IFLA WLIC 2016 COLUMBUS

Submitted on: 16.10.2017

Data in Libraries: the big picture
10 August 2016
University of Chicago, Regenstein location, Chicago, IL, USA

Mitigating the Risk: Identifying Strategic University Partnerships for Compliance Tracking of Research Data and Publications

Andrea Payant
Data Management Metadata Specialist, Utah State University Merrill-Cazier Library, Logan, UT, USA.
andrea.payant@usu.edu

Betty Rozum
Data Services Coordinator and Undergraduate Research Librarian, Utah State University Merrill-Cazier Library, Logan, UT, USA.
betty.rozum@usu.edu

Liz Woolcott
Head of Cataloging and Metadata Services, Utah State University Merrill-Cazier Library, Logan, UT, USA.
liz.woolcott@usu.edu

Copyright © 2016 by Andrea Payant, Betty Rozum and Liz Woolcott. This work is made available under the terms of the Creative Commons Attribution 4.0 International License:
http://creativecommons.org/licenses/by/4.0

Abstract:

As international efforts to develop guidelines for data management emerge, the need for monitoring data deposit compliance increases. Utah State University has a team of university administrators, IT specialists, and librarians that addresses research data management issues. They developed a process to receive reports from faculty with successful grants that ultimately allows creation of records for data. Research grants are tracked though all stages and information about data and publications shared publicly on platforms capable of contributing scholarly work to research outputs aggregators (such as Share) and also to public facing bibliographic aggregators like Worldcat.

Keywords: Academic libraries, Data Management, Research compliance, OSTP memo, Cataloging data

Background

In recent years, much has been written about open data, data sharing, and data mandates. In 2004, the OECD began developing a set of guidelines to facilitate access to digital data from research funded with public financial sources (*OECD Principles and Guidelines for Access to Research Data from Public Funding*, 2007). The organization released these in 2007, and several countries since have implemented data policies, including the UK, US, Canada, Australia, Finland, Germany, the Netherlands and New Zealand (Pryor, 2012).

In 2011, the Research Councils UK published the *RCUK Common Principles on Data Policy*, outlining a framework with seven principles guiding data management. Many agencies, such as the Arts and Humanities Research Council and the Economic and Social Research Council had established data centers long before the OCED or RCUK documents (Dunning, 2006; "50 Years of ESRC Timeline," n.d.). The *RCUK Common Principles* attempts to bring a unifying theme to the existing data policies, stating that research data are a public good, produced in the public interest, and as such should be made as openly available as possible. These principles address many data concerns, including metadata, legal and ethical matters, citing data, and the use of public funds to responsibly manage research data (RUCK, 2011). Accompanying the policy is a guidance document to aid interpretation of the policy, which provides more details for each principle, guidelines for data management plans, and links to specific Research Council guidance for data management plans (RUCK, 2015).

While the UK's has approached making publicly funded data more accessible establishing policies to guide their research centers, to whom researchers submit funding requests, Australia's approach has been quite different. In 2007, the government produced the *Australian Code for the Responsible Conduct of Research*. In addition to addressing allegations of research misconduct, publishing and disseminating research, effective peer review, and managing conflicts of interest, the *Code* has a section devoted to research data. It specifies roles and responsibilities of researchers and institutions, including retention, ownership, storage, security, and confidentiality of data. The *Code* was developed by the National Health and Medical Research Council, the Australian Research Council, and the Universities of Australia. Compliance with the *Code* is a prerequisite for funding from the National Health and Medical Research Council (National Health and Medical Research Council (Australia), Australian Research Council, Universities Australia, & National Health and Medical Research Council (Australia), 2007).

The Australian government's approach to data management places responsibility on institutions to monitor and develop policies for researchers to make the products of their investigations accessible, but it does not mandate public access. They have developed numerous resources and established a single portal to facilitate access to open data via their research institutions, universities, and government agencies through the Research Data Australia portal, https://researchdata.ands.org.au/.

In the United States, the National Institutes of Health and the National Science Foundation implemented data sharing plan requirements in 2003 and 2011, respectively. In 2013 the White House issued a memorandum mandating that any U.S. federal agency receiving over $500 million in research and development funding from the federal government must require the data and publications resulting from funded research to be publicly available (Holdren, 2013). Unlike Australia, no central portal exists to facilitate access to data or publications resulting from U.S. federal funding. Each agency was charged with developing a plan to describe how it would meet the requirements of the White House policy. Several resources were developed, mainly by librarians, to distill the information from the

agency plans as they were released. A valuable resource for tracking data management requirements is a crowd sourced spreadsheet of data and publication responses. (Whitmire et al., 2015). As this spreadsheet shows, in the US, as in many other parts of the world, librarians have been very active in facilitating access to data and participating in the open data movement. Many academic libraries now offer assistance writing data management plans, teaching data literacy, and guiding researchers in locating and depositing data to an appropriate archive.

Questions arise about how compliance with government and funder policies will be monitored, and who is ultimately responsible. Several countries have encouraged open access to research data, namely Germany, Netherlands, New Zealand, Canada, and Finland (Pryor, 2012). Our experiences focus on the US situation and more specifically how our institution assists researchers in complying with policy mandates.

When a grant is awarded and a principal investigator (PI) indicates in a data management plan (DMP) that data and publications will be produced and deposited, the granting agency expects the PI to follow through on this commitment. The responsible party, however, is the PI's institution, not the individual PI. The institution is responsible to ensure that the grant is administered properly, funds are dispersed, reports submitted, and that publications and data are deposited to appropriate depositories.

Academic institutions have managed federal and state grants, plus grants for other funding entities for decades. Adding the responsibility to monitor the deposit of data and publication is a new wrinkle. These products are detailed in the DMPs, but they can be difficult to track down and verify. The publications may go into several different repositories, and data can go into a nearly unlimited number of repositories. If the university were asked to show where the data were deposited, how would it locate and verify that the PI had actually deposited the data related to the grant?

Utah State University

Utah State University (USU) is a public land grant university. In fiscal year 2015, our Carnegie R2 institution received $111 million in grants on the main campus. New funding awards were granted for 1125 projects in FY2015, indicating that USU has a robust research agenda.

For researchers to comply successfully with funder mandates, they must be aware of the requirements and have access to the necessary resources to meet these, such as secure computer storage, appropriate Institutional Board Review approval, grant paperwork submissions and approvals, etc. At USU, most of the individuals involved in managing the researcher grant process are housed in the Office of Research and Graduate Studies (RGS).

The Chief Information Officer and Associate Vice President for Information Technology, who oversees all information technology on campus, is also a key stakeholder as is the Information Security Officer, who reports to the Associate Vice President for Business. This position is responsible for computer security, and HIPAA and FERPA compliance. All of these units and people are involved in grant management and have a role in the successful administration of awards at our campus.

In order to maintain the University's effort to meet funder requirements in a uniform fashion, an informal group was formed in 2013 consisting of representatives from the Library, Office of Research, and Information Technology. In 2015, a new position was created within the Library, the Data Services Coordinator, charged with aligning efforts among these three units to fulfil the requirements set forth by the federal mandate and other agencies. This group, the Data Task Force, consists of the stakeholders listed above plus additional members from each area.

The largest issue the Data Task Force grappled with is how to determine if the PI has fulfilled the requirements of the grant, per the DMP. The Data Services Coordinator receives and reviews all DMPs. Many faculty do not clearly state a repository for publication deposit. Additional review of DMPs from successful awards reveals the same pattern. Even when a repository was indicated, locating the dataset was very time consuming, if not impossible. Further complicating matters, data and publications are often produced after the end of an award. It may take a very long time to see the results published or deposited in a repository.

Given these factors, how would the University know if the data were deposited per the requirements of the grant? If the University were audited, how would it produce evidence that it complied with the terms of the funding agency?

USU needed to establish a mechanism to help create an audit trail and record the data and publication deposits produced from federally funded research in which results are required to be made openly accessible.

This problem of tracking and accounting for the data and publications researchers produce was discussed during an informal meeting between the Dean of the Library, the VPR, and the Data Services Coordinator. The idea for creating catalog records for the outputs resulting from federally funded research was born.

Initial project planning started with those most closely involved with managing the grant awards and the close out process together with the Systems Analyst for the Research Office, who programs the software, Kuali, used by the Division of Sponsored Programs (Sponsored Programs). In the Library, our metadata specialist already participated on the Data Task Force, and we also involved our Head of Metadata and Cataloging.

Sponsored Programs could provide the Library with information about successfully funded federal grants. The Library could create records for the integrated library system and the institutional repository for the data and tag articles in the institutional repository (IR) with funder information. The challenge was determining how to obtain the information about the data and publications produced by researchers for each grant.

This group initially planned to shoulder the burden of identifying and verifying the outputs of funded research between the Library and Sponsored Programs. During discussions among our larger Data Task Force, we realized this was not feasible. We brought two other members to our group, the Associate Vice President for Research and the Director for Research Development. Through more brainstorming, we came up with a plan that places considerable responsibility on the shoulders of the researcher, where it truly belongs.

Certainly, faculty members need to assume responsibility for accounting for their research products, and indeed some US Federal agencies state they expect to see such documentation in their post-award management ("Data Management for NSF SBE Directorate Proposals and Awards," n.d. ;"Data Management Guidance for CISE Proposals and Awards," 2015; Center for Disease Control, 2015). University administrators recognized that a streamlined, efficient method with an intuitive interface would be essential if we were to succeed in capturing the location of data. Also, some prodding was needed beyond the power of the Library to motivate faculty to comply with deposit requirements. By pooling our human and technical resources, we have developed a system we think will offer a fairly high degree of success. Requiring researchers to supply information about data and publications deposit as part of the grant close out process we hope will result in faculty compliance.

Key Institutional Resources

Three key resources are required to accomplish our goals. Kuali, an electronic grant award management system designed to help the research community throughout the award process, will be used to track compliance with the data management standards outlined in researchers' data management plans and adherence to mandates to deposit their data for public access. The information gathered and housed in the Kuali system will be sent to the library for record creation.

Within the Library, we use USU's IR, Digital Commons, which is produced by the company bepress, to host records for each grant award. These will ultimately include, the DMP (if the PI agrees to make it public), primary metadata about the award and its associated research, along with citations for any accompanying publications, location information for related datasets, and any other agency specific requirements. The Library already uses Digital Commons for its IR, so we are leveraging existing resources.

We use the Library's integrated library system (ILS), Sierra, for its impressive reporting capabilities and enhanced searching which supports quick and accurate discoverability. Descriptive records will be made in the system, based on MARC standards, to mirror and link to the grant award metadata found in Digital Commons.

Workflow Overview

The workflow outlined has three different systems sharing similar, if not the same, information. While this may look confusing at the outset, each system has three separate functions. The initial information gathered in Kuali can only be used internally by a select few individuals at USU.

Research office personnel are sent pertinent public information to the library from the Kuali system to be housed in the Digital Commons database. Digital Commons serves two primary functions: 1) it allows a basic record to detail some of the grant information to be paired with files associated with that grant, for instance, the DMP and the primary metadata document that is exported from Kuali. Researchers may deposit data sets in Digital Commons, which would then be associated with the metadata record for their grant. However, we expect that most researchers will deposit their data in discipline-specific repositories. 2) Digital Commons is set up to be harvested by SHARE, a higher education initiative to gather metadata about research projects and make it openly accessible. This allows information about scholarly outputs to be harvested by a national aggregator of metadata for scholarly activities.

Once Digital Commons is updated with the final information, from Kuali, the location and accessibility of the data sets will be checked and the data set will be cataloged in USU's catalog. Metadata is transformed from the Digital Commons Dublin Core format to the MARC format used by the ILS. Due to the simplified nature of the Dublin Core used in Digital Commons, additional information from the Primary Metadata Document will need to be added to the MARC record to create the final permanent record.

Workflow Detailed

There are four phases of the workflow that are associated with the timeline of the award process: Proposal creation and submission, award set-up, award period, and award close-out.

During the proposal creation and submission phase, Kuali will be used to capture the basic elements required for the creation of a record in Digital Commons. Information gathered will include: PIs or

Author(s), project title, key words, University department, funding agency, and sub-award details if relevant.

Next, in the award set-up phase, Sponsored Programs receives agency award grant numbers. PIs will be notified via an automatically generated letter reminding them of federal requirements for public sharing of their publications and data. This letter provides PIs with contacts in both the Research Office (Director or Research Development) and the Library (the Data Services Coordinator) should they need assistance with any aspect of their grant. The letter also informs the PI of the requirement to provide the DMP that was submitted with their grant proposal, which is preferred, or fill out a form that provides basic metadata about their project. Once the investigator has submitted this information, Sponsored Programs sends the Library the data necessary to create preliminary records in Digital Commons.

In the third phase, during the award period, the Kuali system will automatically send notifications to PIs every six months reminding them of the requirement to deposit publications and make their data publicly available in appropriate repositories and encouraging them to update information the primary metadata document. Throughout the award period, this updated information is forwarded to the Library and updated in Digital Commons. The library will verify, if possible, that data associated with research are complete, deposited, and publicly accessible. If so, metadata records for datasets will be created in Digital Commons and the ILS. The timing of record creation for data will depend on when the researcher completes their project and is ready to release data to the public.

In the final phase, during award close-out, Sponsored Programs will notify PIs that updated metadata is required within 30 days of the close out of the award period. Since publications and dataset deposit will often take place after an award officially closes out, a special status of "data pending" will be attached to projects in the Kuali system until staff can confirm that the researcher has complied with the mandate to share their data. When this obligation has been satisfied, the status will be updated and the project will no longer be considered open.

Cataloging Components

To test the viability of cataloging datasets according to MARC standards, a few records were created in the ILS for datasets produced by USU researchers, which had already been deposited in Digital Commons. We created records using the standards set for RDA. Best cataloging judgment guided all decisions regarding appropriate use of fields and other choices. Ultimately, we determined dataset cataloging to be feasible in MARC.

Concerns arose while cataloging these datasets which will need to be considered as we progress. First, the names, or titles, assigned to datasets are not always grouped together and can vary between individual files. Second, there was a lack a clarity regarding the roles and responsibilities of the authors, or researchers. Third, there were uncommon and unspecified data formats and modes of access could not be conveyed in the record due to this fact. Fourth, if no readme file is included as part of the dataset, catalogers may have a difficult time creating the most robust and useful record possible. Finally, locating the datasets on the various funding agency websites was difficult, a problem which was exacerbated by the lack of DOIs.

As the workflow evolved, accommodating Dublin Core required further considerations. Since we create records for projects and datasets in Digital Commons as well as the ILS, we developed a crosswalk for the metadata between the two standards.

MARC Mapping	DC Mapping	Field Description
100	Creator	1st Author/Researcher listed
245 $a	Title	Title/Name assigned to data set
245 $c		All authors/researchers listed
264 $a		Place where data originated
264 $b		Primary institution name
264 $c	Date	Year of publication/deposit
347 $a		Digital characteristics - file type, refer to the file extension
347 $b		Digital file characteristics - encoding format
347$c	Format.Extent	Digital file characteristics - file size
536	Description	Granting Agency, grant award number
500	Description	Any additional information pertinent to the data/dataset that is not otherwise reflected in the record. Refer to the readme file, if available
500	Relation.IsReferencedBy	Citation for original publication based on this data set
538	Relation.Requires	Include information about the characteristic of the files, noting mode of access, software or computer access. Refer to the readme file, if available.
520	Description.Abstract	Include any summary information about the content of the dataset, such as an abstract.
650	Subject	Subject of research data
700	Creator	Name(s) of additional researcher
856	Description	URL for location of dataset
856	Identifier	Dataset DOI
856	Relation.IsReferencedBy	Link to associated Journal Article
856		Associated Journal Article DOI
	Type	Indicate the DCMI type (typically "Dataset")
	Format	Indicate the file format of the dataset, refer to the MIME types
	Publisher	Indicate where the data set is housed
	Coverage.Spatial	If reported, include the spatial coverage for the dataset
	Coverage.Temporal	If reported, include the date coverage for the dataset

We will use the most comprehensive metadata provided by the PIs to create the records. If PIs do not provide sufficient information, Sponsored Programs will ask them to supply the missing information in order to complete the cataloging of their data. After the library finalizes records in Digital Commons

the data can be exported as a spreadsheet. The spreadsheet can then be manipulated to map the Dublin Core elements to MARC fields thereby enabling an import of the data directly into the ILS in MARC format as part of a batch process.

Working with diverse viewpoints

We quickly realized that for our collaboration to succeed, we would all need to learn about each other's areas very quickly. For example, the librarians have become more knowledgeable regarding the general research process, researcher needs, and are now familiar with the award process for grants. RGS and Sponsored Programs have a better understanding of metadata and its importance in creating catalog records. And they have learned more about how the Library can support research efforts at the University.

Despite diverging concepts of metadata, discoverability, archiving, and preservation, our group shared expertise to create a workflow that will encourage researchers to manage their data more effectively and will enable our institution to mitigate the risk of non-compliance with federal mandates to share data

Anticipated outcomes

With the considerable work invested in developing these protocols and processes, the benefit to the University, Library, researchers, and patrons needs to be substantial to justify the efforts. The primary benefit is the ability to track and verify data deposits to ensure compliance with federal mandates. The processes developed at USU provide checkpoints to verify that data have not only been deposited as outlined in the DMPs but also that they are accessible and discoverable by the public. Many items deposited in repositories are technically available but difficult to locate. Library staff will check the information provided to ensure that the data are indeed where reported and will strive to improve discoverability.

Adding records for data to the ILS provides an advantage of enhanced tracking capabilities. The granularity of MARC metadata schema allows for flexibility in choosing which metadata elements will be recorded. The robustness of the software can quickly generate relational reports based on that metadata– not only on what has been deposited and where but also the type of grant, granting agency, university department, researcher, date of deposit, etc.[1] The ability to report these details opens the door to a number of potential assessment opportunities. We anticipate that this management system will allow the University to assess grant related information and workflows in a more substantial and complex format.

In addition to this primary benefit, three tangential benefits to this process have been identified, including: 1) exposing data, 2) DMPs, and 3) strengthened ties between the Library and the University.

The major advantage to cataloging the data is an openly accessible record. Records in the catalog are publicly available – to university administration as well as the public. Additionally, all MARC records

[1] As repository systems are developed for housing and exposing data sets, the emphasis on building robust reporting mechanisms to pull reports on deposited data sets has yet to be fully developed. To be sure, existing systems do allow for exporting data that can then be examined to create reports, but the tools are in their infancy, the metadata schemas are often confined, and the resources for the development of these repositories is often focused on other functions.

for data sets will be contributed to OCLC's WorldCat to enhance global discoverability. Contributing records to WorldCat provides an additional means of discovery for librarians, researchers, and search engines. Open research outputs are increasingly crucial to long term research strategies for universities, and exposing data sets through multiple avenues (Digital Commons, Share, Sierra, and OCLC Worldcat, for example) provides multiple venues for discovery.

The process also allows for DMPs from successful grant awards to be publicly accessible via Digital Commons. This provides future researchers and those interested in applying for grants to see examples of DMPs from funded grants to help construct their own applications. From a collection of public, successful DMPs, exemplary plans can be identified and curated to provide examples to other researchers. Having observed deficits in DMPs from reviewing such plans, librarians have detected a need for education about data management planning.

Overall, the process has afforded a unique opportunity for the library to provide a vital service for the University. Libraries continuously seek to communicate the importance of their role in academic life, and this is a prime juncture to provide a visible and timely service to the University. Developing working groups composed of library staff, faculty, and administrative stakeholders focused on a joint need can only help to strengthen that relationship. Additionally, being involved with the research generated by the faculty of the University will give librarians a deeper understanding of the work being done on campus. This can potentially translate into better collection development policies to support research and more inroads to help faculty understand file management, metadata, and end-user needs for data.

Assessment

Even with these great benefits, libraries embarking on this type of a project must prepare to measure the costs of the project and weigh them against the benefits. Costs can be measured in staff time, equipment/software required, and opportunity trade-offs. While developing this project out of a pilot phase and into a full-blown process, the USU library will be keeping track of staff time hours to ensure that reasonable expectations have been set. In particular, an analysis of the number of staff required will help with planning to adequately meet the demand. Currently no additional staff support is being offered to the Library, but with over 1100 grants awarded in 2015, additional staff may be requested to keep up with demand.

As the project continues, the Library will assess key stakeholders' satisfaction with services at each stage. This will primarily be done through regular meetings of the Data Task Force but will also likely encompass interviews with faculty and surveys of the campus community as well as identified potential users of data. As much as possible, the Library would like to ascertain how most potential users find the data and if there are additional ways we can facilitate that discovery. Key to that assessment is measuring if there is a shift in attitude about library services from administration, faculty, and end-users. Interacting with each of these groups at more prominent levels can provide an opportunity to demonstrate the value of libraries.

Additionally, with this project we also aim to see if there is a change in the quality of data management plans produced at the University. The Data Services Coordinator currently reviews data management plans for grants and maintains an assessment log of the quality. Comparing the DMPs before and after implementing the management system will provide some insight on the effectiveness of openly sharing data management plans.

Conclusion

This method for compliance checking is a holistic, cross-campus approach that both verifies data deposits and increases discoverability of those data sets. By cross-walking grant tracking data from a privately accessed system to a publicly available IR, the DMPs and data deposits of successful grants are available researchers, university administrators, and the public. This mutually-beneficial arrangement provides a much needed safe-guard to the University, discoverability for the researcher, and demonstrates the value of library services and expertise. As the project matures, additional benefits will likely manifest as the library is able to assess how services are impacting researcher's satisfaction, awareness of best practices, and attitudes towards library services and the perceived value we offer

Acknowledgments

The authors gratefully acknowledge the Utah State University Office of Research and Graduate Studies and Division of Sponsored Programs for collaboration in this project.

References

50 Years of ESRC Timeline. (n.d.). Retrieved from http://www.esrc.ac.uk/about-us/50-years-of-esrc/timeline/

Amanda Whitmire, Kristin Briney, Amy Nurnberger, Margaret Henderson, Thea Atwood, Margaret Janz, … Lisa Zilinski. (2015). A table summarizing the Federal public access policies resulting from the US Office of Science and Technology Policy memorandum of February 2013. http://doi.org/10.6084/m9.figshare.1372041

Center for Disease Control. (2015). *CDC PLAN FOR INCREASING ACCESS TO SCIENTIFIC PUBLICATIONS AND DIGITAL SCIENTIFIC DATA GENERATED WITH CDC FUNDING*.

Data Management for NSF SBE Directorate Proposals and Awards. (n.d.). National Science Foundation. Retrieved from http://www.nsf.gov/sbe/SBE_DataMgmtPlanPolicy.pdf

Data Management Guidance for CISE Proposals and Awards. (2015, March 15). Retrieved from http://www.nsf.gov/cise/cise_dmp.jsp

Dunning, A. (2006). The Tasks of the AHDS: Ten Years On. *Ariadne*, (48). Retrieved from http://www.ariadne.ac.uk /issue48/dunning/

Holdren, J. (2013, February 22). Memorandum for the Heads of Executive Departments and Agencies: Increasing Access to the Results of Federally Funded Scientific Research. Retrieved from www.white-house.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

National Health and Medical Research Council (Australia), Australian Research Council, Universities Australia, & National Health and Medical Research Council (Australia). (2007). *Australian code for the responsible conduct of research: revision of the Joint NHMRC/AVCC statement and guidelines on research practice*. Canberra: National

Health and Medical Research Council. Retrieved from http://www.nhmrc.gov.au/publications/synopses/_files/r39.pdf

OECD Principles and Guidelines for Access to Research Data from Public Funding. (2007). OECD Publishing. Retrieved from http://www.oecd-ilibrary.org/science-and-technology/oecd-principles-and-guidelines-for-access-to-research-data-from-public-funding_9789264034020-en-fr

Pryor, G. (Ed.). (2012). *Managing research data*. London: Facet Publ.

RUCK. (2011). *RCUK Common Principles on Data Policy*. UK: Research Councils. Retrieved from http://www.rcuk.ac.uk/research/datapolicy/

RUCK. (2015). *Guidance on best practice in the management of research data*. UK: Research Councils. Retrieved from http://www.rcuk.ac.uk/documents/documents/rcukcommonprinciplesondatapolicy-pdf/