their capacity to adequately collect, preserve, manage and reuse data. The recent data-oriented focus within the library and information science domain attests to the increasing awareness of the critical nature of these efforts. Scientific endeavours in the 21st century have been noted as relying more heavily on data rich and collaborative interactions than at other time in history (Tenopir et al., 2011). An acknowledgement of this situation can be seen in the development of several data preservation, management and sharing policies that have emerged in the United States and internationally in recent years (Sansone, 2010). In particular, numerous funding agencies require recipients to make research outputs accessible, including publications and research data (Charbonneau, 2013; Kozlowski, 2014). Yet, even with these mandates, researchers generally have not "concentrated on the organization, access, reuse, and preservation of data in their day-to-day research" (Wiley, 2014).

Ensuring the accessibility of research data is a critical consideration for scientific purposes. The benefits of sharing data include facilitating new scientific inquiry, promoting potential new uses of data, and encouraging the validation of research (Charbonneau, 2013). While what constitutes data sharing is interpreted differently across disciplines, Borgman (2012) defines data sharing broadly as "the release of research data for use by others." The capacity to share data beyond the informal channels of exchanging data via email attachments and storage media has led to the development of various policies and facilities. For example, a number of funding agencies have included data sharing mandates within their grant opportunities. Funders, publishers, societies and individual research groups have developed "tools, resources, and policies to encourage investigators to make their data publicly available" in an effort to capitalize on these advantages (Piwowar, 2011).

Indeed, journals play an important role in providing the guidance their author-researchers need to ensure research data are made accessible. Yet research that examines the state of data guidance provided to authors by journals is sparse. Further work is needed to better understand existing journal data policies for data sharing and dissemination so that best practices can be identified and further developed. As scientific research continues to increase in both size and scope, codified procedures for the collection, preservation, management and reuse of data have become imperative.

Background

Data Sharing and Reuse Issues

Various aspects of data sharing and data reuse among scientists have been investigated in several recent studies. Tenopir et al. (2011) report on some of the challenges of sharing data among scientists. Issues such as trust, lack of time and anticipated data sharing costs were cited as difficulties in sharing research data. In another study, a

concern about the privacy of human subjects' data was speculated to be a factor in determining a researcher's willingness and ability to share raw study data (Piwowar, 2011). Variation in data sharing behaviours was discovered among disciplines. Basic science researchers reported storing more digital data than other researchers (Akers and Doty, 2013). In addition, basic science researchers were the most familiar with funding agency requirements for data management plans, were most likely to share data with people outside of their research groups, and were more likely to deposit some or all of their data in data repositories or databanks (Akers and Doty, 2013). On the other hand, social scientists have been found to be less likely to report their data was shared or archived, and this group is followed by researchers in the arts and humanities. The latter group of researchers was found to be even less likely to share or archive their data (Akers and Doty, 2013). As these studies suggest, the existing literature has begun to provide insight into some of disciplinary differences and perceived challenges associated with sharing data among various researchers.

Within the realm of cancer research, Rolland and Lee (2013) comment that the "sharing and reuse of data is still complex and fraught with pitfalls." This statement suggests that within this domain clinical researchers face a number of challenges. Technical obstacles may include handling large amounts of data in varying formats, and the lack of uniformity in the data being collected (DeMartino and Larsen, 2013). Consequently, the reuse of data currently contained in various repositories and databases may be problematic for the oncology research community. This situation was discussed by Rolland and Lee (2013), who found that researchers may struggle to make sense of the data collected by others, even when investigators have direct access to all available study documentation. MacMillan (2014) further notes that insufficient technical infrastructure, systems that do not communicate with one another, and a lack of reward incentives can be additional barriers to sharing data. Thus, the extant literature has identified that standards for data sharing, issues of data reuse, and inadequate infrastructure to support data-intensive research are significant concerns.

Despite these challenges, Carpenter et al. (2012) argue that growing repositories of data should be leveraged to conduct cancer research. DeMartino and Larsen (2013) agree with this sentiment, noting that data should be "moved from silos" in order to enhance patient care. These authors clearly indicate the critical nature of data sharing in cancer research by stating that "[t]he oncology community must embrace the idea of collaboration to combine available sources of data and improve the current healthcare system" (DeMartino and Larsen, 2013).

Data Collaboration and Policy Issues

There is a growing understanding in the literature that data management requires the support of a number of individuals and institutions. Parkham and Doty (2012) assert that in order to be effective overall data management efforts "need to be a partnership among librarians, administrators, technologists, and researchers themselves." In addition to these key players in data management, more research is needed to explore the state of data guidance provided by journals in relation to funding agency requirements for access to research data. For example, the National Institutes of Health (NIH) has policies in place for funded researchers to share their research data supported by tax payer dollars (NIH, 2015). The Wellcome Trust also requires funded researchers to share their data and that data should be made available with as few restrictions as possible (Wellcome Trust, 2013). Both examples demonstrate policies developed by funding agencies aimed at increasing public access to digital scientific data.

MacMillan (2014) states that journals are "revising their data-handling policies and publication models to ensure that supplemental data can be validated, preserved, and indexed to facilitate discoverability and reuse by others." This suggests that there needs to be a concerted effort among multiple constituencies for the effective management of research data, and that journals publishers can play a critical role. One such effort to make research data more available for researchers to use can be seen in the number of journals that now request or require investigators share their data (Piwowar and Chapman, 2008). More recently the Public Library of Science (PLOS), an international non-profit supporting the open access to scientific publications, enacted a data sharing requirement. When authors submit a manuscript to any of the PLOS journals, the author must provide a "Data Availability Statement" describing their compliance with PLOS's data policy. This data availability statement is then published with the article (Bloom, Ganley & Winker, 2013). As a result of this step, PLOS's data policy attempts to foster scientific advances by "ensuring access to the underlying data [which] should be an intrinsic part of the scientific publishing process" (Bloom, Ganley & Winker, 2013).

Piwowar and Chapman (2008) examined journal data sharing policies for a single type of data – biological gene expression microarrays – and found "some mention of sharing publication-related data within their Instruction to Author statements." Yet the state of data guidance provided in journal policies to help authors share their data, and to also comply with funder requirements, remains incomplete. Studies of the journals' data guidance statements contribute to a richer understanding of the expansive research data landscape. As MacMillan (2014) states "the scholarly community is moving inexorably toward improved access to research data." While there is clearly a movement toward providing access to research data, how this will be achieved has yet to be determined. Lin and Strasser (2014) suggest that journal publishers can "help build a vibrant research ecosystem in which research data is publicly available for maximum reuse." An additional motivation for data sharing guidelines among scientific journals is the provision of publicly available data, which has been associated with a significant increase in citations (Dorsch, 2012; Piwowar et al., 2007). Accordingly, "data reuse, data sharing, and collaboration [will] become increasingly important to the conduct of scientific research" (Rolland and Lee, 2013).

To summarize, previously conducted research helps highlight several factors related to sharing and reusing research data (Joint Information Systems Committee [JISC], 2012). As research interest in, and funder mandates requiring, the deposit or sharing of research data continue to grow, the need for the development of journal policies aimed at facilitating access to research data is likely to increase. To address this gap in the existing research, the present study examines the state of data guidance provided to journal authors in oncology journals. The study also explores and identifies the data practices which are addressed among these journals.

Research Design

Sample

To examine the state of data guidance provided to authors, policy data were collected from the websites of 50 journals in oncology with the aim of performing a structured content analysis on the collected data. Journals in oncology were selected as a primary

focus of the present study because one of the authors is engaged in cancer-related research and a prospective author for these journals. The journals were selected using Thomson Reuter's Institute for Scientific Information (ISI). ISI was consulted to identify highly cited journals in the domain of oncology. The impact factor of the journals, as provided by ISI, was used as the primary selection criteria in the development of the list of the top-ranked 50 journals. This strategy, which has been used previously to help identify highly ranked journals in oncology for their inclusion in research studies (Kesselheim et al., 2012), was believed to offer several important advantages. As these journals are regarded as having high-impact within the field, their data policies were expected to reflect the current best practices adopted within the field of oncology, and were believed to have the potential to provide insight into the ways oncology researchers handle their study data. Once the 50 top-ranked journals were identified, basic data (e.g., journal title, ISSN, publisher name, impact factor, and URL for journal) about each title was recorded using an Excel spreadsheet.

Data Collection

In June of 2014 the authors collected and recorded information from 50 oncology journals' websites concerning their data guidelines. Using an expanded data collection instrument adapted from Piwowar and Chapman (2008) and JISC (2012), information about the journals were captured into a single Excel spreadsheet and multiple Word documents. Data collection consisted of documenting various dimensions concerning the journals' data guidelines, as presented through each journal's website content. 28 dimensions were sought and examined by the researchers, and data concerning these were entered into an Excel spreadsheet. Dimensions ranged from the location of the data policy on the journals' websites, the data formats the guidelines addressed, and the metadata required for data sharing, to the consequences for not complying with the journals' data policy requirements. For a full account of the dimensions examined for this study see Appendix 1.

In this study, a "data policy" refers to a statement providing guidance to authors about data types, supported file formats, when data should be submitted, and where. When a journal's website text simply offered a generic statement to authors that "supplemental material could be submitted," it was not counted among the group with data policies. The text of the journals' data policies, when available, were collected and copied verbatim into separate Word document files. Each Word document file contained a URL to the website where the data policy content was found, the title of the journal, the heading for the section where the information was found and the copied text. In some instances information pertaining to data guidance was found in multiple sections and, or pages, of a journal's website. In these cases the multiple locations were noted within the Word document. The text of the Word files were searched for key terms, key bits of data were extracted for the specific dimensions under examination, and the corresponding data were entered into the spreadsheet.

Data collection instruments consisted of a coding manual with definitions for the dimensions of data to be examined, Word documents which capture the text of the websites' data policies, and an Excel spreadsheet with columns corresponding to the coding dimensions. These were piloted by the researchers to test for their efficacy in capturing information useful for providing insight into the journals' data guidance practices.

The piloting process revealed several areas needing further clarification and development. For example, when the original spreadsheet was piloted and the data

mentioned in the journals' guidelines were examined, it was discovered that multiple file formats (e.g., CVS, DOC, PDF, JPG, etc.) and data types (e.g., datasets, documentation, metadata, structures, etc.) were discussed. Thus, the data collection instruments were updated to allow for a more granular recording of file formats and data types. The pilot, beyond pointing to the need for increased specificity in the dimensions of data extracted, also alerted the researchers to weaknesses in the definitions for the original codes used to record the strength of the journals' data policies. To address these issues the researchers revised the coding sheet and re-coded previously examined journal policies. This iterative process of revising and re-coding was repeated at several points in the study when newly collected data pointed to the need for additional specificity in the data collection instrument so that additional dimensions relevant to the goals of the present study could be included.

The researchers developed their data collection instrument based on previous work which had examined journal policies (JISC, 2012; Piwowar and Chapman, 2008). Previous research examined the following journal characteristics for sharing gene expression microarray data: journal impact factor, existence of a data sharing policy, and where the data policy was mentioned for authors (Piwowar and Chapman, 2008). The present study builds upon the existing research and extends the line of inquiry to include additional categories to further assess the availability and accessibility of journal data policies. Categories which were added include the amount of time spent locating the journal data sharing policy, the number of mouse clicks required to reach the policy, and the type, or format of data mentioned in the policy, and metadata requirements.

Data Analysis

Analysis of the data was carried out using two methods. The first method consisted of a quantitative analysis of how the journals performed on a single dimension. For example, data were collected from each journal's policy concerning metadata, with the journals discussing the topic being noted and tallied. The second method of analysis consisted of examining the extracted text in further detail, using case ordered displays to reveal variations to be found within a single dimension. For example, during the analysis of the topic of metadata, some journal policies provided directions concerning the specific kinds of metadata to be submitted. This deeper level of analysis revealed variations to be found in the collected data. By examining the variations in the level of detail to be found in the journals' data policies, the analysis revealed patterns in the depth of guidance, and recommended and/or required practices. An examination of the data using these two methods was carried out by the researchers across each dimension for the 50 journals.

Several procedures were used to ensure consistency in the data analysis process. These procedures consisted of frequent communications about the process among the researchers, examining the analysis performed by the other researcher, updating of coding manual with more specific instructions, and the revision of previously examined policies based on agreed upon changes. These efforts helped to ensure uniformity in the data collection and examination.

Limitations

Several limitations of the research methods must be acknowledged. The first of these being that while the researchers systematically collected data from all 50 websites, it is

 doi:10.2218/ijdc.v10i2.375

possible that instances of information concerning the journal's guidelines went undiscovered. As the researchers agreed to terminate searches for data policy information after examining a journal's website for ten minutes, the potential presence of guideline data on several sites and their contents are unknown. Furthermore, it must be stated that the study's researchers do not possess an in-depth clinical expertise in oncology, and so it is possible that additional details concerning the guidelines might be revealed if the data were examined by subject experts.

Findings

The findings from the analysis presented here consist of results pertaining to the type and prevalence of data guidelines, and the findability and location of the guidelines on the journals' websites. These results are followed by the findings concerning data accessibility, data types and file formats, and specific named data repositories. Finally, the findings concerning the metadata associated with the data, the monitoring of researchers' data sharing practices, consequences for not adhering to the journals' policies, and the relative strengths of the data guidelines are provided.

Fundamental Characteristics

The analysis revealed that among the 50 journal websites in the study sample, 72% (n=36) of the journals contained some form of data policy. As seen in Figure 1, data sharing was "required" for 40% (n=20) of journals, "optional" for 18% (n=9), or "other/partial" for 14% (n=7) of the journals with data policies. In the "other/partial" category were journals requiring data sharing for some types of data (i.e., gene sequencing data), while sharing was optional for other types. No data policies were discovered for 28% (n=14) of the examined journals.

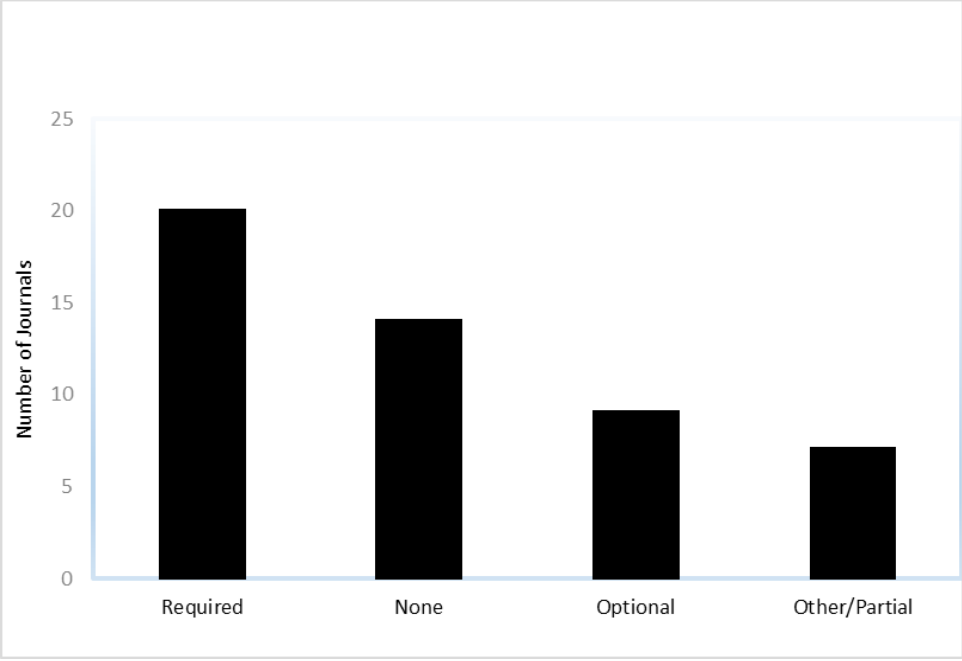

Figure 1. Presence of data policies among journals

Location of Data Policies

As seen in Figure 2, approximately 78% (n=28) of the data policies discovered among the 36 journals' websites were located under "Author Guidelines." This was the most common location for data policies among the journals. Following this, the next most common location for data policies was under "Availability of Data and Materials," found among roughly 19% (n=7) of the journals. In a sole case, data policy information was provided under "About this Journal."

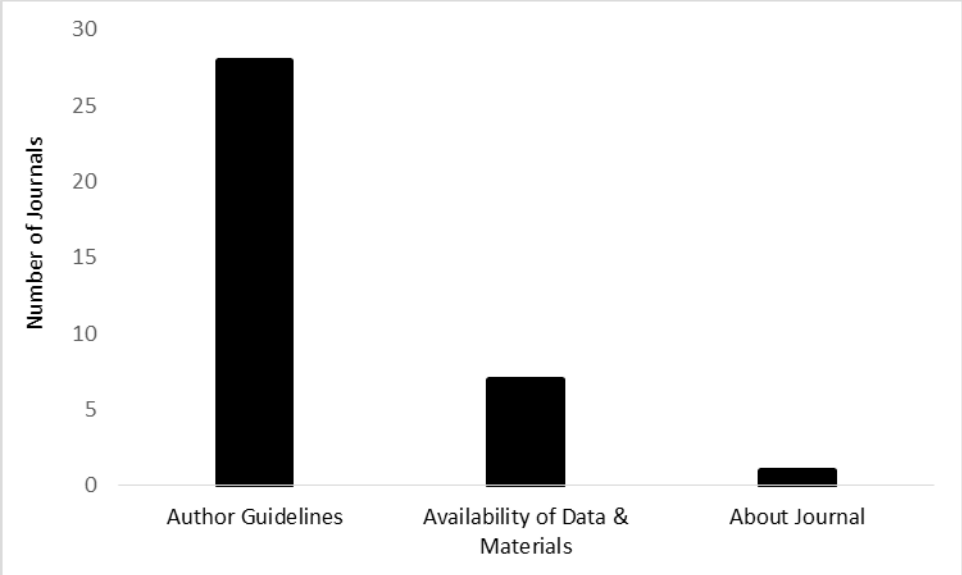

Figure 2. Location of data policies

Effort to Locate Data Policies

The discoverability of the data policies was examined by tracking the elapsed time and the number of mouse clicks required to locate the policies on the journals' websites. The amount of time taken to find the policy documents was found to vary greatly across the websites. The majority (64%, n=23) of the 36 journals' policies were discovered by the researchers in one to three minutes. The policies of another group of journal websites (11%, n=4) required between three minutes to five minutes to locate. Times which fell on the extreme ends of the time to locate spectrum were those websites (11%, n=4) that required less than a minute, and those that took in excess of five minutes (14%, n=5).

The number of mouse clicks required to locate data-related policy information on the journal websites also varied. As seen in Figure 3, the number of mouse clicks ranged from one click (n=10) to seven clicks (n=1) to locate data policies on the journal websites. Clearly the data policies on some of the journal websites in the sample were easier to access. In fact, policy information was located within one or two clicks for 64% of the journals (n=23). In contrast, other journal websites were problematic both in their structure and layout of content, thereby hampering the discoverability of this information. For 13 of the journals (n=13), the number of mouse clicks required to obtain data policy information spanned from three to seven clicks. The information being sought was sometimes more challenging to locate, as is seen in the greater number of mouse clicks and amount of time required to locate the data policies. As

 doi:10.2218/ijdc.v10i2.375

noted above, the logical placement under the "Author Guidelines" on the journal websites might have facilitated easy access and identification of data-related policy information. This is especially true in those journals featuring their author guidelines prominently on their websites.

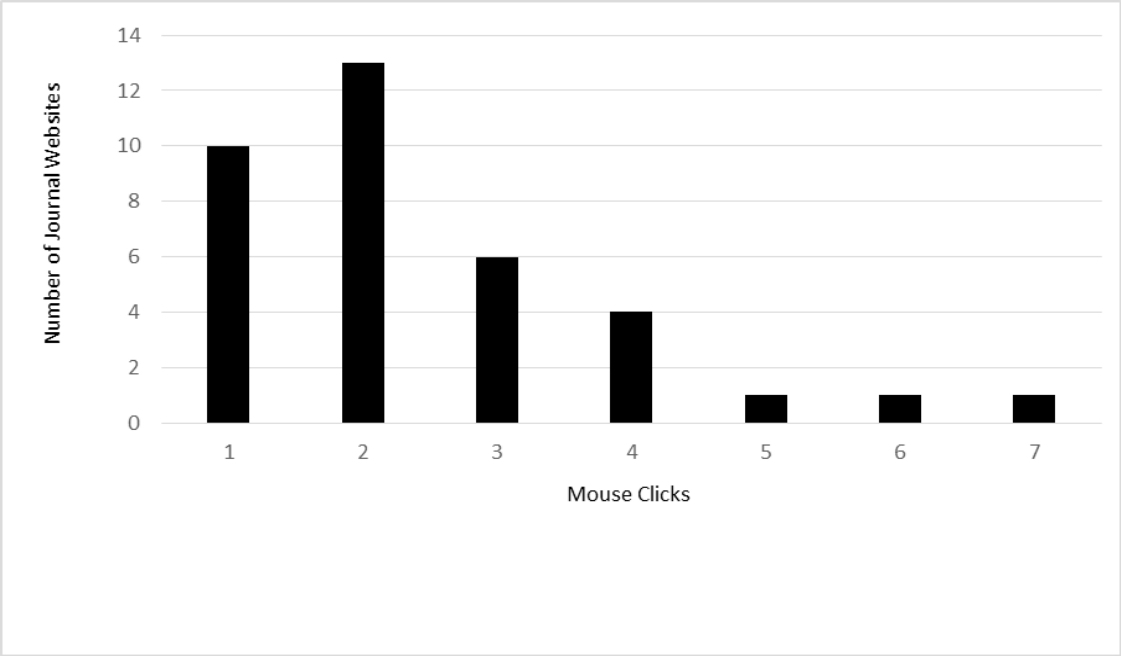

Figure 3. Number of mouse clicks

Accessibility of Data

The collected data policies were examined for statements regarding access to the research data used in the published studies. Of the 36 journals with discoverable data policies, roughly 72% (n=26) discussed the accessibility of data as a part of their policy statements. In those instances where accessibility was discussed, the majority of the journals (n=23) expected the data to be open, and a few noted that the data would be held in closed repositories (n=2), or that access to the data upon request was sufficient (n=1). In those cases where data would be "closed," the journal publishers would host the data. For these, access to the data was limited to individuals or institutions with a journal subscription. An additional journal discussed accessibility in the context of how to accommodate individuals whose abilities restrict their access to the publication, rather than providing access to research data.

Additional details concerning access to the research data were found in the data policies. These included discussions of potential exceptions to sharing the data. Conditions of access associated with the data were indicated by roughly 28% (n=10) of the journals with data policies. This included exemptions from providing data if it ~~was~~ were difficult to obtain, or if reuse was for commercial purposes. Economic issues surrounding data access were found in two of the journals' policies. In these, it was noted that fees could be applied as a way to recoup the costs associated with providing data.

Furthermore, it was discovered that the policies sometimes indicated who should have access to the data (see Figure 4). Discussions of the individuals and groups given

access to the data were found among roughly 53% (n=19) of the 36 journals with discoverable policies. Commonly encountered roles identified in the policies to be given access were editors, reviewers, and researchers/scholars. The most commonly found roles among the data policies were for reviewer, found in 50% of the journals (n=18), and for general reader (44.44%, n=16). Editors, publishers, and researchers were additional roles mentioned, with each appearing within the journal policies of two (5.55%) websites.

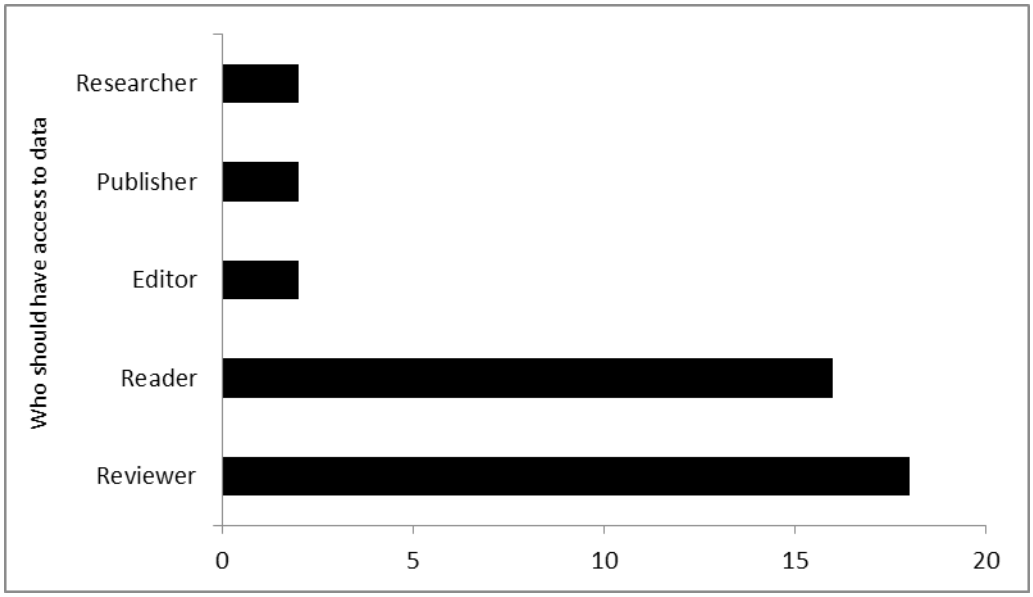

Figure 4. Who should have access to data

A number of journals recommended specific repositories where the data could be deposited. The most frequently mentioned data repositories in the policies are illustrated in Figure 5. Repositories which were commonly noted included Gene Expression Omnibus (GEO), ArrayExpress, and DNA Databank of Japan. Hence, the most frequently recommended data repositories were found to be discipline-specific.

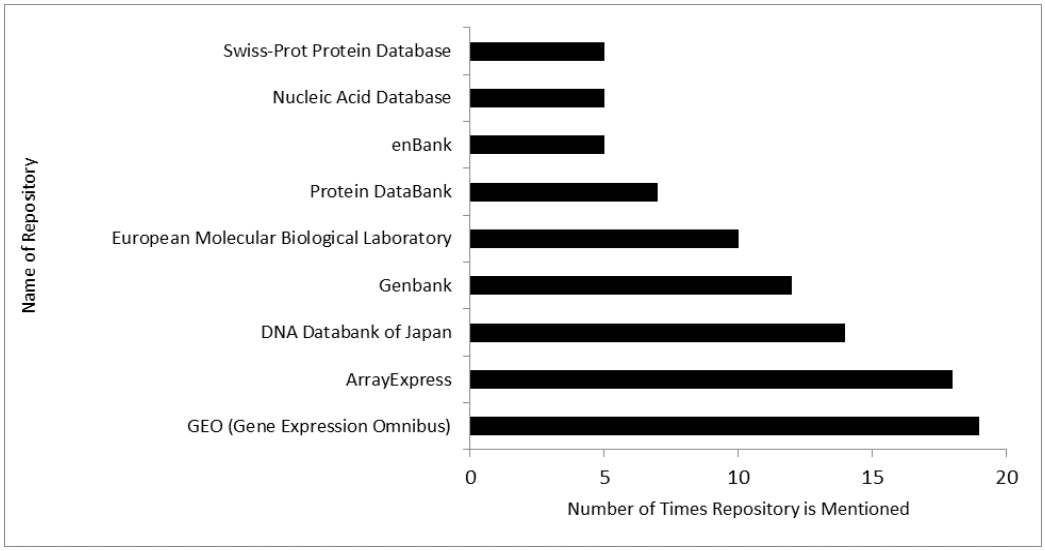

Figure 5. Most frequently named repositories (five or more mentions)

 doi:10.2218/ijdc.v10i2.375

The length of time that the data needed to be available was provided by approximately 22% (n=8) of the journals with data policies. While many of the journals indicated that data needed to be available for three years (n=7), one expected data to be available for ten years. The remaining journals (n=28) with data policies in the study sample did not specify the length of time that data are expected to be available.

Data Types

A remarkable range of data types appeared in the journal policies. As seen in Figure 6, the most often encountered data types in the journal policies were non-specific datasets, found in roughly 78% (n=28) of the journals' policy documentation, and protein/DNA sequences, discovered in approximately 61% (n=22). Additional types of data mentioned in the 36 journals with policies include structures (44.44%, n=16), and specimens (38.88%, n=14). While these results suggest that the primary concern of the policies was the underlying data crucial to the research, a large segment of the content was not specifically identifiable as primary research data. For example, approximately 47% (n=17) of the journals with data policies included the less specific multimedia content, roughly 11% (n=4) identified program code or software, and a few discussed supplemental documentation (n=2), and metadata (n=1). Rounding out the discussion of data types found in the policies were unspecified data types, which were discussed among roughly 17% (n=6) of the journals. Additionally, in several cases the topic of research data failed to be addressed altogether in their policy statements.

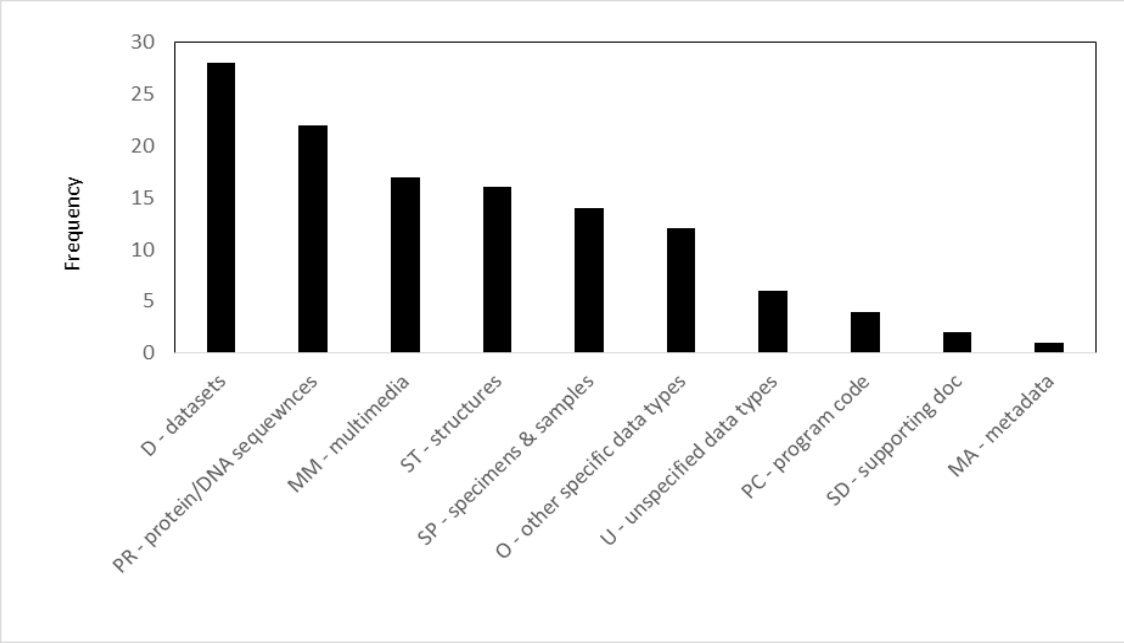

Figure 6. Data types mentioned in data policy

File Formats

The various kinds of file formats that carry the research data is a topic that was explored in conjunction with the journals' data policies. The study examined whether or not the journals address the topic of file formats in their policies and, in those instances where

this was encountered, to note what file formats were included. It was discovered that roughly 44% (n=16) of the journals with policies mention the topic of file formats in the context of the research data (or in the discussions of additional supporting materials). In these discussions the most commonly encountered file types in the data policies were PDF and those in the MS Office suite (DOC, DOCX, PPT, XLX, XLSX, etc.). As seen in Figure 7, PDF and MS Office file formats were identified among a majority (62.5%, n=10) of the 16 journals which mentioned file formats in their data policies.

Additional file formats to figure prominently in the data policies were media-based file types. Half of the journals (n=8) with policies addressing file formats included image file formats. The most commonly identified image file types were JPG and TIFF, although a number of other image and graphical formats were encountered. Other file types associated with image content that were encountered include GIF (CompuServe's Graphics Interchange Format), BMP (Microsoft Windows Bitmap formatted image), EPS (Encapsulated Postscript), WRL (Computer-Aided Design software), and PNG (Portable Network Graphic). Audio and video content were also found in the policies of a number of journals. Several journals (n=5) noted that audio content could be shared, and all but one of these identified specific audio file formats. The identified formats included WAV (Microsoft Wave), WMA (Windows Media Audio), MP3, MP4, and MPG. Several journals (n=3) identified various kinds of video files that could be accommodated, with the specific formats being identified as SWF (Macromedia Flash), and MOV (QuickTime). The file formats that were mentioned in the policies were not limited to media and standard business practice formats, and instead showed a remarkable variety. Additional file formats that were identified in the data policies include: HTML, KML, GIZ, MOL, MOL2, NB, PDB, PS, PSE, TEX, and ZIP.

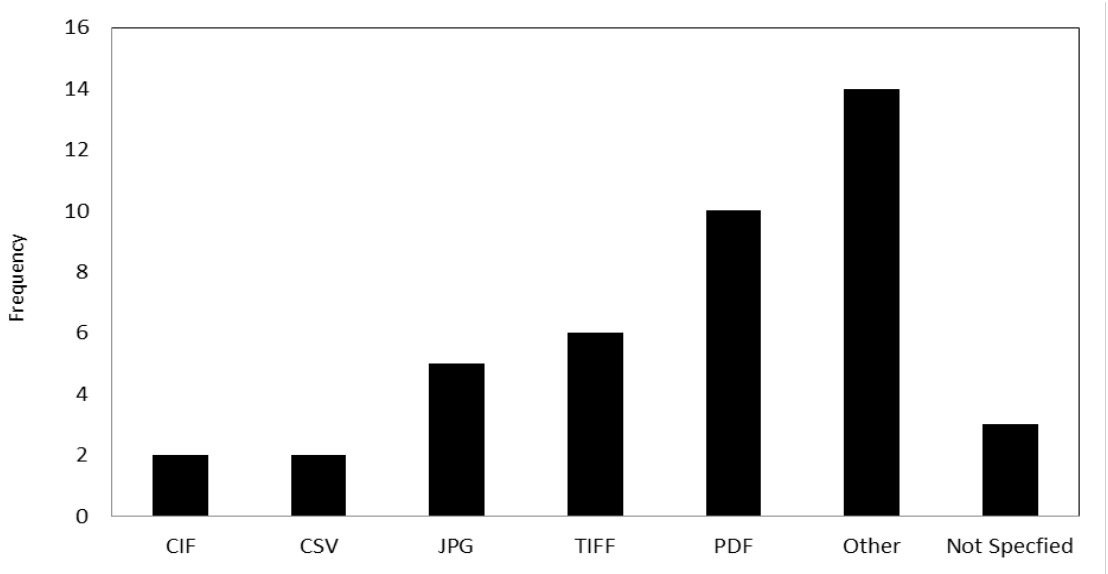

Figure 7. File formats mentioned in data policy

Metadata

The need for metadata – descriptive information about the data helpful for future access to, and understanding of, the research data – does not appear to be a concern for the journals' publishers. Only approximately 17% (n=6) of the 36 journals with data

 doi:10.2218/ijdc.v10i2.375

policies discussed metadata. Of these, several journals (11.11%, n=4) indicated that descriptive captions should accompany the data.

Consequences and Monitoring

Coverage of how the journals will monitor authors' compliance with the stated policies, and the consequences of failing to follow the guidelines, was another topic explored in the data analysis. It was discovered that none of the journals with data policies discussed if, or how, they would monitor authors' compliance with their guidelines. Furthermore, less than half (38.88%, n=14) of the journals stated the actions they would take for author's non-compliance with their data policies. The consequences for failing to comply with publishers' policies consisted of contacting the authors' institution(s) (n=7), and not publishing the manuscript (n=6). A final response, not reviewing the manuscript, was found among several (n=3) of the journals' policies.

Policy Strength

Policy strength was categorized as being either "strong" or "weak" based on previous definitions established by Piwowar and Chapman (2008). Data policies were considered to be "strong" if the journals required data sharing and accompanying evidence, such as accession numbers. In contrast, journal policies were categorized as "weak" if policies merely suggested, or requested, that data be shared, but that data sharing was not enforced. In addition, policies mandating that data be shared but which failed to require evidence that data had been shared were also categorized as "weak" policies. Using this guiding framework, roughly 56% (n=20) of the 36 oncology journals with policies had "strong" data policies whereas approximately 44% (n=16) of the journals had "weak" policies (see Figure 8).

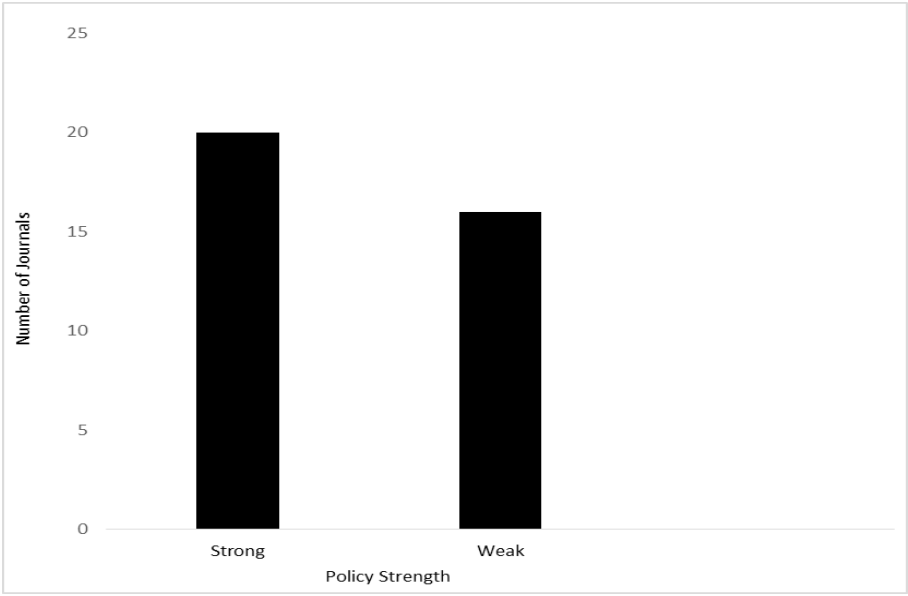

Figure 8. Policy strength

Discussion and Implications

Several implications arise from the findings of this study. First, the results of the study show a low level of support for data sharing and access to research data among oncology journals. Given that less than half of the journals in the study sample required data sharing, it is clear that scholarly publishers can do more to help promote and advance access to research data. As a means of supporting the advancement of research through data sharing, well-defined and easily discoverable data policies should be present on publishers' websites in those cases where their journals' contents regularly report study results. In this study, data policies were most often located in the "Author Guidelines" section on the journal websites. This finding is consistent with Piwowar and Chapman (2008), who discovered journal data sharing policies for microarray data were also commonly addressed in this location.

Second, the method used to evaluate data policy strength needs to be codified. Adapting definitions for "strong" and "weak" data policies from the existing literature was not without its challenges. The differences between the two were sometimes difficult to discern, as policies were found to partially meet the criteria outlined for inclusion within a particular category. Strong policies should include "well-described requirements" for sharing data and evidence for it, such as accession numbers for data sets submitted to formal repositories (Piwowar and Chapman, 2008). Yet, the present study discovered that some journals' policies only addressed data sharing for a particular type of research data. For example, one journal recommended using a public repository for microarray data, but no other data types were mentioned. Although journals may address data sharing, and require evidence from authors that data are shared, the weaknesses found raise questions concerning how comprehensive and helpful these data policies are for authors whose research efforts produce a wider range of data types.

An additional issue was discovered in that even if journals required making materials "freely available" to the editorial team and reviewers, other key stakeholders were noticeably missing from the conversation in the journals' data policies. While access to research data may have been noted as being required for the editorial team and reviewers in the journals' policy statements, researchers and the general public should be included in order to facilitate wider access to research data. Consequently, the original definitions for "strong" and "weak" policy strengths should be revisited to ensure the availability of research data more broadly. Efforts to facilitate wider accessibility to research data, especially federally-sponsored research, is more in keeping with mandates requiring public access to research funded by taxpayers. For these reasons, it became evident that it was challenging to categorize the data policies found in the sample of oncology journals as only strong or weak. To remedy this situation, the definitions used for prescribing what constitutes a policy as either "strong" or "weak" should be updated and expanded.

Our findings suggest that categorizing data policy strength is complex and nuanced; therefore, additional categories in the overall schema to capture these policy variations may be needed. More specifically, a category for "partial" strength may be warranted to reflect the conflicting or ambiguous messages contained in existing journal policies. Currently, a "partial" policy may describe a number of scenarios. For example, a policy could mandate data sharing but the placement of data in a public database and evidence of this via accession number(s) for the submissions may only be "preferable" (not required or enforced). Another case of a "partial" policy is when data sharing is only

required for certain types of data. Furthermore, when data are required to be made available to limited audiences, such as journal editors or reviewers as opposed to open to all audiences, this too constitutes a "partial" policy. Thus, a new category representing "partial" policies is needed to better identify data guidelines that are incomplete, vague, or otherwise limited in some capacity.

Third, journal publishers do not currently provide adequate direction through policy documentation and guidance. Publishers can play an important role in promoting access to research data by having well-crafted "data availability" policies (Lin and Strasser, 2014). Moreover, these policies should clearly require making research data publicly available, rather than merely recommending or suggesting that the data should be shared. Dryad's Joint Data Archiving Policy (JDAP) describes an effort endorsed by leading journals in the field of evolution to require, "as a condition for publication, that data supporting the results in the paper should be archived in an appropriate public archive" (Dryad, 2011). Coordinated efforts among journal publishers are expanding to other disciplines, and could potentially be adopted by the oncology journals in the study sample.

In addition, journal publishers could support the development of more useful policies by working with data repositories to provide specific procedures concerning data deposit. Acknowledging the leading role that journal publishers can play in the development of data sharing policies, Lin and Strasser (2014) recommend that journal publishers work more closely with data repositories to benefit authors, the research community, and the journals alike by making data sharing a more seamless process. The National Institutes of Health, National Science Foundation, and Wellcome Trust are among many agencies with policies and expectations that funded researchers will make their research data publicly available in "an appropriate repository" (Association of Research Libraries, 2015). Failure to comply with the terms and conditions of funding agreements can lead to serious consequences, "including the withholding of funding" (NIH, 2015). With this mind, these findings have implications for policy compliance and guidance for funded investigators as oncology journals can better direct authors to appropriate data repositories to help meet funder mandates for public access digital research data.

While funding agencies expect researchers to have data sharing, management, and preservation plans in place, the key publishers of this research currently do not provide adequate direction to authors through policy documentation and/or guidance. For example, a lack of expectations about, and guidance for, the metadata accompanying research data that was found in the current study has implications for the future discoverability and reuse of data. Will future researchers be able to understand and reuse data that lack basic descriptive details? This is not an inconsequential question. Efforts to ensure that research data are properly managed, preserved, and shared are of limited use if the data cannot be identified, retrieved, and analysed based on their accompanying metadata. Metadata practices for data sharing, as discussed in data policies, requires future work to ensure that comprehensive guidelines are provided to researchers.

Fourth, monitoring expected author actions regarding their research data during, and after, the journal article has been published, has an implication for the long-term availability and access to research materials. If no one is watching, as is suggested by the limited coverage of monitoring in the journals' policy language, will authors ensure their research data can be accessed and reused into the future? Expectations concerning what are considered to be sufficient, even among individuals and teams firmly

committed to sharing their research data, are likely to vary. Thus, policies with clearly delineated details concerning the availability and access of research data are needed.

Finally, the language used in the data policies would benefit from additional clarification, as it was often difficult to discern whether or not the journals were supporting the preservation and exchange of research data in their policy, or if they were discussing additional kinds of materials with a more supportive, and less critical function, for the publication. For example, does the term data refer to all information in digital form (i.e., charts or videos illustrating research results, author's biographical information presented as an audio file, and so on), or only the data points gathered in the context of a scientific experiment? The broad range of data types found in the data policies examined in the sample suggests that the journals are not distinguishing research data from the various forms of supplementary content associated with publication. Furthermore, as the current language surrounding data management activities and processes lacks specificity, it would be useful to develop an agreed upon vocabulary which supports efforts to maintain research data.

Conclusion and Future Research

This research contributes to the growing body of work surrounding the level of support and depth of data guidance provided to authors in journal policies. While the study sample consisted of 50 journals in oncology, these findings provide insight into a range of data issues addressed by journal policies and so are likely to have wider application. Future research that examines how journals in other disciplines are addressing issues around data sharing, reuse, and dissemination would be useful. An analysis of how journals' data policies compare to the requirements outlined by funding agencies may reveal additional areas needing to be addressed. Furthermore, the views of individuals engaged in cancer research about the journals' data policies would be a solid next step for future research. Research investigating what particular policies and guidance they deem useful would be helpful for supporting oncology researchers' efforts to improve the health care and outcomes of cancer patients. Finally, research that explores the point of view of journal editors would also contribute to a fuller understanding of data guidance practices.

This study is the first to examine the state of data guidance provided to authors in the policies of top-ranked journals in oncology. While a number of studies have examined data sharing practices among researchers, little is known about how journals are addressing this topic. Thus, the findings of this study are likely to be of interest to the broader library and information science community, as well as a number of key stakeholders such as funders, institutions, and researchers. The findings have practical implications for journal publishers, editors, and researchers. First and foremost, journal publishers should provide meaningful data guidance to prospective authors. Data policies that mandate sharing, guide researchers to relevant data repositories, and offer specific direction for details surrounding research data (e.g., requirements for metadata, coding manuals, file types, etc.) would be essential enhancements. These improvements will advance the preservation of, and access to, research data while also providing guidance for editors, journal reviewers and authors. As the ultimate aim of mandated data management and sharing policies among federal and other funding agencies is the encouragement of future research, a major step in ensuring this becomes a reality is the

provision of clearly defined practices in the data policies of the journals reporting research.

References

Akers, K.G. & Doty, J. (2013). Disciplinary differences in faculty research data management practices and perspectives. *International Journal of Digital Curation*, *8*(2), 5-26.

Association of Research Libraries. (2015). *Public access plans and policies for US government agencies.* Retrieved from http://www.arl.org/storage/documents/Public-Access-Plans-and-Policies-for-US-Government-Agencies-as-of-April-2015.pdf

Bloom, T., Ganley, E., & Winker, M. (2013). Data access for the open access literature: PLOS's data policy. Retrieved from http://www.plos.org/data-access-for-the-open-access-literature-ploss-data-policy

Borgman, C.L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology, 63*(6), 1059-1078. doi:10.1002/asi.22634

Carpenter, W.R., Meyer, A.M., Abernethy, A.P., Stürmer, T., & Kosorok, M.R. (2012). A framework for understanding cancer comparative effectiveness research data needs. *Journal of Clinical Epidemiology, 65*(11), 1150-1158. doi:10.1016/j.jclinepi.2012.06.005

Charbonneau, D.H. (2013). Strategies for data management engagement. *Medical Reference Services Quarterly*, *32*(3), 365-374. doi:10.1080/02763869.2013.807089

DeMartino, J.K. & Larsen, J.K. (2013). Data needs in oncology: "Making sense of the big data soup." *Journal of the National Comprehensive Cancer Network*, *11*(2), 1-12. Retrieved from http://www.jnccn.org/content/11/suppl_2/S-1.full.pdf

Dorsch, B. (2012). On the citation advantage of linking to data. *hprints*. Retrieved from http://hprints.org/hprints-00714715

Dryad. (2011). *Joint data archiving policy (JDAP)*. Retrieved from http://datadryad.org/pages/jdap

Joint Information Systems Committee (JISC). (2012). Journal Research Data Policy Bank (JoRD). Retrieved from https://jordproject.wordpress.com

Kesselheim, A.S., Lee, J.L., Avorn, J., Servi, A., Shrank, W.H., & Choudhry, N.K. (2012). Conflict of interest in oncology publications: A survey of disclosure policies and statements. *Cancer, 118*, 188–195. doi:10.1002/cncr.26237

Kozlowski, W. (2014). Funding agency responses to federal requirements for public access to research results. *Bulletin of the American Society for Information Science and Technology, 40*(6), 26-30. Retrieved from http://www.asis.org/Bulletin/Aug-14/AugSep14_Kozlowski.pdf

Lin, J. & Strasser, C. (2014). Recommendations for the role of publishers in access to data. *Public Library of Science (PLoS) Biology, 12*(10), e1001975. doi:10.1371/journal.pbio.1001975

MacMillan, D. (2014). Data sharing and discovery: What librarians need to know. *The Journal of Academic Librarianship, 40*(5), 541–549.

National Institutes of Health (NIH). (2015). *Plan for increasing access to scientific publications and digital scientific data from NIH funded scientific research*. Retrieved from http://grants.nih.gov/grants/NIH-Public-Access-Plan.pdf

Parkham, S.W. & Doty, C. (2012). NSF DMP content analysis: What are researchers saying? *Bulletin of the American Society for Information Science and Technology, 39*(1), 37-38. Retrieved from http://www.asis.org/Bulletin/Oct-12/OctNov12_Parham_Doty.pdf

Parsons, M.A. & Berman, F. (2013). The research data alliance: Implementing the technology, practice and connections of a data infrastructure. *Bulletin of the American Society for Information Science and Technology, 39*(6), 33-36. Retrieved from http://www.asis.org/Bulletin/Aug-13/AugSep13_Parsons_Berman.pdf

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

Piwowar, H.A., & Chapman, W.W. (2008). A review of journal policies for sharing research data. *Nature Precedings*. Retrieved from http://precedings.nature.com/documents/1700/version/1

Piwowar, H.A., Day, R.S., & Fridsma, D.B. (2007). Sharing detailed research data is associated with increased citation rate. *Public Library of Science (PLoS) ONE, 2*(3), e308. doi:10.1371/journal.pone.0000308

Rolland, B. & Lee, C.P. (2013). Beyond trust and reliability: Reusing data in collaborative cancer epidemiology research. In *CSCW '13, Proceedings of the 2013 Conference on Computer-Supported Cooperative Work, San Antonio, TX, USA, February 23-27, 2013* (pp. 435-444). New York: Association for Computing Machinery. doi:10.1145/2441776.2441826

Sansone, S. (2010). Omics data sharing – BioSharing: On data policies's plans and reporting standards. *Nature Precedings*. doi:10.1038/npre.2010.5049.1

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A.U., Wu, L., Manoff, M., Read, E., & Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *Public Library of Science (PLoS) ONE, 6*(6), e21101. doi:10.1371/journal.pone.0021101

Wellcome Trust. (2013). *Guidance for researchers: Developing a data management and sharing plan.* Retrieved from http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Guidance-for-researchers/

Wiley, C. (2014). Metadata use in research data management. *Bulletin of the American Society for Information Science and Technology, 40*(6), 38-40. Retrieved from http://www.asis.org/Bulletin/Aug-14/AugSep14_Wiley.pdf

Appendix 1

Table 1. Journal Data Policy Coding Manual

Dimension	Dimension Definition
A - Coder	Initials of coder
B - Full Title	Title of journal
C - ISSN	ISSN as provided by ISI Web of Knowledge (IWOK)
D - Impact Factor	Impact factor as provided by IWOK
E - Publisher Name	Name as provided by IWOK
F - Client	Governing body (If journal is published by a professional society, etc.)
G - Country	Country listed in address of publisher according to IWOK
H - Journal URL	URL of homepage for journal
I - Policy URL	URL where policy is found
J - Policy Location	Details concerning where the policy is found within page
K - Policy Name	Title on the policy page, section, or document
L - Policy Language	Text of policy copied and pasted into cell
M - Policy Term(s)	On the web site: what is the policy listed as – e.g., headings categories or labels such as data storage, data requirements, etc. (What they are calling it on the web site.)
N - Time Spent to Find	Time in minutes and seconds to find policy - 10 minutes max.
O - Number of Clicks	Number of mouse clicks needed to find policy
P - Funding Agency Noted	List any funding agency appearing within the policy
N = None	There is no policy stated.
S = Suggested	There is a suggested method of data handling.
R = Required	There is a required method of data handling.
O = Other	Partial, elusive, etc. – this definition will be developed as data is collected.

R - Policy Comment	Other potentially useful information
S - Date Reviewed	day-month-year
T - Reviewed by	Give initials
U - Data Types	Definition – issue of original and recorded info about originals
D = Datasets	
MA = Metadata	Is it required, and if so what kind (about original, about digital form, about creation, etc.)
MM = Multimedia	X-ray, MRI, sonograms, videos
PC = Program Code/Software	
PR = Protein/DNA Sequences	DNA models
SP = Specimens and Samples /Materials	
ST = Structures	Molecular models, etc.
SD = Supporting Doc	Code books, data collection methods and instruments, normalization of data
U = Unspecified	No data type defined
O = Other	Quirky and not covered above
V - Data Guidelines	Formal published guidelines
W - Where	Aspects surrounding where the data is, or will be placed
D = discussed	Is the storage of the data discussed at all? y/n
I = indicated	Do they indicate where the data needs to go? y/n
N = named	If yes, is there a specific named Database / Repository? Please state.
If yes, what repository type is indicated? Spell out after named code e.g., N(subject)	
U = Unspecified	No location information provided
O = Other	Not covered above
X = When Available	Discussion of time in association with availability of data
A = Availability	Is availability data mentioned? y/n
If yes, to whom?	Reviewers, publishers, colleagues, researchers, funding agencies, general public, etc.?
If yes, when must data be available? List choice:	Prior to / on publication, On submission, On acceptance, After publication, etc. Enter as - A(prior to publication)
D = Duration	Is duration of data availability specified? y/n ; enter as DY or DN
If yes, how long will data be made available?	
L = Lifecycle mention	Data Storage vs. Archive, if noted indicate stage, i.e., active use of data vs. preservation state indicated, raw vs. analysed
O = Other	Quirky and not covered above
Y = Accessibility	Discussion of the accessibility of the data
A = Accessibility discussed	Is accessibility discussed?
If yes, what kind of access is mentioned? List choice:	Open, Closed Enter as - A(open) or A(closed)
C = Conditions of accessibility.	Embargoed (time restriction), Redacted (data

List choice(s):	restriction), Limited to specific users (user restriction) ; enter as C(embargoed), etc.
E = Economics	Free vs. Paid ; enter as E(free) or E(paid)
O = Other	Quirky and not covered above
Z – Metadata	Discussion of metadata
S = Statement	Is metadata discussed? y/n
If yes, what is discussed? List choice:	(A)Required (B) Recommended (C) Mentioned, not specified (U) Not mentioned, not specified
D = DOI – url for dataset	DOI or other identifiers discussed? y/n
O = Other	Quirky and not covered above
AA - File Format	Definition
S = Statement	Are file formats for data discussed? y/n
If yes, what format is mentioned (e.g., CIF, CASTOR, CSV, PDF, JPG, RAW, etc.)	(A) CIF (B) CSV (C) JPG (D) PDF (E) Other (describe/explain) Note: enter and then at end tally up which ones are mentioned
O = Other	Quirky and not covered above
AB - Consequences	If the consequences of not following policy are provided, please list
AC - Monitoring	Monitoring of data sharing
AD - Policy Strength	List if weak or strong. Give details
AE - Note	Discussion of things that are noteworthy but that do not fit elsewhere
AF – Number	Primary key for each journal