A Balancing Act: The Ideal and the Realistic in Dryad's Preservation Policy Development

Sara Mannheimer, Ayoung Yoon, Jane Greenberg, Elena Feinstein, and Ryan Scherle

PEER-REVIEWED JOURNAL ON THE INTERNET

A balancing act: The ideal and the realistic in developing Dryad's preservation policy
by Sara Mannheimer, Ayoung Yoon, Jane Greenberg, Elena Feinstein, and Ryan Scherle

Abstract

Data preservation has gained momentum and visibility in connection with the growth in digital data and data sharing policies. The Dryad Repository, a curated general–purpose repository for preserving and sharing the data underlying scientific publications, has taken steps to develop a preservation policy to ensure the long–term persistence of this archived data. In 2013, a Preservation Working Group, consisting of Dryad staff and national and international experts in data management and preservation, was convened to guide the development of a preservation policy. This paper describes the policy development process, outcomes, and lessons learned in the process. To meet Dryad's specific needs, Dryad's preservation policy negotiates between the ideal and the realistic, including complying with broader governing policies, matching current practices, and working within system constraints.

Contents

Introduction

The growth in digital data and data–intensive computing has contributed to the phenomenon often referred to as "the data deluge" (Hey and Trefethen, 2003). These data are increasingly being recognized as valuable, and there are numerous efforts pushing for open access and facilitating data reuse. The goal of these efforts is not only to maintain but also to ensure the enduring value of such data. Recent research has also highlighted the benefits of open data by showing that data sharing leads to increased scientific discourse and discovery (Pienta, *et al.*, 2010), and increases citations and impact (Gleditsch, *et al.*, 2003; Piwowar and Vision, 2013).

Data preservation has gained momentum and visibility in connection with the growth in digital data and the increasing prevalence of data–sharing policies. Researchers archiving data want to confirm that it will be preserved in order to support data sharing and data reuse (Beagrie, *et al.*, 2010). National and international funding agencies have begun to require data management plans for grant applications (*e.g.*, the National Institutes of Health (NIH) in 2003; National Science Foundation (NSF) in 2011; Natural Environment Research Council in the U.K in 2011; Biotechnology and Biological Sciences Research Council in the U.K. in 2010). A February 2013 memo from the White House Office of Science and Technology Policy stated that data resulting from

> "unclassified research that are published in peer–reviewed
> publications directly arising from federal funding should be
> stored for long–term preservation and publicly accessible to
> search, retrieve, and analyze in ways that maximize the

impact and accountability of the Federal research investment" (Holdren, 2013).

All of these efforts, together with the rapid growth of data repositories, support the need for reliable repositories with effective preservation policies. The information and library science community has expert knowledge in these areas and has been engaged in advancing practices and policies specific to preserving research data. Sharing these developments is important. This paper serves to document the development of a preservation policy for the Dryad Repository, which stands out in the area of scholarly communication as a partner to scientific journals with established data archiving requirements (Whitlock, *et al.*, 2010).

This article begins with a description of the Dryad Repository, then reviews background information that has shaped Dryad's preservation policy. The central part of the paper presents the process of developing a preservation policy for Dryad, the challenges, and the lessons learned. The conclusion summarizes the paper's highlights and discusses the next steps.

The Dryad Repository

Dryad is a "curated, general–purpose repository that makes the data underlying scientific and medical publications discoverable, freely reusable, and citable" (Dryad Repository, 2013). Dryad was officially launched in in September 2008, funded by a National Science Foundation grant, as a joint project between the Metadata Research Center (MRC) at the University of North Carolina at Chapel Hill (UNC–CH) and the National Center for Evolutionary Synthesis (NESCent). NESCent is a collaboration involving Duke University (Duke), UNC–CH, and North Carolina State University (NC State). Dryad was envisioned as an easy–to–use, sustainable, community–governed data infrastructure, and preservation is a key element of Dryad's mission to ensure access, facilitate data availability, support data sharing, and enhance scholarly communication. Dryad accepts data underlying peer–reviewed publications, and all data files in Dryad are available for download and reuse, except those that are under a temporary embargo period, as permitted by editors of the relevant journals. As of June 2014, Dryad held data from more than 300 different journals. Journals are encouraged to become "integrated," a service that links the manuscript submission process with the data submission process for ease of deposit and richness of description. As of April 2014, Dryad had more than 70 integrated partner journals. Some of Dryad's earliest partners were a group of leading journals and scientific societies in evolutionary biology and ecology, whose affiliation with Dryad supported a Joint Data Archiving Policy (Whitlock, *et al.*, 2010). Since 2009, Dryad has been a membership organization, governed by a Board of Directors. In 2013, Dryad began transitioning away from its grant–funded status, incorporating as a not–for–profit and instituting a data publishing charge in order to support sustainability.

Literature review

Data preservation

Digital preservation is not a new concern in the library and archives fields, and considerable research has been done over the past 20 years regarding long–term preservation of digital information (Thibodeau, 2002; Burda and Teuteberg, 2013). The term digital preservation has been defined in various ways in the literature, but there is a general consensus in archives and preservation communities that digital preservation aims to ensure authenticity and access for digital objects over time. The Research Libraries Group (RLG) and Online Computer Library Center (OCLC) (RLG and OCLC, 2002) defined digital preservation as "the managed activities necessary for ensuring both the long–term maintenance of a bytestream and continued accessibility of its contents" [1]. The Joint Information Systems Committee's (Jisc) briefing paper (Pennock, 2006) provides a more detailed definition, noting that digital preservation encompasses "not just technical activities, but also all of the strategic and organizational considerations that relate to the survival and management of digital material" [2]. Jisc's definition supports the purpose of this paper, encouraging digital preservation activities not only in the course of digital object processing, but also on a broader policy level.

Digital preservation research has focused on a few key areas: intellectual property concerns (Charlesworth, 2012; Lor and Britz, 2012); technology obsolescence, including format/software and media/hardware (Caplan, 2007; Guttenbrunner and Rauber, 2012); physical threats such as bit rot, human error, and natural events (Weinstein, 1999; Martyniak, 2010); security (Qian, *et al.*, 2011); and description/metadata (Lavoie, 2004; Groenewald and Breytenbach, 2011). Many data preservation issues align with this existing research. However, data preservation also has its own

distinct challenges. Perhaps most challenging from a preservation perspective is that research data have a wide variety of formats (McGath, 2013). These formats may require rare or proprietary programs to interpret them, making emulation or migration difficult and expensive. Second, data are raw materials, rather than finished products; data creators need to be able to add new versions of datasets to the repository as they continue to conduct their research. Because data are the basis of scholarly output, they may not be immediately appropriate for public archiving; data may contain sensitive information, may not be formatted appropriately, or could potentially be used to scoop research findings (Laakso and Björk, 2013, Beagrie, *et al.*, 2010). Lastly, data sets can be very large and consist of multiple files; so while metadata creation is important to facilitate discoverability, it is also time consuming to curate large data sets (Greenberg, *et al.*, 2009).

Data repositories

Data curation and preservation in social science repositories dates back to the 1960s [3], but the scientific research community has been slower to enter the public discourse on data preservation. The earliest reports addressing scientific data sharing were published in the 1980s (Fienberg, *et al.*, 1985), and national organizations did not begin to publish about research data preservation until the 2000s (National Science Board, 2005; Friedlander and Adler, 2006). Since then, the growth of e–science and data–intensive computing have resulted in a "data deluge" — a term that has risen to popularity in both the academic and popular presses (Hey and Trefethen, 2003; Vardi, 2008; *Economist*, 2010). The massive and growing amounts of data have underscored the importance of data management and preservation.

GenBank (https://www.ncbi.nlm.nih.gov/genbank), one of the first data repositories on the Web, was established by the National Institutes of Health in 1982 to hold nucleic acid sequences. GenBank was used to archive Human Genome Project data in the 1990s–2000s, and it continues to accept sequencing data for more than 240,000 named organisms. Another wave of online scientific data repositories arose in the 2000s, including the Knowledge Network for Biocomplexity (https://knb.ecoinformatics.org). The number of repositories has continued to grow. The Registry of Research Data Repositories (http://www.re3data.org/) currently lists hundreds of scientific data repositories in the United States, and Marcial and Hemminger (2010) estimate the worldwide number of research data repositories to be in the thousands.

Preservation polices and trustworthy repositories

The Roman philosopher Lucius Annaeus Seneca wrote, "if one does not know to which port one is sailing, no wind is favorable" (Knowles, 2009). Written policy is vital to any organization in order to convey a clear mission, develop a strategic roadmap, and encourage a proactive approach to company practice. By drafting policy, an organization lays a foundation that facilitates future growth.

Policy development is essential to the success of project implementation, including digital preservation (Beagrie and Jones, 2008). Preservation policy is also required in order to be certified as a Trusted Digital Repository. The Consultative Committee for Space Data Systems (CCSDS) Audit and Certification of Trustworthy Digital Repositories (2011) explains, "documentation assures stakeholders (consumers, producers, and contributors of digital content) that the repository is meeting its requirements and fully performing its role as a trustworthy digital repository" [4]. Policy development also provides support for strategic planning, and it is critical that policy be continually reevaluated as repositories grow, holdings change, and preservation theory evolves (Bergmeyer, *et al.*, 2009).

While preservation policies have become increasingly commonplace in digital repositories over the past few years, data repositories lag behind the curve. Social science data repositories tend to have more robust policies [5]. However, there are fewer preservation policies for scientific data repositories. This may be because many scientific data repositories are developed by smaller, discipline–specific communities in order to serve their own needs [6], and it is more difficult for preservation policy development to happen on such small scales. Many existing scientific data preservation policies are associated with larger, more established data repositories. Four policies that informed Dryad's policy development are summarized in Table 1. Our selection of these four policies for review was based on the recommendation of the Dryad Preservation Working Group.

Table 1: Preservation policies informing Dryad.			
Repository name, location	**Focus**	**Last updated**	**Overview of policy**
			Like Dryad, ADS has internal policy documents and host institution

Archaeology Data Service (ADS), University of York, U.K. (http://archaeologydataservice.ac.uk/advice/preservation)	Archaeological data	27 January 2011	policies. The preservation policy points out that it "does not exist in isolation" (p. 1), but as part of a suite of policy documents, and in partnership with several organizations with whom ADS has service level agreements and memoranda of understanding. The ADS preservation policy is the only one surveyed here that acknowledges the Open Archival Information System (OAIS) reference model (CCSDS, 2012); it adheres closely to the model and relies heavily on OAIS terminology. It includes a guidance and implementation section that specifies preservation activities, indicating who in the organization conducts each activity.
European Organization for Nuclear Research (CERN): Large Hadron Collider, Geneva, Switzerland (https://cms-docdb.cern.ch/cgi-	Compact Muon Solenoid (CMS) data	3 May 2012	The policy identifies four levels of complexity in the CMS content, and outlines distinct preservation policy for each level of data. The policy also emphasizes CERN's commitment to open access, writing that "open access to the data will, in the long term,

bin/PublicDocDB/ShowDocument?docid=6032			allow the maximum realization of their scientific potential" (p. 1). CERN plans to archive data via third parties, so this policy does not include repository–specific preservation planning.
National Snow and Ice Data Center (NSIDC), Colorado, U.S. (link unavailable)	General scientific data	11 February 2004	The preservation policy acknowledges the importance of keeping policy up to date, explicitly stating that "to keep this document pertinent and effective, it should be reviewed frequently" (p. 3). However, reflecting the challenge of reconciling policy with practice, NSIDC's policy has not been revised since its development in 2004, leading to outdated technical information. Apart from this issue, the policy is robust and detailed. Its sections include: Data Solicitation and Acceptance, Levels of Service, Metadata and Data Format Standards, Data Set Documentation, Archive Policies, Architecture and Security, Data Tools, Data Deletion and Retirement, and Data

			Access.
DataOne, University of New Mexico, U.S. (http://mule1.dataone.org/ArchitectureDocs-current/design/PreservationStrategy.html)	General scientific data	December 2010	Of the policies surveyed, DataONE's preservation policy comes closest to the kind of general scientific data policy that Dryad aimed to develop in order to fit its broad collecting policy. DataONE's policy does not speak to specific implementation, but rather outlines a broad preservation practice using a three–tiered system: "keep the bits safe," "protect the form, meaning, and behavior of the bits," and "safeguard the guardians."

Marcial and Hemminger (2010) used Google–based searches to determine that 62 percent of scientific data repositories had "a clear mention of a preservation policy or similar" [7]. However, the authors of this paper found that a mention of preservation policy did not necessarily equate to a published policy. We found few written policies available online. We conclude that there is a continuing need for general scientific data preservation policies that can be used to guide other data repositories with similar holdings.

Data preservation policy development process

Dryad's preservation policy is the result of recent efforts that have built on basic, ongoing preservation activity. To understand the full activity, this section will first review basic, ongoing preservation practices for Dryad, then the process for development of a more sophisticated preservation policy are reviewed and, finally, Dryad's preservation best practice policy to date is presented.

Dryad's basic preservation practices

A number of basic preservation activities were already being implemented in Dryad before an official policy was developed.

1. An MD5 checksum is created for every bitstream upon upload. Checksums are verified nightly, and a verification email is sent to repository administrators.
2. Provenance metadata is automatically created as submissions progress through the Dryad workflow. This metadata provides the name and e–mail address of each actor — usually the depositor and Dryad curators; the date that each action occurred; and a description of the action. Common actions are submission, rejection with documented reason, acceptance into publication blackout, and archiving into DSpace. The provenance metadata is private, and is only visible to logged–in administrators. One problem with the current creation of provenance metadata is that these fields are only automatically created when a submission progresses normally through the workflow. In rare cases, the Dryad Curator moves submissions in

uncommon directions across workflow states. In the future, we hope that provenance metadata can be automatically generated for all actions. Also, until February 2014, all assistant curators worked from the same Dryad account; unique accounts for each assistant curator provide more detailed provenance information.

3. Historically, Dryad curators have informally encouraged preferred formats. This practice developed organically and is used sparingly. The most common reason for a curator to request an alternative file type from a depositor is if the submitted file is unreadable by free software, and the curator is therefore unable to review the file. When the opportunity arises, preferred file types are also encouraged over the course of normal correspondence with depositors.

4. Dryad's servers are hosted by NC State, where regular maintenance is conducted. There is also a server mirror at Duke, which is updated every five minutes and has its own backup processes.

5. In February 2014, Dryad finalized partnerships with DataONE (http://www.dataone.org/) and CLOCKSS (http://www.clockss.org/clockss/Home), which will replicate Dryad's entire content.

Policy development process

Dryad's preservation policy builds on the basic practices reviewed above. The first preservation policy was drafted in 2012 by then–curator Elena Feinstein (version 1.0), based on a Digital Curation Centre template (http://www.dcc.ac.uk/sites/default/files/documents/Preservation%20policy%20template.pdf), and with the assistance of Dryad staff. To further pursue the preservation mission, Dryad convened a Preservation Working Group (http://wiki.datadryad.org/Preservation_working_group_2013), comprised of preservation experts in the United States and United Kingdom. The Working Group's initial meeting reviewed Dryad's current practices, identified preservation priorities, and outlined broad goals for policy development. Official policy development was initiated in February 2013, led by two authors of this paper (Sara Mannheimer and Ayoung Yoon). See Table 2 for main phases of policy development.

Table 2: Three main phases of policy development.		
Phase 1	**Phase 2**	**Phase 3**
Literature review of digital preservation literature and practice	Review of other data repositories' preservation activities	Drafting the preservation policy document, informed by the Open Archival Information Systems (OAIS) reference model, but ultimately structured according to Dryad's specific needs

The Working Group recommended resources and advised on several drafts of the policy. Version 2.0 was finalized and presented to the Dryad Board of Directors in May 2013. Taking into account the Board's responses, the policy was revised in cooperation with Dryad staff (July 2013) and the Preservation Working Group (August 2013). In November 2013, the Board reviewed and approved the policy, dissolved the Preservation Working Group, and made plans to convene a new Preservation Task Force in 2014 to address specific implementation issues, including versioning (including pricing, DOI considerations, and corrections vs. concatenations), preferred file formats, file type migration strategies, and implementation budget.

Results: Dryad preservation policy

The final preservation policy is structured as follows:

1. Purpose
   - The policy states that its purpose is "to ensure authenticity, reliability, and integrity of research data over the long term so that data can be re–used for research, education, or any other purpose."

2. Scope and content coverage
   - Content criteria (corresponds to 3.1 in Dryad's Terms of Service)

— Must underlie a publication
— No sensitive or illegal material
— No technical problems or viruses

3. Overview and preservation strategy
- Bit–level preservation
- Collection of adequate descriptive information to facilitate preservation
- File–meaning preservation: maintaining the functionality, usability, and understandability of the content

4. Format support and levels of preservation
- Preferred file formats: non–proprietary and open source
- Full list of preferred formats and support levels is still pending and will be finalized by the 2014 Preservation Task Force

5. Implementing the strategy
- Procedures for preservation
  — Ingest, curation and publication: current practice, short–term plan, long term plan
  — Archival storage: current practice, short–term plan, long–term plan
- Data integrity
  — Fixity checks, replication
- Security
  — Server room security
  — IT infrastructure
- Versioning
  — This section will be clarified and expanded by the 2014 Preservation Task Force
- Withdrawal of content
  — Withdrawal "should be rare and is not part of the normal life cycle for any archived file."

6. Sustainability plans
- Technical sustainability
  — Dryad software: open source, periodically synchronized with DSpace
  — Dryad data: in the event that Dryad is no longer able to maintain content at the primary location, responsibility will pass to a mirror location
- Persistent data accessibility: DOIs
- Institutional and financial sustainability
  — Governance, business model, and sustainability plan for long term organizational stability and viability: Not–for–profit, overseen by a membership–elected board, partnership between UNC–CH, Duke, and NC State
  — Dryad members have a strong interest in seeing repository persist
  — Participation in DataONE network ensures future data availability

---

Discussion

Development of data preservation policy is still a relatively new activity, and we believe it is important to share what was learned from the policy development experience. The following discussion shares specific lessons learned, as well as some open questions raised from the experience.

1. Negotiating what is ideal and what is realistic
Policy development is a negotiation between what is ideal and what is realistic. International standards, models, and best practices exist for long–term preservation. While these standards provide useful insights and guidance for the big picture, implementation should be considered within the boundaries of organizational capability, and adopting these standards also requires embedding local context during implementation.

At the initial policy development stage, we identified standards and models relevant to the Dryad preservation policy, including the OAIS reference model (ISO 14721:2003) (CCSDS, 2012) and PREservation Metadata: Implementation Strategies (PREMIS) (PREMIS Editorial Committee, 2012). Due to its fundamental significance for developing repositories for long–term preservation, Dryad adopted the OAIS model and concepts at a high level. Dryad preservation policy uses the definition of "long term" as stated in the OAIS reference model: "a period of time long enough for there to be concern about the impacts of changing technologies, including support for new media and data formats, and of a changing user community, on the information being held in a repository." [8] Because the model is conceptual and high level, however, sometimes there was a need to clarify the terms and localize the use of concepts in the context of Dryad. For instance, while the initial Dryad preservation policy included statements about SIPs and AIPs, a question was raised by Dryad staff about how to distinguish SIPs and AIPs during the day–to–day curation activities already in place. PREMIS was also one of Dryad's considerations from the early development of this project. However, PREMIS did not align with the priorities at that time, and as already discussed, Dryad already had provenance metadata requirements for datasets. Thus, full implementation of PREMIS remains a long–term goal for Dryad.

Dryad is also aware of other standards and guidelines about audit and certification for building a trusted digital repository, such as *Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC)* (Ambacher, *et al.*, 2007) and *Data Seal of Approval (DSA)* (Sesink, *et al.*, 2008), and has kept certification as a consideration when developing policies and procedures. However, Dryad's current efforts will focus more on putting the planned preservation infrastructure in place, such as format migration, before going through a formal self or external auditing process.

2. Aligning internal and external policies
A second lesson learned was that preservation policy must align with other internal and institutional policies. In order to follow Dryad's internal policies, we looked primarily to Dryad's Terms of Service document (https://datadryad.org/pages/policies), which includes policies on submission, content, payment, usage, and privacy. We aimed to have no overlap between the preservation policy and the Terms of Service, and therefore chose to link to the Terms of Service when necessary.

We also aimed to comply with Dryad's unofficial policies, and policies that have yet to be finalized. One example of a policy–in–progress is Dryad's policy on versioning. An initial policy on versioning was developed in the first quarter of 2009 (http://wiki.datadryad.org/Track_Version_Changes). The versioning feature is currently deployed for curator use only, and in a few rare cases, updated files have been submitted. However, the tool is still in the process of being streamlined and debugged before being deployed publicly. The versioning policy is currently being revisited by the Dryad Board of Directors, and will be one of the specific issues addressed by the 2014 Preservation Task Force.

Dryad has strong operational activities at both the Metadata Research Center (UNC–CH) and Duke University, and receives server support from NC State. It follows that Dryad's policies are influenced, in part, by the policies at each of these three institutions. Knowledge of the policies at NC State was especially important, since Dryad's servers are maintained and housed there. Historically, Dryad has been relatively hands–off with the servers, but in order to ensure preservation, NC State's policies and procedures surrounding server maintenance and security must agree with Dryad's preservation policy.

3. Meeting Dryad's specific needs
Preservation policy should ultimately be structured according to Dryad's specific needs. Meeting specific organizational needs is fundamentally important and should be the first consideration in all work, as each organization has different goals, priorities, and capabilities.

This becomes apparent when preservation policy deals with data depositors' requirements. Dryad has minimal requirements for depositors, which might seem to be a very different approach compared to some other data repositories that require more submission information. Making submission as easy as possible for data depositors has been important in supporting Dryad's mission of encouraging data sharing through open and widespread availability. The minimal requirements may lead to questions, such as: is the minimal representation information sufficient for preservation and reuse? All self–deposit repositories must strike a balance between "minimum efforts" and having "enough" representation information. Dryad's decision to require minimum efforts from depositors is compensated, on some level, by other factors related to Dryad's submission process. Unlike some other data repositories, Dryad only accepts data that are linked to papers or publications, making it possible to acquire metadata mandated by journals. This link provides a richer context for understanding the data in Dryad, and helps to enhance the metadata experience from both curation

and reuse perspectives.

4. DSpace constraints
Dryad is built on a DSpace platform, and some development–related constraints are unavoidable. Thus far, Dryad has been committed to customizing DSpace to function according to its specifications, developing whatever special features are required to create a database with wide functionality. However, many preservation activities (*i.e.*, PREMIS implementation, versioning, and format migration) would require substantial additional customization of DSpace, and could therefore require extensive developer hours to develop, implement, and debug. Developer time is spread between many projects at Dryad, so prioritizing time for preservation is an issue that must be addressed by Dryad administrators.

Conclusion

Dryad's initial steps in developing a preservation policy required compromise and balance. International standards, models, and best practices that exist for long–term preservation (*i.e.*, the OAIS reference model and PREMIS) provide valuable guidelines for preservation activities. During our policy development process, we came to understand that some departure from these standards was necessary, and that policy development would need to unfold in increments in order for long–term preservation practices to be manageable into the future.

The development of Dryad's preservation policy is just the first step in an ongoing process of preservation efforts. The preservation policy, as a living document, should evolve alongside the repository. Dryad will make an ongoing effort to adopt preservation standards, models, and best practices, and the policy will be continually revised and re–evaluated as preservation practices are refined. A key goal of the 2014 Preservation Task Force will be to develop a strategic plan to complement the preservation policy. After the strategic plan is in place to bridge policy and practice, TRAC or DSA may be implemented.

Preservation policy does not exist in a vacuum; it must interoperate with other organizational functions. To address this, we aimed to structure Dryad's preservation policy to fit the real world of the Dryad Repository, including broad governing policies, current practices, and system constraints. We also designed the policy to be aspirational. We included short– and long–term goals in order to facilitate growth and begin to mold Dryad's workflow to approach the ideal.

As documented in Table 1, Dryad's policy development was aided significantly by a review of existing data preservation policies. However, at this point in time, the Dryad staff were unable to identify a policy that could be immediately adopted. Just as existing policies informed our process, Dryad's policy has the potential to inform preservation policy development at other repositories. ▮

About the authors

Sara Mannheimer is the Data Management Librarian at Montana State University, supporting research data services and promoting open access. Her research focuses on digital preservation, digital curation, and data sharing. She was Dryad's curator from June 2013–February 2014.
E–mail: sara [dot] mannheimer [at] montana [dot] edu

Ayoung Yoon is a doctoral candidate at the School of Information and Library Science, University of North Carolina at Chapel Hill. Her research interests include data reusers' trust in data and data repositories, data curation, and personal digital archiving. She has an M.S.I. in both preservation and archives and record management from the University of Michigan School of Information, and B.A. in history from Ewha Womans University, South Korea.
E–mail: ayyoon [at] email [dot] unc [dot] edu

Jane Greenberg is a professor at the School of Information and Library Science and the director of the Metadata Research Center at the University of North Carolina at Chapel Hill. She is co–Principal Investigator of the Dryad grants. Her research focuses on metadata, knowledge organization/ontological engineering, data science, and linked data.
E–mail: janeg [at] email [dot] unc [dot] edu

Elena Feinstein was Dryad's curator from 2009–2013. She is currently the Librarian for Chemistry and Biological Sciences at Duke University, where she continues to promote data sharing and open science.
E–mail: elena [dot] feinstein [at] duke [dot] edu

Ryan Scherle is the repository architect at Dryad Repository. He has a Ph.D. in computer science

from Indiana University, Bloomington.
E-mail: ryan [at] datadryad [dot] org

Acknowledgements

Notes

1. RLG and OCLC, 2002, p. 3.

2. Pennock, 2006, p. 1.

3. See the IASIST bibliography for details at http://www.iassistdata.org/publications/bibart.html.

4. Consultative Committee for Space Data Systems (CCSDS) Audit and Certification of Trustworthy Digital Repositories, 2011, p. 3–5.

5. *E.g.*, Odum Institute, at http://www.irss.unc.edu/odum/contentSubpage.jsp?nodeid=629 and ICPSR, at
http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/preservation/policies/index.html.

6. For example, www.re3data.org lists eight unique repositories for data on the ecology of agricultural landscapes, and 50 repositories for immunology data.

7. Marcial and Hemminger, 2010, p. 2,038.

8. Consultative Committee for Space Data Systems (CCSDS) Reference Model for an Open Archival Information System (OAIS), 2012, p. 1–1.

References

Bruce Ambacher, Kevin Ashley, John Berry, Connie Brooks, Robin L. Dale, Dale Flecker, David Giaretta, Babak Hamidzadeh, Keith Johnson, Maggie Jones, Nancy McGovern, Andrew McHugh, Don Sawyer, and Johan Steenbakkers, 2007. *Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC)*. Chicago: Center for Research Libraries, at
http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf, accessed 23 April 2014.

Neil Beagrie and Maggie Jones, 2008. *Preservation management of digital materials: The handbook*. Heslington, York: Digital Preservation Coalition.

Neil Beagrie, Brian F. Lavoie, and Matthew Woollard, 2010. *Keeping research data Safe 2*. Bristol: Higher Education Funding Council for England (HEFCE).

Winfried Bergmeyer, Susanne Dobratz, Andrea Hanger, Karsten Huth, Christian Keitel, Jens Klump, Nikola Korb, Peter Rodig, Stefan Rohde–Enslin, Astrid Schoger, Kathrin Schoder, Tobias Steinke,

Stefan Strathmann, and Heidrun Wiesenmuller, 2009. *Catalogue of criteria for trusted digital repositories, Nestor criteria*. Frankfurt: Deutsche Nationalbibliothek, at http://nbn-resolving.de/urn:nbn:de:0008-2010030806, accessed 23 April 2014.

Biotechnology and Biological Sciences Research Council, 2010. *BBSRC data sharing policy*. Swindon: BBSRC, at http://www.bbsrc.ac.uk/web/FILES/Policies/data-sharing-policy.pdf, accessed 23 April 2014.

Daniel Burda and Frank Teuteberg, 2013. "Sustaining accessibility of information through digital preservation: A literature review," *Journal of Information Science*, volume 39, number 4, pp. 442–458.
doi: http://dx.doi.org/10.1177/0165551513480107, accessed 27 July 2014.

Priscilla Caplan, 2007. "The Florida Digital Archive and DAITSS: A working preservation repository based on format migration," *International Journal On Digital Libraries*, volume 6, number 4, pp. 305–311.
doi: http://dx.doi.org/10.1007/s00799-007-0009-6, accessed 27 July 2014.

Andres Charlesworth, 2012. "Intellectual property rights for digital preservation," In: Charles Beagrie Ltd. (editors). *DPC Technology Watch Report 12–02*. Heslington, York: Digital Preservation Coalition.
doi: http://dx.doi.org/10.7207/twr12-02, accessed 27 July 2014.

Consultative Committee for Space Data Systems, 2012. *Reference Model for an Open Archival Information System (OAIS): CCSDS 650.0–M–2 Magenta Book*, at http://public.ccsds.org/publications/archive/650x0m2.pdf, accessed 5 August 2014.

Consultative Committee for Space Data Systems, 2011. *Audit and Certification of Trustworthy Digital Repositories: CCSDS 652.0–M–1 Magenta Book*, at http://public.ccsds.org/publications/archive/652x0m1.pdf, accessed 23 April 2014.

Dryad Repository, 2013. "The organization: Overview," at http://datadryad.org/pages/organization, accessed 23 April 2014.

Economist, 2010. "The data deluge," *Economist*, volume 394, number 8671 (27 February), p. 11, and at http://www.economist.com/node/15579717, accessed 27 July 2014.

Stephen E. Fienberg, Margaret E. Martin, and Miron L. Straf (editors), 1985. *Sharing research data*. Washington, D.C.: National Academies Press, at http://www.nap.edu/openbook.php?record_id=2033, accessed 27 July 2014.

Amy Friedlander and Prudence Adler, 2006. "To stand the test of time: Long term stewardship of digital data sets in science and engineering," Washington, D.C.: Association of Research Libraries, at http://arl.nonprofitsoapbox.com/storage/documents/publications/digital-data-report-2006.pdf, accessed 23 April 2014.

Nils Petter Gleditsch, Claire Metelits, and Håvard Strand, 2003. "Posting your data: Will you be scooped or will you be famous?" *International Studies Perspectives*, volume 4, number 1, pp. 89–97.

Jane Greenberg, Hollie C. White, Sarah Carrier, and Ryan Scherle, 2009. "A metadata best practice for a scientific data repository," *Journal of Library Metadata*, volume 9, numbers 3–4, pp. 194–212.
doi: http://dx.doi.org/10.1080/19386380903405090, accessed 27 July 2014.

Ria Groenewald and Amelia Breytenbach, 2011. "The use of metadata and preservation methods for continuous access to digital data," *Electronic Library*, volume 29, number 2, pp. 236–248.
doi: http://dx.doi.org/10.1108/02640471111125195, accessed 27 July 2014.

Mark Guttenbrunner and Andreas Rauber, 2012. "A measurement framework for evaluating emulators for digital preservation," *ACM Transactions on Information Systems*, volume 30, number 2, article number 14.
doi: http://dx.doi.org/10.1145/2180868.2180876, accessed 27 July 2014.

Anthony J.G. Hey and Anne E. Trefethen, 2003. "The data deluge: An e–science perspective," In: Fran Berman, Geoffrey Fox, and Anthony J.G. Hey (editors). *Grid computing: Making the global infrastructure a reality*. New York: Wiley, pp. 809–824.
doi: http://dx.doi.org/10.1002/0470867167.ch36, accessed 27 July 2014.

John P. Holdren, 2013. "Increasing access to the results of federally funded scientific research," *Memorandum for the Heads of Executive Departments and Agencies, Executive Office of the President, Office of Science and Technology Policy* (22 February), at http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf, accessed 24 April 2014.

Elizabeth Knowles (editor), 2009. *Oxford dictionary of quotations*. Seventh edition. Oxford: Oxford University Press.

Mikael Laakso and Bo–Christer Björk, 2013. "Delayed open access: An overlooked high-impact category of openly available scientific literature," *Journal of the American Society for Information Science and Technology*, volume 64, number 7, pp. 1,323–1,329.
doi: http://dx.doi.org/10.1002/asi.22856, accessed 27 July 2014.

Brian F. Lavoie, 2004. "Implementing metadata in digital preservation systems: The PREMIS activity," *D–Lib Magazine*, volume 10, number 4, at http://www.dlib.org/dlib/april04/lavoie/04lavoie.html, accessed 27 July 2014.
doi: http://dx.doi.org/10.1045/april2004-lavoie, accessed 27 July 2014.

Peter Johan Lor and J.J. Britz, 2012. "An ethical perspective on political–economic issues in the long–term preservation of digital heritage," *Journal of the American Society for Information Science and Technology*, volume 63, number 11, pp. 2,153–2,164.
doi: http://dx.doi.org/10.1002/asi.22725, accessed 27 July 2014.

Laura Haak Marcial and Bradley M. Hemminger, 2010. "Scientific data repositories on the Web: An initial survey," *Journal of the American Society for Information Science and Technology*, volume 61, number 10, pp. 2,029–2,048.
doi: http://dx.doi.org/10.1002/asi.21339, accessed 27 July 2014.

Cathleen Martyniak, 2010. "When preservation moves off campus: Trends and effective practices in ARL libraries," *Library Resources & Technical Services*, volume 54, number 4, pp. 183–199.

Gary McGath, 2013. "The format registry problem," *Code4Lib Journal*, number 19, at http://journal.code4lib.org/articles/8029, accessed 27 July 2014.

National Institutes of Health, 2003. "Data sharing policy and implementation guidance" (5 March), at http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm, accessed 24 April 2014.

Natural Environment Research Council, 2011. "Data policy," at http://www.nerc.ac.uk/research/sites/data/policy/, accessed 24 April 2014.

National Science Board, 2005. "NSB–05–40, Long–lived digital data collections enabling research and education in the 21st century," at http://www.nsf.gov/pubs/2005/nsb0540, accessed 24 April 2014.

National Science Foundation, 2011. "NSF data management plan requirements," at http://www.nsf.gov/eng/general/dmp.jsp, accessed 24 April 2014.

Maureen Pennock, 2006. "Continued access to authentic digital assets," at http://jisc.ac.uk/publications/briefingpapers/2006/pub_digipreservationbp.aspx, accessed 24 April 2014.

Amy M. Pienta, George C. Alter, and Jared A. Lyle, 2010. "The enduring value of social science research: The use and reuse of primary research data," paper presented at "The Organisation, Economics and Policy of Scientific Research" Workshop, Torino, Italy, in April, at http://hdl.handle.net/2027.42/78307, accessed 24 April 2014.

Heather A. Piwowar and Todd J. Vision, 2013. "Data reuse and the open data citation advantage," *PeerJ*, volume 1, p. e175.
doi: http://dx.doi.org/10.7717/peerj.175, accessed 27 July 2014.

PREMIS Editorial Committee, 2012. *PREMIS Data Dictionary for Preservation Metadata*, version 2.2, at http://www.loc.gov/standards/premis/v2/premis-2-2.pdf, accessed 5 August 2014.

Kim Qian, Maik Schott, Christian Kraetzer, and Matthias Hemmje, 2011. "Contextualizing security for digital long–term preservation," *MM&Sec '11: Proceedings of the Thirteenth ACM Multimedia Workshop on Multimedia and Security*, pp. 115–120.
doi: http://dx.doi.org/10.1145/2037252.2037274, accessed 27 July 2014.

Research Libraries Group and Online Computer Library Center, 2002. "Trusted digital repositories: Attributes and responsibilities," at http://www.oclc.org/content/dam/research/activities/trustedrep/repositories.pdf, accessed 24 April 2014.

Laurents Sesink, René van Horik, and Henk Harmsen, 2008. "Data seal of approval," at http://www.datasealofapproval.org/en/, accessed 24 April 2014.

Kenneth Thibodeau, 2002. "Overview of technological approaches to digital preservation and challenges in coming years," *The state of digital preservation: An international perspective:*

Conference proceedings, at http://www.clir.org/pubs/reports/pub107/thibodeau.html, accessed 24 April 2014.

Moshe Y. Vardi (editor), 2008. "Surviving the data deluge" (special issue). *Communications of the ACM*, volume 51, number 12, pp. 50–74.

Lauren Weinstein, 1999. "Bit–rot roulette," *Communications of the ACM*, volume 42, number 3, p. 144.
doi: http://dx.doi.org/10.1145/295685.295727, accessed 27 July 2014.

Michael C. Whitlock, Mark A. McPeek, Mark D. Rausher, Loren Rieseberg, and Allen J. Moore, 2010. "Data archiving," *American Naturalist*, volume 175, number 2, pp. 145–146.
doi: http://dx.doi.org/10.1086/650340, accessed 27 July 2014.

Editorial history