Research Data Repositories: Review of current features, gap analysis, and recommendations for minimum requirements

by Claire Austin[1,2], Susan Brown[1,3], Nancy Fong[4], Chuck Humphrey[1,5], Amber Leahey[1,6*], Peter Webster[1,7]

[1]Research Data Canada, Standards and Interoperability Committee, [2]Environment Canada, [3]University of Guelph, [4]University of Toronto, [5]University of Alberta, [6]Scholars Portal, [7]Saint Mary's University

*Contact: Amber.Leahey@utoronto.ca

Abstract

Data sharing is increasingly recognized as integral to scientific research and publishing. This requires informed and thoughtful preparation from initial research planning to collection of data/metadata, interoperability, deposit in data repositories, and curation. Research Data Canada (RDC) is a collaborative, non-government organization that promotes access to and preservation of Canadian research data. The RDC Standards and Interoperability Committee (RDC-SINC) surveyed 32 Canadian and International online data platforms for storage, data transfer, curation activities, preservation, access, and sharing features. We developed a checklist to compare criteria and features between platforms. The survey revealed a heterogeneity of features and services across platforms, non-standardized use of terms, uneven compliance with relevant standards, and a paucity of certified data repositories. Recommendations for online digital infrastructure development to meet evolving researcher and end-user needs centre around persistent identification and citation of datasets, data reliability, version control, metadata, data sharing, privacy controls, long-term preservation of data, and certification of data repositories. We identified a need in Canada for investment in an integrated, comprehensive national digital infrastructure for research data.

Keywords

Data sharing, data publication, digital repository, data interoperability, data standards, data deposit, digital infrastructure

INTRODUCTION

Research data sharing is increasingly recognized as an essential component of scholarly and scientific research. Increased sharing improves the ability to reproduce results, replicate findings, and generate new knowledge (Parr & Cummings, 2005; Hernan & Wilcox 2009; Peng 2011; Poisot et al. 2013; Stodden et al. 2014, 2015). Although some disciplines (e.g., astronomy) have a long established practice of sharing and citing scientific data sets (CODATA-ICSTI, 2013), a very large number of researchers are still very reluctant to do so. Perceived risks in data sharing sometimes put forth by researchers, such as damage to the researcher's reputation, misinterpretation of the data, or misappropriation of the data (CODATA-ICSTI, 2013), all immediately disappear the moment the data are properly managed and documented. Some surveys have found that approximately half of

researchers share data (Alsheikh-Ali et al. 2011; Vines et al, 2014). However, this most probably follows publication of results in peer-reviewed journals, often years after the data were originally collected, and data sharing does not necessarily mean the data are useable by another researcher. The usability of shared data relates to 'best practices' in data management, data structure, interoperability, metadata, licensing, and accessibility (Jones et al. 2006; Peer & Green, 2014). In Canada, increasing public access to scientific research data will help drive innovation and discovery across the broader scientific community, as well as implementation of better data management practices (GoC 2014).

A major source of research funding in Canada is Tri-Council Plus (TC3+): the Social Science and Humanities Research Council (SSHRC), the Natural Sciences and Engineering Research Council (NSERC), the Canadian Institutes of Health Research (CIHR), and the Canada Foundation for Innovation (CFI). Tri-Council Plus has stated that, "*the potential of data-intensive research is progressively and rapidly outstripping our ability to manage and to grow the digital ecosystem to meet 21st century needs*" (GoC 2013). In an effort to establish a greater culture of data stewardship, the Canadian granting councils agreed to promote and develop appropriate data management systems and capabilities, in line with existing data and best practises globally.

In early 2015, the Canadian research councils formulated a harmonized *Open Access Policy* that requires all peer-reviewed journal publications funded by one of the three granting agencies to be made freely available online by depositing the manuscript(s) in an online repository within 12 months of publication (GoC 2015). CIHR-funded researchers are also required to deposit their research data into a relevant disciplinary repository immediately after publication of research results, and they must retain original data sets for a minimum of five years. This is enormous progress, but it also begs some important questions. Why are original datasets required to be kept for only five years? Why are NSERC- and SSHRC-funded researchers not also required to deposit their research data in a digital repository? When and how will Tri-Council provide incentive to researchers and reward them for '*data publication'*, elevating the practice to a first-class research output on par with traditional forms of journal publication and thereby lead the way for needed change in the academic reward system? Would it not benefit the researcher, the broader scientific community, and the common good if data publication were to precede journal publication, even?

Data publication should be peer reviewed as rigorously as journal articles in the academic and scientific literature, and data should be openly shared in curated data repositories. Data are the foundation of everything else that follows, and researchers must receive credit for producing reliable data (Costello 2009; Atici et al, 2013; Kratz and Strasser, 2014). We recognize that Principal Investigators (P.I.'s) have a primary responsibility in data management and data publication (see Endnotes 1&2). It must also be emphasized that there needs to be a robust digital infrastructure in place to support proper data management and to ensure that data are preserved in a useable form for people other than the creators of the data – whether or not the P.I.'s care about this, although they should.

Credible data publication requires effective data management and a robust digital infrastructure. Is such an infrastructure currently in place so that governments and funding agencies can take that next step in requiring robust data management plans and deposit of research data in data repositories? This is the question that the present paper seeks to answer, at least in part.

METHODS

The Research Data Canada (RDC) Standards and Interoperability Committee (SINC) surveyed Canadian and international online data platforms to identify currently implemented standards, requirements, and features related to the management and sharing of research data across a variety of academic disciplines. This work was done in parallel with the development of, '*Guidelines for the deposit and preservation of research data in Canada*' (RDC 2015). The categories for assessment used in the present work were developed from community guidelines and digital preservation literature (see References).

Online data platforms that were publicly accessible via the world wide web and that allowed data upload were included in the survey. The survey was performed during the period October 2014-February 2015. The first phase focused on a group of large, established, general platforms (specifically Dryad, FigShare, Dataverse, ICPSR, Pangaea). DataCite, although a metadata platform – not a data repository – was also included. Publicly available information, including upload and submission instructions, data requirements, recommended metadata and file naming conventions, data sharing and deposit policies, user guidelines and documents, data dissemination formats, persistent identifiers, and stated data preservation activities were reviewed. In some cases, online platforms restricted user access and did not have openly available documentation regarding metadata and data submission requirements. In those cases, we created a user account and password and attempted to load a sample dataset into the data platform for the purposes of the review.

In total 32 online platforms were surveyed for the deposit and submission, storage, description, curation, preservation and archiving, dissemination policies and features, collaboration options, and open access (see Table 1). These included platforms in the Biological & Life Sciences, Social Sciences (economics, sociology, political science, etc.), Medical & Life Sciences, Earth & Environmental Sciences, one from Physics, and one from Astronomy. There were 19 multidisciplinary platforms covering multiple disciplines in the same general domain area (e.g. medical sciences, social sciences).

Comparison of the 32 online platforms was a challenge due to the heterogeneity of features and the non-standardized use of terms. Platform features and data criteria to be surveyed were developed based on *Data Seal of Approval* guidelines (DSA 2013), *Trustworthy Repositories Audit & Certification* (TRAC) criteria and checklists (CRL/OCLC 2007), and an initial survey of features observed in the selected online platforms. These features and data criteria were compiled in a checklist that was used as a tool to compare features and requirements across platforms (see Table 2). The use of the checklist to identify majority practice (i.e. >50% across platforms) with respect to any feature or data criteria was still exceedingly difficult. Therefore, for the summary results we used a lower threshold, arbitrarily set at 40%, as a more informative indicator of relatively common practice with respect to their implementation (or not).

RESULTS

Summary results from our survey of the 32 online platforms are found in Table 3. Detailed results can be viewed online in the, "*Repository Requirements Features Review Spreadsheet*" found in the RDC-SINC Dataverse Repository (see RDC-SINC 2015).

Subject areas

We found that a large number of platforms surveyed handled a variety of data and were multidisciplinary in scope. However, the majority identified with a particular domain or area of study (e.g. Earth and Environmental Sciences, Social Sciences, Medical and Life Science, etc.). Online platforms survyed often had strong government and academic affiliations, with nearly 41% (13 out of 32) being supported directly by government. An additional seven were NGO's, six were institutional (academic), three were corporate or commercial, and for some the affiliation was unclear.

Metadata

As to the kinds of features the platforms supported for metadata and description of datasets, we noted that they generally recognized depositors as being central to the data publication process. Datasets and metadata uploaded to these platforms often contained information concerning authors, publishers, subject matter, dates of collection, abstract etc..

Support for metadata ingestion and creation was a feature that we looked at particularly closely. The majority of platforms surveyed, 69% (22 out of 32), used some kind of local or custom metadata profile or schema for description and documentation of datasets. Nearly 38% of platforms surveyed (12 out of 32) supported or were mapped to a standard metadata set for resource description, e.g., Dublin Core (DC) or DataCite. Additional support was noted among some of the platforms for discipline specific standards such as the FGDC and/or ISO 19115 for geographic information (7 out of 32 platforms), or the Data Documentation Initiative (DDI) (6 out of the 32 platforms). While many platforms used standardized metadata, a number of the major platforms used non-standardized, internally devised metadata schemas which could not be cross-searched and that were not interoperable with any other system or resource. The granularity of metadata provided also varied significantly across platforms, with some support for dataset or file-level metadata descriptions available for only a few.

Persistent identifiers

Typically, the platforms surveyed ensured that uploaded datasets were assigned a unique or persistent identifier (e.g. URI, PID) for proper online identification and access. However, they varied in their approach to the use of persistent identifiers, with some providing a resolvable URL to the dataset's associated metadata.

Approximately half (17 out of 32) of the platforms surveyed supported the Digital Object Identifier (DOI) standard for persistent identification of datasets. Other persistent identifier standards that were used included DSpace Handles and URNs (6%, or 2 out of 32), with the majority using a local or some unknown unique identification system. Typically, identifiers were assigned at the level of metadata description for the dataset or study. Concerning the ease of data citation, we found that close to 63% of the platforms (20 out of 32) provided a direct data citation and/or some other mechanism to cite stored data.

Version control

We found that version control, although an important issue, was still an unresolved problem. More than two thirds of the platforms surveyed allowed depositors to edit files after they had been uploaded. Fewer than half offered any standard version control system, other recommendation for versioning, or version statement. Approximately two-thirds of the platforms provided time stamping of uploaded files. Time stamping appears to be the most common practice employed to identify

changed files, but this does not constitute version control. Only one platform offered a systematic and persistent method for identifying versions of datasets (Universal Numeric Fingerprint (UNF)).

Ownership and data reuse

Approximately three quarters of the platforms surveyed (24 out of 32) associated a Creative Commons or other open license with the datasets. The majority also supported other data use licenses – often customized to the specific platform – but not meeting any standards. These included restricted licences where ownership rights were retained and that defined limited terms of use for datasets. Provision for access to data with restrictions was noted in close to 84% of the platforms (27 out of 32). Nearly 66% of the platforms (21 out of 32) published a specific policy on data sharing, terms of use, and ownership. Three provided no information concerning terms of use of shared or downloaded data.

Open access

There was a the sense that data should be made available online when there were no legal or ethical reasons not to do so. Although open data access was not universal, nearly all of the platforms provided some public information concerning terms of use. When open access was provided it was not always strictly anonymous. Most of the platforms surveyed, 78% (25 out of 32) offered some form of authentication whereby users needed to "sign in" in some way to gain full access. More work is needed to understand the kind of restrictions applied and the reasons for them.

Data usage

With regard to tracking data usage, approximately half (15 out of 32) of the platforms surveyed indicated that they offered download or other usage statistics to demonstrate access to and reuse of datasets. The remainder provided no information related to usage.

Fees

Nearly all of the platforms surveyed offered some form of open, free, or anonymous access for data download. Two thirds of of them also offered free data deposit. One third sought some form of payment or funding from some or all data depositors for services such as data preparation, curation or preservation.

Dataset publication

In general, the platforms surveyed offered data providers some level of support for dataset publication, although these activities varied greatly between platforms and across disciplines. Approximately two thirds indicated that they offered some sort of data preparation, metadata support, or review of the data prior to publication.

Data curation

One third of the platforms surveyed did not provide any information concerning data curation activities or data support. For those that did state that there was some data curation activity, the detail and extent of the curation services provided were vague or unclear.

Interoperability

In general, in terms of standards for the effective access and exchange of data and metadata, we note that support for open and interoperable standards is not widespread. Only 34% of the platforms surveyed (11 out of 32) supported Open Archives Initiative (OAI) protocols, such as the OAI-PMH protocol for the open exchange and harvesting of data and metadata. However, nearly

two-thirds (19 out of 32) offered alternative access to data and metadata through some form of Application Programming Interface (API) for online access and exchange. Sixteen of the platforms surveyed supported either XML or JSON format for export and exchange.

Preservation

We were able to extract very little detailed information from the information provided concerning preservation. Nonetheless, nearly 56% (18 out of 32) indicated that they did offer long-term storage and preservation of data and had a preservation policy and practices statement. Additionally, 44% (14 out of 32) indicated that the platform set-up included multiple redundancy and backup for files. Fewer than 13% (4 out of 32) indicated the use of standard file transfer and copy systems such as LOCKSS or CLOCKSS (LOCKSS 2015; CLOCKSS 2015).

Certification

Only 20% (6 out of the 32) of the platforms surveyed were certified under some form of community assessment or certification body such as *World Data System* (WDS) or *Data Seal of Approval* (DSA). With only two of the platforms providing information concerning their succession plans, we note that statements and policies concerning plans for data after the online platform ceases to exist were virtually non-existent.

DISCUSSION

Increased data sharing and greater openness of scientific research requires robust data infrastructure and sound data and metadata management practices. The present survey is a broad overview of the current features of Canadian and international repositories and data sharing platforms. This work is not a comprehensive list of available online data platforms or data repository requirements and features, nor is it a replacement for repository assessment or accreditation. It has, however, identified areas where action is needed to develop the necessary national digital infrastructure in Canada to support researchers with management, sharing, and preservation of research data. The checklist and findings may also assist further study and development of 'best practices.'

Moving forward, investment is needed to develop an integrated, comprehensive digital infracture and to improve data sharing and reuse of research data in Canada. See, for example, the initiative funded under the European Union's Horizon 2020 research and innovation programme resulting in EUdat (EUdata 2015). For data sharing to be effective, data must be reliable, usable, easily discoverable, accessible, and stored in a persistent manner for the long-term. Most importantly, datasets must be considered legitimate research outputs and be appropriately acknowledged for their value in promotion, tenure, and funding decisions to the same degree as are other peer-reviewed publications. The emergence of data journals publishing peer-reviewed scholarly and scientific datasets is a step in this direction. However, this needs to be accompanied by a significant culture change in the academic community in order to become a reality.

New data journals recommend or use existing digital online platform infrastructures (Figshare 2015). Their data policies vary in terms of standards, compliance enforcement, and data review (Stodden et al, 2013; Peer & Green 2015). Data sharing is frequently a 'self-deposit' model, whereby the publisher recommends a list of online data platforms that may or may not perform quality control or review of the data deposited (NPG 2015). Support for standard metadata is highly variable

between repositories and data sharing platforms. More than two thirds (23) of the online platforms surveyed provided some support for metadata creation (i.e. guidelines, templates, review etc.), but most large ones still left metadata quality control largely in the hands of the data providers. Metadata are the backbone of any dataset and ongoing quality control of metadata is as important as the data. Metadata are vital in ensuring that the data are correctly understood and can be effectively used. Given the importance of quality control, it is noteworthy that the majority of the platforms surveyed did not address this issue.

Data curation is the activity of managing and promoting the use of data from the point of creation to ensure that the data are fit for contemporary purpose and available for discovery and reuse (RDC 2014). For dynamic datasets this may mean continuous enrichment or updating to maintain fitness for purpose. Higher levels of curation also involve links with annotation and with other published materials. One third of the platforms surveyed provided no information concerning data curation, and the remainder provided only vague or unclear information. Additional work is needed to understand the curation process used by different online platforms in much greater detail, to understand what is meant by curation in each case, how data selection, retention and quality control decisions are made and what processes are in place.

In the development of research data management services and support, the primary focus has been at the institutional level. This often coincides with the need to develop institutional online repositories such as those that now exist at Harvard, Hong Kong University of Science and Technology, John Hopkins, Monash University, and Purdue (Wong, 2009). Beyond the needs of repository managers and organizations who are primarily interested in digital preservation, few resources are available for researchers, survey managers, granting agencies, publishers, librarians, or archivists to assess the suitability of online platforms for research data deposit and sharing (Humphrey 2015; Guindon 2014). However, there do exist excellent repository assessment and best practice guidelines, such as the *Trusted Repository Audit Checklist* (TRAC), *Trustworthy Digital Repository Checklist* (TDR), *Digital Repository Audit Method Based on Risk Assessment* (DRAMBORA), and the *Data Seal of Approval* (DSA). Managers and researchers can adapt and use these for self-assessment, to perform internal evaluations, and to develop sound data management and preservation practices.

Clearly, costs associated with data and metadata infrastructure and curation services are considerable, and these will increase with the success and growth of each repository. DataCite Canada, for example, is providing its DOI minting service for free to non-profit organizations until March 31, 2016. This business model is currently under review.

Conclusion

Although Principal Investigators are ultimately responsible for the integrity of the data upon which their research findings are based, few have the knowledge, time, or resources to implement state-of-the-art data management practices or evaluate online data storage and sharing options (RDC 2014; Guindon, 2014). The results of the present survey suggest that there is still a great deal of work to be done to ensure that online data platforms meet minimum standards for reliable curation and sharing of data. We believe that Canada's Tri-Council is wise in being cautious about what it requires from researchers in terms of data management and online deposit of research data until a

robust national digital infrastructure, including supported data management, is established in Canada.

Academic libraries and archives already have experience with client service and with storage of a vast array of file types: audio, images, software code, and datasets. A logical next step for improving digital infrastructure in Canada would be the expansion of existing library and archive services in the development of a national data infrastructure, including institutional repositories, with complementary data management consultation services to support researchers (Wong, 2009). We also recommend that 'best practices' for data management and the systematic use of data repositories be incorporated into the curriculum at university undergraduate and graduate levels in the humanities, business, sciences, engineering, computer science, mathematics and statistics, and medical sciences, to begin to building capacity and skills in this area.

Author statement

All authors contributed equally to the writing of this paper. All authors declare no conflict of interest. Opinions expressed in this paper are those of the authors and do not necessarily reflect the policies of the organizations with which they are affiliated.

Feedback

Readers are invited to provide feedback to the authors at the following link: https://www.surveymonkey.com/s/RDC-SINC_IASSIST2015_paper_approval

Funding

This work was supported by Research Data Canada (RDC).

REFERENCES

Alsheikh-Ali, A., Qureshi, W., Al-Mallah, M., & Ioannidis, J. P. A. (2011). *Public availability of published research data in high-impact journals*. PLoS One, 6(9) doi: http://dx.doi.org/10.1371/journal.pone.0024357

Atici, L., Kansa, S. W., Lev-Tov, J., & Kansa, E. C. (2013). *Other people's data: A demonstration of the imperative of publishing primary data*. Journal of Archaeological Method and Theory, 20(4), 663-681. doi:http://dx.doi.org/10.1007/s10816-012-9132-9

CODATA-ICSTI (2013). *Out of cite, out of mind: The current state of practice, policy, and technology for the citation of data*. Task Group on Data Citation Standards and Practices. Data Science Journal, Volume 12. https://www.jstage.jst.go.jp/article/dsj/12/0/12_OSOM13-043/_pdf Accessed 2015-04-27.

CLOCKSS (2015). *Controlled LOCKSS*. http://www.clockss.org/clockss/Home

Costello, M. J. (2009). *Motivating online publication of data.* Bioscience, 59(5), 418-427.

CRL/OCLC (February 2007). *TRAC – Trustworthy Repositories Audit & Certification: Criteria and Checklist v 1.0.* Center for Research Libraries & Online Computer Library Center. http://www.crl.edu/sites/default/files/d6/attachments/pages/trac_0.pdf

DCC (2009). *Digital Repository Audit Method Based on Risk Assessment – DRAMBORA*. The Digital Curation Centre. http://www.repositoryaudit.eu

DSA (July 2013). *Repositories – Data Seal of Approval Guidelines, v2*. Data Seal of Approval. http://datasealofapproval.org/media/filer_public/2013/09/27/guidelines_2014-2015.pdf

EUdata (2015). *Research Data Services, Expertise & Technology Solutions*. http://www.eudat.eu

Figshare (2015). *The rise of the 'Data journal'*. The FigShare Blog. http://figshare.com/blog/The_rise_of_the_Data_Journal_/149?utm_source=Users+%2B+Advisors&utm_campaign=7d388f3383-figshare_integrates_with_Projects8_9_2013&utm_medium=email&utm_term=0_e5f7149158-7d388f3383-97189037

GoC (2013). *Capitalizing on Big Data: Toward a policy framework for advancing digital scholarship in Canada*. Government of Canada, Tri-Council Plus. http://www.sshrc-crsh.gc.ca/about-au_sujet/publications/digital_scholarship_consultation_e.pdf Accessed 2015-04-27.

GoC (2014). Canada's Action Plan on Open Government 2014-2016. Government of Canada. http://open.canada.ca/en/content/canadas-action-plan-open-government-2014-16 Accessed 2015-04-27.

GoC (2015). *Tri-Agency Open Access Policy on Publications*. Government of Canada http://www.science.gc.ca/default.asp?lang=En&n=F6765465-1

Guindon, A. (2014). *Research Data Management at Concordia University: A Survey of Current Practices*. Feliciter 60(2), p15.

Harvard (2015). Privacy tools project - Data tags. Harvard School of Engineering and Applied Sciences. http://privacytools.seas.harvard.edu/datatags

Hernan, M. A. & Wilcox, A. J. (2009). *Epidemiology, Data Sharing, and the Challenge of Scientific Replication*. *Epidemiology*, 20(2), 167-168. doi: http://journals.lww.com/epidem/Fulltext/2009/03000/Epidemiology,_Data_Sharing,_and_the_Challenge_of.3.aspx

Humphrey, C. (2015). *Preserving research data in Canada - The long tale of data*. Chuck Humphrey Blog. http://preservingresearchdataincanada.net Accessed 2015-04-27.

Jones, M. B., Schildhauer, M. P., Reichman, O. J., and Bowers, S. (2006). *The new bioinformatics: Integrating ecological data from the gene to the biosphere*. Annual Review of Ecology, Evolution and Systematics, 37, 519-544. doi:10.1146/annurev.ecolsys.37.091305.110031

Kratz, John; Strasser, Carly (2014). Data publication consensus and controversies, v3. http://f1000research.com/articles/3-94/v3

LOCKSS (2015). *Lots of Copies Keep Stuff Safe*. http://www.lockss.org

NPG (2015). *Questionnaire to assist with Scientific Data repository evaluation.* Scientific Data, Nature Publishing Group. http://www.nature.com/uploads/ckeditor/attachments/1301/SciData_respository_evaluation_March2015.docx

Parr, C. S., and Cummings, M. (2005). Data sharing in ecology and evolution. *Trends in Ecology & Evolution*, *20*(7), 362 - 363. http://www.sciencedirect.com/science/article/pii/S0169534705001308

Peer, L., Green, A. (2014). *Committing to data quality review.* Presented at the 9th International Digital Curation Conference (DCC). http://isps.yale.edu/sites/default/files/files/CommitingToDataQualityReview_idcc14-PrePrint.pdf

Peer, L., Green, A. (2015). *Research data review is gaining ground*. Political Science Replication Blog. https://politicalsciencereplication.wordpress.com/2015/03/26/guest-post-research-data-review-is-gaining-ground-by-l-peer-and-a-green/

Peng, Roger D. (2011). *Reproducible research in computational science*. Science, 334(6060), 1226-1227.

Poisot, T., Mounce, R. and D. Gravel. 2013. Moving toward a sustainable ecological science: don't let data go to waste! Ideas in Ecology and Evolution, Vol 2, No. 6. http://library.queensu.ca/ojs/index.php/IEE/article/view/4632/4992

RDA (2015). *Repository Platforms for Research Data*. Research Data Alliance Interest Group. https://rd-alliance.org/group/repository-platforms-research-data/case-statement/repository-platforms-research-data-case

RDC (2014). *Glossary of terms and definitions*. Research Data Canada. http://www.rdc-drc.ca/glossary

RDC (2015). *Guidelines for the deposit and preservation of research data in Canada*. Research Data Canada, Standards and Interoperability Committee. In press. http://www.rdc-drc.ca

RDC-SINC (2015). *Research Data Repository Requirements and Features Review*. Research Data Canada, Standards and Interoperability Committee. http://dataverse.scholarsportal.info/dvn/dv/rdcsinc

Stodden, Victoria; Guo P, Ma Z (2013) *Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals*. PLoS ONE 8(6): e67111. doi:10.1371/journal.pone.0067111

Stodden, Victoria; Leisch, Friedrich; Peng , Roger D. (2014). *Implementing Reproducible Research*. CRC Press. ISBN 9781466561595.

Stodden, Victoria; Miguez, Sheila; Seiler, Jennifer (2015). *ResearchCompendia.org: Cyberinfrastructure for Reproducibility and Collaboration in Computational Science*. Computing in Science & Engineering, 17(1), 12-19.

Vines, T. H., Albert, A.Y.K., Andrew, R. L., De´barre, F., Bock, D.G., Franklin, M. T., . . . Rennison, D. J. (2014). *The Availability of Research Data Declines Rapidly with Article Age*. Current Biology 24(1), 94–97.

Wong, G. (2009). *Exploring Research Data Hosting at the HKUST Institutional Repository*. Serials Review, 35(3), 125–132.

Endnotes

1. RDC defines data management as, "The activities of data policies, data planning, data element standardization, information management control, data synchronization, data sharing, and database development, including practices and projects that acquire, control, protect, deliver and enhance the value of data and information" (RDC 2014). RDC views the Principal Investigator (P.I.) as having responsibility in this area, his or her role being defined as the person who "has a research leadership role and is the point of contact for a project or partnership that applies the scientific method, historical method, or other research methodology for the advancement of knowledge resulting in independent, objective, high quality, traceable, and reproducible results. The P.I. has primary responsibility for the intellectual direction and integrity of the research or research-related activity, including data production, findings and results, and ensures ethical conduct in all aspects of the research process including but not limited to the treatment of human and animal subjects, conflicts of interest, data acquisition, sharing and ownership, publication practices, responsible authorship, and collaborative research and reporting. While various tasks may be delegated to team members, some of whom may have greater expertise in specific areas, the P.I. is familiar with the various technical and scientific aspects of a project and how they fit together, is able to identify and remediate gaps, and ensure communication within the team and with users of the research data and results" (RDC 2014).

2. RDC uses the following terms and definitions relevant to the deposit and preservation of research data (RDC 2014): *Data centre* - A facility providing IT services, such as servers, massive storage, and network connectivity. *Data repository* - An archival service providing the long-term care for digital objects with research value. The standard for such repositories is the Open Archival Information System reference model (ISO 14721:2003). *Repository* - Repositories preserve, manage, and provide access to many types of digital materials in a variety of formats. Materials in online repositories are curated to enable search, discovery, and reuse. There must be sufficient control for the digital material to be authentic, reliable, accessible and usable on a continuing basis. *Trusted Digital Repository (TDR)* - A repository whose mission is to provide its designated community with reliable, long-term access to managed digital resources." Please see the Glossary for definitions of other related terms (RDC 2014).

TABLE 1. Online data platforms surveyed

3TU.Datacentrum http://datacentrum.3tu.nl/en/home/	ICPSR https://www.icpsr.umich.edu/icpsrweb/landing.jsp
ArcGIS Online http://doc.arcgis.com/en/arcgis-online/share-maps/share-items.htm	ImmPort http://www.immport.org/immport-open/public/home/home
Archaeology Data Service http://archaeologydataservice.ac.uk/	IRIS http://www.iris.edu/hq/
B.C. Conservation Data Centre http://www.env.gov.bc.ca/cdc/	Journal of Applied Econometrics Data Archive (Queen's University) http://qed.econ.queensu.ca/jae/
Barcode of Life Data Systems (BOLD) http://www.boldsystems.org/	LabArchives http://www.labarchives.com/
BioLINC (Biologic Specimen and Data Repository Information Coordinating Center) https://biolincc.nhlbi.nih.gov/home/	National Snow and Ice Data Centre http://nsidc.org
Canadian Astronomy Data Centre (CANFAR) http://www.canfar.phys.uvic.ca/canfar/	NESSTAR* (<odesi>, CESSDA) http://www.nesstar.com http://www.cessda.net
CERN Open data portal http://opendata.cern.ch/?ln=en	Ocean Networks Canada http://www.oceannetworks.ca/information
CKAN* http://ckan.org	OpenAIRE / Zenodo repository http://www.zenodo.org/
DataCite https://www.datacite.org	Opencontext.org http://opencontext.org/
Dataverse* (OCUL Dataverse, Harvard) http://dataverse.org	OpenICPSR https://www.openicpsr.org/
Dryad http://datadryad.org/	Pangaea http://www.pangaea.de/
EASY (DANS) https://easy.dans.knaw.nl/ui/home	Polar Data Catalogue https://www.polardata.ca/
Figshare http://figshare.com	Scratch Pads http://scratchpads.eu/
FlowRepository https://flowrepository.org/	SDA http://sda.berkeley.edu/archive.htm
GEOSS portal http://www.geoportal.org/web/guest/geo_home_stp	UK Data Archive / RESHARE (E-prints) http://www.data-archive.ac.uk/home

* Data repository software

TABLE 2. Features checklist used to compare 32 online data platforms

Category	Sub-category	Detailed features
Hardware & Infrastructure	Server (server resources, platforms etc.)	● Cloud (i.e. Amazon S3) ● Dspace ● Local - IBM server and storage, VMware ESXi virtualized redundant server farm ● Other
	Cost	● Free to access, download, and deposit data ● Free to access, but contribution suggested or required for deposit, i.e. funding structure for access/deposit beyond the threshold ● Publishing charge $ ● Formal agreement with research/monitoring program for funding to support archiving and serving the datasets
	Size	● Size of repository (number of files, datasets)
Description	Domain	● Multidisciplinary ● Earth & environmental science ● Medical & life sciences ● Social Sciences (Economics, Sociology, Political Science, etc.) ● Physics ● Biological and Life Sciences
Preservation	Redundancy	● Multiple redundant copies ● CLOCKSS - Geographically and geopolitically distributed network of redundant archive nodes
	Persistent identifiers	● DOI (specify where possible) ● DSpace Handle (HDL) ● Other persistent IDs ● Other unique resource identifiers (i.e. URIs) (not persistent) ● EZID registration management or other persistent identifier registration
	Persistent data deposit	● Long Term preservation of data
	Curation	● Data curation (specify where possible)
Privacy & Security	Security	● Authentication mechanisms ● Distinction between public and private data
Archiving	Author identifier	● ORCID ID ● SCOPUS ID ● Digital Author Identifier
	Timestamping and version control	● Timestamped upon upload ● Data can be edited following upload ● Version statement ● Universal Numeric Fingerprint (UNF)
	Citation and references	● Citation provided (specify format)
Submission	Data types accepted (list exceptions per repo)	● Datasets ● Metadata (supported upload of exchange formats (XML)) ● Computer code ● Other files

Category	Sub-category	Detailed features
		• Figures • Audio (MP3, WAV) • Video (MPG: MPEG2 for PAL, VLC, MP4: AAC, MPEG-4 for HDTV) • Photo (tiff, jpeg) • File sets • Formatted documents (PDF(A), odf, ASCII) • Geospatial (KML/KMZ, Web Map Service/Context, GeoRSS, GML) • Raster/Matrix • Vector • Most kinds of data (text, spreadsheets, video, photographs, software code, compressed archives of multiple files, non-data files) • Publications (Papers, Posters, Presentation) • Compressed (zip)
	Size (storage allocated, upload limits etc.)	• Specify size
	Metadata data submission (where applicable)	• Metadata • Other (local schema, discipline specific) • Digital Resource Description (Dublin Core, DataCite, MARC21) • Geospatial Metadata (ISO 19139, FGDC, ISO 19115, INSPIRE, etc.) • Health (NIH CDE etc.) • DDI (DDI v2, Lifecycle etc.) • Controlled language - terminology • Readme file (data description, definitions of column & row headings, data codes including missing data, units, data processing steps, contact info, etc.)
	Support	• Support for data preparation and quality control • Formal review and approval of submitted metadata and data before availability online
Access & Sharing	Online access	• Data available for free and open download (no registration, must anonymously "Agree" to terms of use)
	Web services	• API for harvesting & search access, Proprietary (REST or SOAP) • OAI PMH harvesting and search access, OAI-PMH exchange format • Other web service
	Exchange Metadata	• Exchange formats (XML, JSON)
	License	• Creative Commons License (Attribution or Zero) • Government License • Other license (open) • Other License (restricted)
	Linkages	• Linkage between data and publication, and / or citation indexes
Collaboration	Multiple user collaboration	• Collaboration (project workspace for multiple users)
Policy	Mandate	• Under what authority does the repository operate (i.e. governing entity)

Category	Sub-category	Detailed features
	Guidelines	● Terminology or Glossary of Terms
	Data sharing policy	● Data rights and usage statement (data for use) ● Data sharing policy available (data for deposit)
	Data deposit policy	● Terms & conditions by which data are ingested into the repository
	Data ownership policy	A repository's statement about ownership of the data it ingests
	Formats policy	● The digital formats accepted by the repository and whether of formats is performed
	Preservation policy	● A repository's statement about its preservation practices
	Succession plan	● Actions to be taken in the event that the repository is closed
Administration	Tracking	● Counts views & downloads
Tabular data	View Data	● Tabular data view, map view etc.
	File conversion formats	● File conversion options i.e. formats, projection etc.
	Download	● Download options
Certification status	Trusted Repository status	● ICSU World Data System ● Data Seal of Approval

TABLE 3. Summary of detailed features found in 32 online data platforms surveyed*

Features and data criteria	Yes	No	Not avail-able	Other
Cloud (i.e. Amazon S3)	7	8	17	
Free to access, download data, and deposit data	23	7	0	2
Free to access, but contribution suggested or required for deposit (i.e. funding structure for access / deposit beyond the threshold)	13	19		
Publishing charge $	8	24		
Formal agreement with research/monitoring program for funding to support archiving and serving their resulting datasets	7	23	2	
Size of repository (number of files, datasets)	Note 1			
Multidisciplinary	20	12		
Earth & environmental science	21	11		
Medical & life sciences	15	17		
Social Sciences (Economics, Sociology, Political Science, etc.)	17	15		
Biological and Life Sciences	17	14		
Physics	1	31		
Multiple redundant copies	14	10	7	1
LOCKSS/CLOCKSS - Geographically and geopolitically distributed network of redundant archive nodes	4	18	10	
DOI (specify where possible)	17	14	1	
DSpace Handle (HDL)	1	30	1	
Other persistent identifiers (URNs; PURLs)	1	2		
Other unique resource identifiers (i.e. IDs) (not persistent)	13	17	2	
EZID registration management or other persistent identifier registration	3	28	1	
Long term preservation of data	18	9	4	1
Data curation (specify where possible)	23	7	2	
Authentication mechanisms	26	3	3	
Distinction between public and private data	28	3	1	
ORCID ID	6	24	2	
SCOPUS ID	1	27	4	
Digital Author Identifier	1	30	1	
Timestamped upon upload	23	7	2	
Data can be edited once uploaded	22	3	7	
Version statement	13	12	7	
Universal Numeric Fingerprint (UNF)	1	29	2	
Citation provided (specify format)	20	12		
Datasets	31	1		
Metadata (supported upload of exchange formats (XML))	12	11	9	
Computer code	19	6	7	
Other files	22	4	6	
Figures	16	8	8	
Audio (MP3, WAV)	13	9	10	
Video (MPG: MPEG2 for PAL, VLC, MP4: AAC, MPEG-4 for HDTV)	12	10	10	
Publications (Papers, Posters, Presentation)	21	4	7	
File sets ("multiple related")	22	3	7	
Compressed (zip)	22	3	7	
Most kinds of data (text, spreadsheets, video, photographs, software code, compressed archives of multiple files, non-data files)	21	5	6	
Photo (tiff, jpeg)	20	4	8	
Formatted documents (PDF(A), odf, ASCII)	22	4	6	

Features and data criteria	Yes	No	Not avail-able	Other
Geospatial (KML/KMZ, Web Map Service/Context, GeoRSS, GML)	19	4	9	
Raster/Matrix	17	6	9	
Vector	19	4	9	
Specify size			19	
Metadata - Other (local schema, discipline specific)	22	7	3	
Metadata - Digital Resource Description (Dublin Core or DataCite)	11	20		
Metadata - Geospatial Metadata (ISO 19139, FGDC, ISO 19115, INSPIRE, etc.)	7	25		
Metadata - Health (NIH CDE etc.)	2	30		
Metadata - DDI (DDI v2, Lifecycle etc.)	6	26		
Metadata (Controlled language - terminology)	8	21	3	
Readme file (data description, definitions of column headings & row labels, data codes including missing data, units, data processing steps, contact info)	14	14	4	
Support for data prep and quality control	23	7	2	
Formal review and approval of submitted metadata and data before availability online	20	9	3	
Data available for free and open download (no registration required, must anonymously "Agree" to terms of use)	22	9	1	
API for harvesting & search access, Proprietory (REST or SOAP) API	16	14	2	
OAI PMH harvesting and search access, OAI-PMH exchange format	11	19	2	
Other web service	10	20	2	
Exchange formats (XML, JSON)	17	13	2	
Creative Commons (Attribution or Zero)	19	11	2	
Open Government License	9	21	2	
Other license (open)	20	11	1	
Other License (restricted)	19	11	2	
Linkage between data and publication, and / or citation indexes	18	10	4	
Collaboration (project workspace for multiple users)	15	15	2	
Under what authority does the repository operate (i.e. governing entity)	Note 1			
Terminology or Glossary of Terms	10	19	2	
Data rights and usage statement (data for use)	14	17	1	
Data sharing policy available (data for deposit)	21	10	1	
Terms and conditions by which data are ingested into the repository	21	7	4	
A repository's statement about the ownership of the data it ingests	19	10	3	
The digital formats accepted by the repository and whether normalization of formats is performed	18	9	4	
A repository's statement about its preservation practices	16	15	1	
Actions to be taken in the event that the repository is closed	2	26	4	
Counts views & downloads	15	14	3	
Tabular data view, map view etc.	18	12	2	
File conversion options i.e. formats, projection etc.	9	20	3	
Download available	27	3	2	
Certified as a trusted repository? (i.e. Data Seal of Approval)	7	24	1	

* The use of the checklist to identify majority practice (i.e. >50% across platforms) with respect to any feature or data criteria was still exceedingly difficult. Therefore, a lower threshold, arbitrarily set at 40%, was used as a more informative indicator of relatively common practice with respect to their implementation (or not). Relatively common practice across platforms is identified by the cells highlighted in colour in the Table.
Note1: See repository spreadsheet RDC-SINC 2015
Note 2: The majority indicated up to 2GB upload (remote submission)