Bishop, B, et al. 2020. Data Curation Profiling to Assess Data Management Training Needs and Practices to Inform a Toolkit. *Data Science Journal,* 19: 4, pp. 1–8. DOI: https://doi.org/10.5334/dsj-2020-004

RESEARCH PAPER

Data Curation Profiling to Assess Data Management Training Needs and Practices to Inform a Toolkit

Bradley Bishop[1], Hannah Gunderman[2], Rowena Davis[3], Tina Lee[3], Rebecca Howard[1], Robert Samors[3], Fiona Murphy[4] and Judit Ungvari[3]

[1] University of Tennessee, US

[2] Carnegie Mellon University, US

[3] Belmont Forum

[4] Fiona Murphy and MMC Ltd, GB

Corresponding author: Bradley Bishop (wade.bishop@utk.edu)

The purpose of this paper is to explore current data management training needs and practices for Belmont Forum member agencies and researchers to inform a Toolkit. Fourteen Belmont Forum affiliated individuals were interviewed following a predetermined set of questions to create data curation profiles of their funded work. The data curation profile questionnaire includes questions related to data management, storage, stakeholders, costs, training, and credentials. The interview findings highlight gaps in existing knowledge of data management theory and practice that could impact data re-use. Results of these interviews were used to populate a Toolkit of data management training and effective practice resources specifically developed to train Belmont Forum grant awardees. The results also highlight some attitudes and behaviours of current scientists, researchers, and agency representatives, and the impact of the implementation of data management plans on the open science movement.

Keywords: data management plan; data curation profile; open data; open data policy; research data management

Introduction

Data management planning is neither a new concept nor a new practice. The "data deluge" described by Hey and Trefethen (2003) necessitated a change to the scholarly output infrastructure and data creation in science to meet the demands of data-driven discoveries. A data management plan (DMP) is a structured, formal document describing the roles, responsibilities and activities for managing data during and after research (Bishop & Hank 2020). Alongside encouraging more public-facing scientific research, many funding agencies (86% of UK Research Councils and 63% of US funding bodies) require DMPs within the initial funding application (Smale et al. 2018), creating accountability for sharing data and managing digital outputs particularly in cases of publicly-funded research.

The purpose of this paper is to explore current data management training needs and practices for Belmont Forum member agencies and researchers to inform a Toolkit. Fourteen Belmont Forum affiliated individuals were interviewed following a predetermined set of questions to create data curation profiles of the funded work they were affiliated with. Nine participants were in the role of funders and five were principle investigators on funded projects. The data curation profile questionnaire includes questions related to data management, storage, stakeholders, costs, training, and credentials. The interview findings highlight gaps in existing knowledge of data management theory and practice that could impact data re-use. Results of these interviews were used to populate a Toolkit of data management training and effective practice resources specifically developed to train Belmont Forum grant awardees. The results also highlight some attitudes and behaviours of current scientists, researchers, and agency representatives, and the impact of the

implementation of data management plans on the open science movement. The following literature review provides more insight into previous studies of DMP implementation, awareness, and training.

Literature Review

Prior to funding agency requirements for DMPs, data management planning simply meant capturing details at data collection and data deposit into a repository to prevent information entropy from impeding re-use by the creators and other data consumers (Michener, Brunt, Helly, Kirchner, & Stafford 1997). Information entropy explains the real factors that lead to data loss, such as format obsolescence, hardware/software failures, legal encumbrance, among others. Projects are at greatest risk for data loss or damage "when it is produced by a small group or single person" (Sweetkind-Singer et al. 2006, p. 311). In addition, failure to collect details at the time of data creation makes it difficult to determine the reliability of the data, validate results using the data, and maximize the data's overall re-usability for new studies (Higgins, 2012). DMPs were promoted in concert with open data movements with the assumption that sharing science data would reap many benefits. For example, data sharing purports, without much evidence to back up these claims, that open data would lead to an increase in the visibility of research, allow the maximum transparency in science, reduce the cost of duplicating data collection, as well as new discoveries and uses (Smale et al. 2018).

The need for comprehensive Research Data Management (RDM) support for researchers is accelerating alongside a digital revolution in which evolving technologies have allowed for greater volume and diversity of data outputs. Within this revolution, scientific data originate in digital form and theoretically enable greater opportunities for storing, sharing, and reusing these data (Ray 2013), supporting a culture of open science and widespread access to scientific knowledge. These funder mandates are theorized to be the key drivers for positive shifts in organizational perceptions towards RDM (Cox & Pinfield 2013). Further, several academic journals now require researchers to make public the data and digital outputs associated with a publication (The Royal Society 2017; PLOS, n.d.). While a culture supporting RDM practices in scientific research and publishing is thriving, in practice, there exist numerous barriers to RDM for enabling open science and long-term scientific preservation, including governmental, institutional, technological, cultural, and economic influences and restrictions. Supporting sustainable, widespread RDM adoption in scientific research will require an understanding of how these barriers in practice serve to reduce, inhibit, or even block RDM uptake in research endeavors.

Implementing effective RDM practices in the early stages of a research project is a proactive approach to understanding data in their active use environment, increasing scholarly control of researchers and their engagement with their own research outputs (Bishop & Hank 2016). Therefore, from a researcher's perspective, RDM may help them reach their scholarly goals and achieve their desired research agenda. However, scientists often operate under the umbrella of a research institution or government operating under their own specific RDM guidelines, or lack thereof. Many organizations require the formation of a DMP, a document providing researchers with a "mechanism for stating how they will manage data associated with at least part of a research project's data lifecycle" (Smale et al. 2018, p. 2). However, at the time of writing, whether and how researchers actually adhere to these DMPs throughout the lifecycle of the research project are topics which remain unexplored in the literature, particularly in the area of data sharing.

While it is unclear whether researchers are using these DMPs in practice, the return on investment for those who do may be low. Dietrich et al. (2012) reviewed funding organizations and found varying disjointedness of coverage in data management policies concerning storage, licensing, metadata, and sharing. Therefore, while funders may require DMPs from researchers, the required elements within these DMPs vary significantly across organizations, and as a result, following required RDM guidelines for one organization may render the data and digital outputs difficult to use (or even unusable) by researchers from another organization. This complication is magnified when working across multiple nations, as required by all Belmont Forum-funded projects.

Currently, library and information science professionals, higher education institutions, and scientific funding organizations are investing time and funds towards remediating barriers to RDM through educational resources including online and in-person modules, courses, and Toolkits. For example, the Data Observation Network for Earth (DataONE) project[1] provides education modules and resources for research data management across the data lifecycle. While DataONE is holistically targeted towards researchers using Earth and environmental data, the RDM resources are applicable within any disciplinary setting. Similarly, the Earth

[1] https://www.dataone.org/.

Sciences Information Partnership (ESIP) hosts an online Data Management Training Clearinghouse[2] with RDM learning resources comparable to those offered by DataONE. These training resources reference RDM principles and effective practices in a broader context without disciplinary or organization-specific bounds. While these resources certainly create a digital infrastructure encouraging RDM adoption and accountability among scientists, it remains to be seen in the literature if uptake of these resources results in concrete, measurable advances in RDM adoption in scientific research. The interviews undertaken in this research help frame how Belmont Forum affiliated researchers use (or do *not* use) RDM principles in their digital outputs and whether a digital infrastructure of RDM resources proves beneficial within their workflow.

Methods

This study used a qualitative, semi-structured interview approach, using the Data Curation Profile (DCP) approach. After Institutional Review Board approval was gained, interviews with researchers and agency representatives were conducted. The sampling frame included any former or current Belmont Forum affiliated researcher or agency representative. Each institution, organization, country, and individual project has a different approach to data management; therefore, we attempted to sample across roles, projects, and countries. Ultimately, fourteen interviews across multiple projects with five researchers and nine agency representatives from six countries (four continents) occurred. Participants informed consent included anonymized data and removal of indirect identifiers; therefore, specific organization are not mentioned here or in the data. (Belmont Forum Data Curation Profiles, 2019).

The DCP approach was created to capture the step-by-step data lifecycle from scientists for digital curation. This DCP questionnaire was informed by literature, interviews with scientists, and validation from a panel of expert reviewers (Witt et al. 2009). The interview schedule contains questions related to how the data is collected and other critical data attributes, including size, format types, organization, description and representation, and storage. The interview schedule consisted of the following questions:

1. Please provide a brief overview of the research associated with the data we will be discussing.
2. Description of the data
   a. Approximately, how many data files exist?
   b. What is the average size of the data files? (units) and/or overall (total file size)
   c. What format(s) are the data stored in?
3. Data Flow & Use
   a. How was the data acquired/collected?
   b. What specific software programs or tools/hardware were used in the collection/generation of the data?
   c. How was locality determined? Place/time?
   d. What specific software programs or tools/hardware are required to utilize this data? (proprietary file formats)
   e. Describe briefly the way the data is currently organized:
      i. file name conventions,
      ii. any existing metadata, units, etc. (e.g. "detailed annotations", "a code book", "a data dictionary", "column headings in a spreadsheet", etc.).
4. Storage
   a. Where are the files currently stored? Include the storage media(s) and any tools used in your management of the data
   b. Are there backups of the data?
   c. Who is primarily responsible for managing these files?
5. Stakeholders
   a. Who is the intellectual property owner of this data?
   b. Who is the intended audience of this data? Is the data intended to be made available for re-use by others?
   c. Who might you imagine would be interested in this data? (e.g., other researchers in my field, researchers outside of my field, policy makers, etc.)
      i. How might this data be used by these people?

[2] http://dmtclearinghouse.esipfed.org/.

6. Costs
   a. How are data management efforts funded?
   b. What is the percentage of a project's total budget allocated for data management?
7. Training
   a. What are your top three research data management needs?
   b. Have you received any research data management training?
      i. If yes, what types of data research management training did you receive?
   c. What delivery formats do you prefer for training?
8. Please indicate your credentials and degrees.

The interviews were recorded and transcribed. The transcriptions were analyzed using NVivo. Categories and broad themes across responses to the questions emerged using terminology from the questions and responses. The informed consent process for participants included open data language and and the transcripts are available through the Open Inter-university Consortium for Political and Social Research (ICPSR).[3]

There are limitations of this study as agency representatives could not answer specific data questions. Also, more participants may have introduced further variety in responses. Saturation for DCPs may not be attainable given how unique every DMP implementation story is and the international scope of this population. The domains and disciplines participating in the transdisciplinary research funded by the Belmont Forum are broad, but the academic backgrounds and experience of these participants is almost entirely from STEM fields. Further work in this area is required to discover differences in RDM training and practices across areas. Ultimately, the convenience sample resulted from those willing to participate from a call throughout the Belmont Forum network.

Results and Discussion

This section presents major themes of the responses from the fourteen interview participants. Not all interview participants had the experience/authority to answer each question in the questionnaire shown, or their specific research/project parameters rendered the questions irrelevant. Therefore, some data-specific sections of the DCP are not discussed because the researchers and funding agency representatives were too far removed from those aspects of the RDM. Of those questions obtained from participants, keywords, phrases, and responses were coded in accordance with a controlled vocabulary that corresponded with the original questions. Coding the interviews allowed for quicker identification of similarities and differences in the participants' responses. Below, each thematic area from the Data Curation Profile (storage, stakeholders, costs, and credentials) is discussed with specific excerpts from the interviews.

Storage

Storage presents one element of RDM that may be solved for individual projects, but becomes complex across time, space, and multi-agency projects. Data sharing requires data storage, and the in-perpetuity assumptions of open data policies must include long-term support through institutions or governments. For example, most respondents indicated that data storage relied on institutional repositories or data centers. Without probing, it is difficult to assess how the data storage may affect or impede re-use. Most participants answered "yes" to having backups, although the location of backups varied and included cloud based services (e.g., Amazon and Google storage) as well as institutional repositories or other local storage facilities.

Some participants had clear knowledge of how their data is being stored, such as participant 12, stating, "there is a central kind of human infrastructure that looks after the management of this interoperability platform. So, specifically what happens is that there are each of these polls that I told you about-they all have their own internal data storage facilities which are then merged into a structure which is a core centralized data repository." Other participants were not entirely sure of their storage protocols, or they described storage plans as having various approaches. As participant 13 noted, "We have a very dispersed system of storage right now. Different universities have storages. Some disciplines have developed storage, the physical sciences. So, it's a bit of a hodgepodge of different approaches." Participant 7, a researcher, demonstrated no awareness of the location of their research data, stating "Who knows where the discs are." Funding agency representatives were ambivalent about storage as long as the data could be re-used. "I personally don't care where folks store things, as long as it's accessible to others and

[3] https://www.openicpsr.org/, https://doi.org/10.3886/E110203V1.

people can re-use this" (P. 8). The variability of both storage practices and individuals responsible for data management presents inherent sustainability access problems for some projects. However, one researcher participant expressed hope for a more streamlined storage solution: "The idea is to have a national plan where [European Country] will have about, let's say 7 to 8 very large national data centers, will host the resources for the different community and provide a minimum level of services in terms of networking [and] access" (P. 10).

The uncertainty of where data were stored and if it was being backed-up highlight basic questions that researchers and funding agency representatives should be able to answer. RDM compliance could, therefore, be streamlined with additional storage guidance. A dearth of clear answers to the question about individual(s) primarily responsible for data management may demonstrate a cultural assumption that the primary investigator (PI) is ultimately responsible for the tasks related to data management. Further training of researchers that RDM is essential, therefore a dedicated person or at least the person charged with those tasks should be named. The problems stemming from not having a dedicated data manager on a single project, but also across projects presents inherent access problems for sharing in-perpetuity. Therefore, as noted above, it may be advantageous to hire a dedicated data manager instead of charging all responsibility to researchers with a multitude of other professional duties.

Stakeholders

Although answers once again varied for the question regarding the intellectual property owner, a majority of participants considered the public to have these rights, stating "if the research has been done by public funds, then the data is open" (P. 14). Other common answers were the university or institution, and that the ownership would vary, as well as who can access such data, as participant 6 noted "if people come and ask us for the data then we will kind of give them access to it" (P. 6). When asked about licensing requirements, many participants were unsure of the policies surrounding the issue. A couple of participants had knowledge about these requirements, as participant 5 stated when asked, "Not that I'm aware of because I don't think there are any commercial aspects to it. I think all the data that was being used was available data. It was all through government agencies as opposed to government agencies working with a private company." Participant 5 further went on to say the funding agency did not recommend any licensing related to the project. It is worth noting here that for data to be machine-actionable some licensing must be indicated; even if data are public domain, a machine will not make the same assumption a human user would and a machine-actionable license such as the Creative Commons CC0 would be needed in the metadata.

The responses for who would be interested in using the data showed other discrepancies between researchers and funding agency representatives. Researchers' response to data re-use retains a narrow scope of other researchers in the same field as anticipated users, but a few expanded beyond their fields of study. Researchers' initial thoughts of similar uses to the original purpose of data collection make sense to address the question about other uses and users. Funding agency representatives took a wider view on stakeholders, but the statement that anyone might be interested in science data needs further validation.

The audience for undiscovered knowledge that is possible through open data is difficult to speculate, and defining all stakeholders is a challenging question. Still, the usability of discovery tools must be designed for likely potential re-users to ensure discovery and re-use. For example, some decision-makers could benefit from dashboards with analyzed data to inform actions and policy. Other researchers might need raw data in a useable data type, but with enough descriptive metadata that enables discovery and assessment for re-use.

Costs

The issue of costs also seemed to be an area of uncertainty for several participants. In examples where funding was already built into the project, there seemed to be a better support system for researchers, particularly when data management is seen as a natural addition to the workflow in a research project: "I would say, the problem is just that. It's allocating funding as a percentage of a project. The context of data management has to be fundamentally integrated for the entire project. It should not be seen as a separate piece of a project." (P. 11). When funding is built into the project, there are already anticipated costs related to RDM, including administrative costs: "It is funded through the initial project. So the money is initially taken through the university, certain [amounts] for administration [purposes]" (P. 5). One participant stressed the importance of including RDM costs in the budget, saying "Of course we have to want to maximize the money, but we don't have restrictions for that. We can pay for [RDM], if it's needed, we pay for that." (P. 15). One researcher participant mentioned not including data management in the budget, although they included funds for the collection of the data itself: "Well actually I plan the budget, but I didn't put an item code [for] data

management. I have item codes [for] data collection, but not management." (P. 4). Two participants provided estimates of the amounts in the budget allocated towards data management, as participant 4 noted "If I include the software, I think it will be about ¼," and participant 9 stated "if you're taking sort of multi-million [dollar] projects, it racks up. So, I think it's around 2 percent."

Costing RDM varies across data types and project. The non-responses in this section suggest a lack of knowledge and/or experience about budgeting for RDM that may be pervasive. As one researcher participant said, truncating a budget for an action that must be integrated into the entirety of a project for an indefinite amount of time is the problem. Not all researchers have the skills to allocate time and resources to this activity, yet truncating a budget for RDM beyond the data collection phase, as well as considering long-term funding that is sustainable for RDM beyond any single project, is key for open science. Any resources that funding agency representatives could develop to provide guidelines or expectations on how to budget for long-term data sustainability would be helpful to both enhance the accuracy of project budgets and to reinforce the importance of making data open.

Credentials

When asked for their credentials, 9 participants indicated they obtained a Ph.D. in their field. Three participants indicated their highest earned degree was a master's degree, and the rest of the participants did not indicate a degree level, but cited their fields of study. These disciplines included civil engineering, environmental sciences, computer science, planning and landscape, physics, geophysics, geochemistry, conservation ecology, and oceanography.

Training

Capacity building is a crucial, but often overlooked area of the journey towards open data. RDM training and materials exist, but like many optional items may be skipped over by research teams. Funding agencies could require attendance at data management trainings or completion of online modules relevant to each area of research. Due to the vast variation in RDM requirements across funding agencies, it is likely that each RDM training needs to be tailored to each institution, government entity, and funding agency. If institutions mandate training through clear RDM goals and expectations or by providing materials and especially funding, then the researchers will follow. Development of a common core of RDM expectations for funded projects across agencies may facilitate the creation of efficient basic training materials to introduce researchers to basic skills. The Toolkit brings together many existing resources to address a variety of basic needs, but synchronous assistance, working in groups, and other elements of education may be more effective than static options like canned modules.

Based on the interviews and guidance from other online RDM resource repositories, the Toolkit, hosted in GitHub,[4] was designed with three major sections: the Data and Digital Outputs Management Plan Annex (DDOMP) Researcher Guide, Data Management Training, and Best Practices & Standards.[5] The DDOMP Researcher Guide features step-by-step guidance on creating the DDOMP at the pre-proposal, full proposal, and awarded projects stages. The Data Management Training section provides a collection of workshops, webinars, and resources available by geographic region, resource type (professional certifications, virtual trainings, workshops, etc.), and by professional role. The Best Practices & Standards section offers a collection of RDM best practices by organization, by category (licensing, metadata, sharing data, and storage) and by geographic region. Now, there are resources to help funded projects address the Belmont Forum Open Data Policy and Principles adopted in 2015 that strives to ensure that data should be (1) Discoverable through catalogues and search engines; (2) Accessible as open data by default, and made available with minimum time delay; (3) Understandable in a way that allows researchers—including those outside the discipline of origin—to use them; and (4) Manageable and protected from loss for future use in sustainable, trustworthy repositories.

Conclusion

The scientific research landscape is experiencing a globalization of concepts within RDM including open data, data sharing, and data re-use. This globalization offers both challenges and opportunities for advancing open science. The current scholarly landscape pertaining to RDM adoption within scientific research largely approaches the topic through a binary lens: theoretical (open data leading to open science) versus

[4] https://github.com/bfe-inf/Toolkit.
[5] https://bfe-inf.github.io/Toolkit/.

practical (overcoming barriers that do exist to making data open). On the one hand, many scholars applaud RDM practices, particularly those involving data sharing, as a step towards accessible and liberatory scientific research. This inquiry makes clear the need for more research on the specific barriers faced by scientific researchers within an environment of increased advocacy for data sharing, data re-use, and general RDM adoption. Additionally, funding agency representatives' perspectives on how their policies may or may not be implemented as idealized requires further study.

The results of the interviews demonstrated marked differences in RDM policies and practices across disciplines and organizations, suggesting a need for standardization of RDM requirements in an effort to achieve data sharing opportunities more easily. A scorecard for evaluating data management plans with instructions on how to evaluate these vital components of data sharing could prove to be a useful tool to both agencies and researchers when individual researchers might not have answers (Bishop et al. 2019). Further, more research is needed on creating online Toolkits for RDM that are useful in a broad sense despite the staggering differences in RDM adoption and policies across institutions, governments, and countries. This inquiry lends weight to initiatives for online, editable Toolkits specific to an institution/agency rather than static training materials engaging with RDM in a broad sense.

On the other hand, scholars express concern with data sharing and open science from an intellectual property and innovation standpoint. As noted through this research, on the ground level exists the scientists themselves, who face real financial, logistical, institutional, and cultural barriers to implement many RDM practices aimed at promoting open science. In contrast, from a bird's eye view, representatives from funding agencies attempt to incentivize data sharing through policy. These perspectives may not see the same level of detail necessary to overcome the potential barriers on the ground. Further complicating the discussion is a third perspective of the information professionals—digital curator, data manager, data librarian, digital preservationist, or whatever title is given to those building and maintaining the infrastructure that makes data sharing possible. This emerging group is set to play a key role in RDM, but the lack of definition of that role means that the opportunity to harness their skills is currently missed by both researchers and funding agencies. In light of these differences, it is clear a more pragmatic approach to RDM infrastructure-building and education is needed, in which real barriers to RDM by institutional culture, technology, and process are acknowledged and addressed. Some ground truth barriers for data sharing from the researcher perspective can be improved through potential solutions from information professionals, and maintained and extended by the actions of funding agencies through the enforcement of data sharing compliance and clearer mandates for more transparent assessment of DMPs. These efforts may incentivize further investment from researchers' own institutes to engineer services and processes that fully integrate data sharing into existing infrastructures.

Data Accessibility Statement

Bradley Wade Bishop. 2019. Belmont Forum Data Curation Profiles. https://doi.org/10.3886/E110203V1.

Competing Interests

The authors have no competing interests to declare.

References

Belmont Forum. e-Infrastructures and Data Management. *The Data and Digital Outputs Management Plan Annex (DDOMP)*. http://www.bfe-inf.org/resource/data-and-digital-outputs-management-annex-full.

Bishop, BW and **Hank, CF.** 2016. Data curation profiling of biocollections. *Proceedings of the Association for Information Science and Technology*, 53(1): 1–9. DOI: https://doi.org/10.1002/pra2.2016.14505301046

Bishop, BW and **Hank, CF.** 2020. Curation, Digital. In: Kobayashi, A (ed.), *International Encyclopedia of Human Geography*, 2e. Amsterdam, Netherlands: Elsevier.

Bishop, BW, Ungvari, J, Davis, RI, Lee, T, Goudeseune, L, Virapongse, A and **Samors, RJ.** 2019. Belmont Forum Data Management Plan Scorecard (Version v.20190819_final). DOI: http://doi.org/10.5281/zenodo.3530933

Cox, AM and **Pinfield, S.** 2013. Research data management and libraries: Current activities and future priorities. *Journal of Librarianship and Information Science*, 46(4): 299–316. DOI: https://doi.org/10.1177/0961000613492542

Dietrich, D, Adamus, T, Miner, A and **Steinhart, G.** 2012. De-mystifying the data management requirements of research funders. *Issues in Science and Technology Librarianship*, 70(1): n.p. DOI: https://doi.org/10.5062/F44M92G2

Hey, AJG and **Trefethen, AE.** 2003. The data deluge: An e-Science perspective. In: Berman, F, Fox, GC and Hey, AJG (eds.), *Grid Computing – Making the Global Infrastructure a Reality*, 809–824. Wiley and Sons. DOI: https://doi.org/10.1002/0470867167.ch36

Higgins, S. 2012. The lifecycle of data management. In: Pryor, G (ed.), *Managing Research Data*, 57–61. London: Facet Publishing.

Michener, WK, Brunt, JW, Helly, JJ, Kirchner, TB and **Stafford, SG.** 1997. Nongeospatial metadata for the ecological sciences. *Ecological Applications*, 7(1): 330– 342. DOI: https://doi.org/10.2307/2269427

PLOS. n.d. Data availability. http://journals.plos.org/plosone/s/data-availability.

Ray, J. 2013. Introduction to Research Data Management. In: Ray, J (ed.), *Research Data Management: Practical Strategies for Information Professionals*, 1–22. West Lafayette: Purdue University Press.

Smale, N, Unsworth, K, Denyer, G and **Barr, D.** 2018. The history, advocacy and efficacy of data management plans. *bioRxiv* (pre-print). DOI: https://doi.org/10.1101/443499

Sweetkind-Singer, J, Larsgaard, ML and **Erwin, T.** 2006. Digital preservation of geospatial data. *Library Trends,* 55(2): 304–314. DOI: https://doi.org/10.1353/lib.2006.0065

The Royal Society. 2017. Data sharing and mining. Available at https://royalsociety.org/journals/ethics-policies/data-sharing-mining/ [Last accessed 21 May 2019].

Witt, M, Carlson, J, Brandt, DS and **Cragin, MH.** 2009. Constructing Data Curation Profiles. *International Journal of Digital Curation*, 4(3): 93–103. DOI: https://doi.org/10.2218/ijdc.v4i3.117

How to cite this article: Bishop, B, Gunderman, H, Davis, R, Lee, T, Howard, R, Samors, R, Murphy, F and Ungvari, J. 2020. Data Curation Profiling to Assess Data Management Training Needs and Practices to Inform a Toolkit. *Data Science Journal,* 19: 4, pp. 1–8. DOI: https://doi.org/10.5334/dsj-2020-004

Submitted: 20 June 2019 **Accepted:** 07 January 2020 **Published:** 27 January 2020

Copyright: © 2020 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See http://creativecommons.org/licenses/by/4.0/.

]u[*Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press. OPEN ACCESS