

2017

Extending Data Curation Service Models for Academic Library and Institutional Repositories

Jon Wheeler

University of New Mexico - Main Campus, jwheel01@unm.edu

Follow this and additional works at: https://digitalrepository.unm.edu/ulls_fsp



Part of the [Scholarly Communication Commons](#)

Recommended Citation

Wheeler, Jonathan. "Extending Data Curation Service Models for Academic Library and Institutional Repositories." In *Curating Research Data*, ed. Lisa Johnston 1:171–92. Chicago, Illinois: Association of College and Research Libraries, 2017.

This Book Chapter is brought to you for free and open access by the Scholarly Communication - Departments at UNM Digital Repository. It has been accepted for inclusion in University Libraries & Learning Sciences Faculty and Staff Publications by an authorized administrator of UNM Digital Repository. For more information, please contact disc@unm.edu.



CHAPTER 7*

Extending Data Curation Service Models for Academic Library and Institutional Repositories

Jon Wheeler

Introduction

Development of research data management (RDM) and curation services remains both a priority and a challenge for many academic research libraries. Broadly speaking, while library service models continue to evolve to meet the data management needs of researchers accountable to emerging funder requirements, it remains true that many librarians seek clarification about their role in support of data curation.¹ Discussion by Antell and colleagues and Nielsen and Hjørland highlight in particular some of the contradictions librarians perceive between the drive to develop data management skills and the practicality of situating these skills within libraries generally.²

A similar contrast exists between the technical capabilities and expectations regarding the use of institutional repositories for data publication and preservation.

* This work is licensed under a Creative Commons Attribution 4.0 License, CC BY (<https://creativecommons.org/licenses/by/4.0/>).

Because IRs may include content from multiple disciplines and a variety of types—for example electronic theses and dissertations (ETD) and research posters—their utility as data repositories may be legitimately called into question. As noted by Don MacMillan, IRs as data repositories may “further fragment the data landscape and may result in making data more difficult to find than it would be in larger subject-specific or interdisciplinary repositories.”³ Even in cases where one may argue that an IR is a better-than-nothing option, issues described in McGovern and McKay and Jain illustrate how the diversity of IR service models can inhibit their utility when publication workflows are not based on best practices or otherwise modeled against disciplinary standards.⁴ However, with such concerns in mind, a consideration of established library data management services, functions, and roles provides context for the description and development of an IR data service focused on archiving and mirroring collections previously published within domain repositories. Beginning with an overview of the suitability of IRs for this purpose, the chapter further addresses how such a service aligns with existing capabilities and provides illustrative scenarios and strategies for implementation.

Conceptual Models and Rationale

Establishing IRs as mirrors of data collections held by domain repositories is a service capability described at least implicitly in the literature. In particular, the “web of repositories” model presented by Baker and Yarmey and elaborated upon by Baker and Millerand is relevant because of the model’s emphasis on situated, role-based services oriented toward data management within local and nonlocal contexts.⁵ Locality and distance are here understood not as spatial or geographic constraints, but rather refer to a repository’s support for or contribution to data management at different stages in the research life cycle. Additionally, the model is understood to be nonlinear in the sense that data do not necessarily move in sequence from one repository setting to another, but may be hosted or mirrored across systems. As the context changes, so does the community served by the corresponding repository, which necessarily impacts the services provided to manage data in that context. A recent and innovative example of this model is provided by Walters, in which IRs are not preservation end points in themselves, but rather act as communication layers between production services and preservation architectures.⁶

The proposed mirroring service is further informed by the “preservation as relay” model described by Janée and colleagues.⁷ Whereas the web of repositories model explicitly includes mirroring collections across multiple sites, the preservation as relay model more narrowly refers to a complete handoff or asset transfer wherein different types of repositories fulfill preservation requirements at different times. As an example, a short- or near-term repository may commit to providing services or taking necessary actions to transform, migrate, or curate data for five

years. In the event long-term support is required, the migration to another repository altogether, perhaps one with a five-to-ten-year remit, becomes a relay or a handoff. This model includes dark archiving, or periods during which no archive is able to provide public access to the data, with the expectation that appropriate preservation practices are in place to successfully recover the data when necessary.

While these two models provide a broad rationale for establishing IRs as complementary services to domain repositories, further justification for a mirroring service is provided by drivers including funder policies and the DR ecosystem. Policy-wise, the specification within funder data management plan (DMP) recommendations of “data archiving” or “data preservation” as distinct from “data sharing” strategies is relevant (see for example recommendations from the National Science Foundation and the Department of Energy⁸), as the practical difference between publishing and archiving—not to mention between archiving and backup—is not intuitive across disciplines. This is a significant issue, as a lack of distinction between these concepts can result in noncompliance and put data at risk. An illustrative example is provided by Choudhury, who relates how project team members from the Sloan Digital Sky Survey assumed that their data were sufficiently archived because they had been securely stored and backed up.⁹ Even putting aside compliance concerns, the risk of data loss in such circumstances is further illustrated by Uhler’s concept of “information gulags” in which data are “preserved” within systems that are “highly distributed, silent, and invisible.”¹⁰ Here the conflation of archival with sharing and backup processes contributes to a proliferation of these invisible data silos when systems and strategies are adopted that negatively impact the discoverability and usability of data. Establishing IRs as complementary archives of DR collections is one means of preventing information gulags by enlarging the context of discovery, accessibility, and exposure of data to users.

Finally, further practical justification is found among concerns about the sustainability and preservation-readiness of many DRs. As noted in the literature, coverage for data curation across the life cycle is well established within disciplines such as astronomy and certain subfields of biology.¹¹ However, the existence of established, trustworthy repositories across disciplines is the exception rather than the rule. This is a two-fold problem in that a given discipline may on the one hand lack established repositories, while on the other hand available repositories may not provide sufficient preservation support to satisfy funder expectations. For example, Castelli and colleagues enumerate multiple barriers that impact data discovery and preservation among data centers and research digital libraries.¹² In particular, that data may be documented only enough to support discovery or citation,¹³ and the protocols in place for export or federation of resources may be limited or otherwise not based on best practices or standards such as OAI-PMH.¹⁴ Sustainability concerns due to loss of funding are likewise an ongoing concern.¹⁵

Alignment with Existing Roles and Capabilities

In addition to exploring the overall suitability of IRs to mirror DR collections, we further consider how the proposed service model aligns with data management roles and activities among libraries and librarians. Alignment is here considered from multiple perspectives, including administrative-level collaborations, the participation of functional and subject area librarians, and system capabilities.

With regard to collaboration, the development of sustainable data services can benefit from the engagement of library administration with stakeholders from their respective campus IT units and sponsored research offices. As a notable example, Witt describes the development of the Purdue University Research Repository (PURR),¹⁶ an effort which was steered by a working group whose members included, among others, the Associate Vice President for Research, two Associate Deans from the Libraries, liaison librarians, and technical specialists.¹⁷ Similarly composed groupings are proposed by Block and colleagues at Cornell University,¹⁸ and the 2012 ACRL study by Tenopir and colleagues likewise highlights the experience among library directors that sponsored research units in particular are necessary contributors to the development of impactful RDM services.¹⁹

By collaborating at administrative levels to strategically position data and repository services within the research practice of an institution, the identification and promotion of the IR as a complementary service to DRs can become part of the research planning strategy. For example, Choudhury notes a particularly promising outcome of engaging university administrators in the development of the Johns Hopkins University Data Management Services (JHUDMS).²⁰ As a demonstration of the anticipated value the service may provide to researchers, the JHU administration opted to directly fund preproposal consultations between JHUDMS and researchers applying for grants. This consultation includes a review of domain repository options together with information about the JHU Data Archive.²¹ Optional, grant-funded post-award services are also available that can include eventual transfer of data to the archive. This and similar arrangements are of particular import to the service proposed here as they logically extend to proactively defining complementary roles between DRs and IRs. By thoroughly reviewing repository options with researchers and mapping repository capabilities and features to different phases of the data life cycle, librarians are positioned to make strategic recommendations about when and under what circumstances the IR represents a viable option for data archiving. Although such consultations represent librarian rather than administrator activity, the sponsorship of the JHUDMS by university administration in this case demonstrates how a successful collaboration can lead to better promotion of the IR and facilitate collection development.

At the grassroots level, discussions of librarian roles in support of data curation may distinguish between subject area and functional expertise.²² While both contexts may overlap within particular positions, a pairing or collaboration between subject and functional specialists as described by Jaguszewski and Williams is a promising strategy for providing both the domain and technical expertise to effectively support researchers.²³ For example, the composition of data curation project teams at Purdue, as reported by Newton and colleagues, demonstrate a distribution of functional skills and subject area expertise across an organization.²⁴ Other models exist, but the overall implication for IR building in this context is the importance of linking tangible capabilities with researcher needs.

On the functional side, as described for example by Tenopir and colleagues, Sands and colleagues, and Lyon, services performed by IR managers and data curation librarians can include transforming proprietary files to open file formats, conducting file integrity and format validation routines, creating or transforming metadata, and packaging data for submission to the IR.²⁵ These processes and activities will necessarily be important components of an IR data mirroring service. However, as noted by Kim, there remains nonetheless a growing imperative for technical assistance and “a more proactive role in support of digital scholarship” that is relevant to extending IR services.²⁶ Because the proposed model is focused on the batch transfer and repository ingest of complete data set collections, it may be necessary to scale up workflows that are currently oriented toward the curation of single or small collections of data sets. At minimum, adapting workflows in this way will require some scripting capabilities and familiarity with application programming interfaces (APIs).

It has likewise been shown that IR managers and data curation librarians are not necessarily technicians and that the duties of librarians in these positions may focus on assisting researchers with the identification and implementation of best practices in content, data, and metadata management. Lyle and colleagues, for example, describe a series of collaborations between the Inter-university Consortium for Political and Social Research (ICPSR) and multiple IRs to curate and publish legacy datasets.²⁷ Noting at the outset that many IR managers have “limited experience dealing with quantitative or qualitative data,”²⁸ the authors proceed through a series of case studies that highlight the types of functional support IR managers may need in preparing data for archiving. However, in lieu of technical skills, the strengths in relationship and resource building that participating librarians brought to the case studies indicated that IR managers and data curation librarians are well-positioned to mediate between data owners and developers or technicians in support of collection-scale curation and archiving.²⁹

Established data management activities of subject area librarians can be likewise aligned with the proposed IR data mirroring service. As reported by Antell and colleagues, data management skills practiced with some regularity among librarians include consultation about DRs as well as providing information about

data life cycle management and funder requirements.³⁰ Similar to the JHUDMS example above, such consultations provide an opportunity for librarians to identify publication and archiving requirements that a local IR may appropriately provide in the absence of, or in addition to, an established DR. Additionally, Newton and colleagues described the value of the domain expertise that subject librarians bring to the selection and appraisal of data sets for IR inclusion,³¹ while discussion in Bracke further illustrated the application of domain knowledge to support data curation and metadata development.³²

All of these activities are relevant to extending IR service models, as subject librarians are well-positioned to know which DRs their faculty utilize and the long-term preservation capabilities and funding prospects of those repositories. This awareness is essential to identifying published data sets that may benefit from mirroring within the IR, as well as identifying “value add” services that the IR can provide, like supplementary documentation, citation linking, or other services. Similarly, because the DR mirroring service is oriented toward the batch curation and archiving of collections rather than toward individual datasets, the expertise that subject librarians bring to smaller-scale appraisals may more broadly carry over to assessing the long-term value of DR collections based on uniqueness or impact.

A final area of interest with regard to aligning existing IR capabilities with the proposed service relates to technical infrastructure. As noted above, repository solutions with wide adoption among libraries are strongly oriented toward traditional scholarly document types such as preprints and ETDs, with out-of-the-box support for a limited metadata profile based on the simple or qualified Dublin Core schemas.³³ Nonetheless, as reported by Carlson and colleagues and Johnston,³⁴ workflows have been developed that support data curation and publication within common IR platforms including Digital Commons and DSpace.³⁵

In many cases, a lack of data-ready features within IRs can result in a flattening of complex metadata and a format-agnostic presentation of data formats and file types. Even so, expressed priorities and concerns of researchers demonstrate that the publication, permanent identification, and preservation features common among IR platforms can contribute to their adoption as data repositories. For example, Cragin and colleagues and McLure and colleagues described the differing perceptions of researchers regarding concerns and expectations for sharing data and the corresponding service implications for repository builders.³⁶ Limitations aside, important service capabilities as identified by Cragin and colleagues are well-supported by IR platforms, including embargoes and specification of use requirements with preferred citations.³⁷ McLure and colleagues likewise documented researcher views on the potential benefit of IRs as locally managed dissemination and preservation platforms.³⁸ By identifying service requirements of researchers that map to the general purpose, discipline-agnostic nature of IRs, such findings suggest a selective use of IRs to mirror DR collections is a valid use case in alignment with researcher priorities. Taken together with the conceptual

rationale provided above, establishing IR mirrors of DR collections can be of particular benefit when the partnering DR or its data providers lose funding. Additionally, when storage limitations or competing priorities require DRs to concentrate resources around high-use data, mirroring or transferring less in-demand data to an IR offers a means to maintain access through a distribution of management and stewardship duties.

Applications: Requirements and Example Use Cases

Based on the above discussion a case can be made that IRs are suitable platforms to serve as mirrors of DR published data collections. That said, it's important to reiterate that a mirroring service is likely to be practical only if implemented through batch workflows, the development of which will be dependent upon differing DR architectures. However, for the purpose of defining an extensible process model, the scenarios and strategies below are organized into three broadly defined phases: defining stakeholder interactions and requirements, harvesting and metadata processing, and content curation and packaging.

Defining Stakeholder Interactions and Requirements

The first phase of a mirroring service to reflect the contexts of a DR in your IR involves defining the stakeholder interactions and baseline requirements for harvest and ingest procedures. Among other things, the IR or project manager must determine how to satisfy the use, access, and attribution requirements of stakeholders representing the source DR. Minimally, this involves securing permission to harvest and republish the data, either formally via a submission agreement or informally through e-mail or verbal agreement. Additionally, details about which data to transfer along with a proposed schedule should be documented with the necessary authorizations. This documentation is similar to using a submission agreement.

If the data to be mirrored are not subject to restrictions that would prevent mirroring, such formal agreements may not be necessary. However, IR managers should be sensitive to the potential for confusion among researchers who originally contributed their data to the DR. While communicating about the project directly with the researchers or contributors may not be feasible or practical, regular communication with key DR stakeholders about the project

time line and milestones can help prevent misunderstandings. For example, following a collection ingest, IR managers may want to promote the mirroring project via a press release or mass e-mail to their campus community. Such communications should be timed so that researchers who contributed data to the DR are well-informed before any broader announcements are made to potential users.

Regarding use and access permissions, DRs may explicitly include permission information within the corresponding item-level metadata, or else the IR must work to translate this information from implicit repository or collection-level policies. For example, in 2015 the University of New Mexico (UNM) Libraries collaborated with the Sevilleta Long Term Ecological Research (LTER) program to archive and mirror data sets previously published in the LTER Network Data Portal.³⁹ Establishing the authority of the libraries to republish the Sevilleta LTER data via the IR was a multistep process of exploring different strategies for incorporating the LTER data policy. Ultimately, boilerplate language was included as rights metadata within item records with a reference to the full policy online.⁴⁰ Preferred citations referencing the original LTER version of the data were also copied into item records within the IR. Throughout the process, librarians consulted with LTER stakeholders and developed test collections to model different ways of presenting the information.

Another example is provided by Geographic Storage and Retrieval Engine (GSToRE), maintained by the Earth Data Analysis Center (EDAC) of UNM.⁴¹ The GSToRE data are collected from a variety of sources, and there are no overarching access or use policies. Item-level permissions vary, and many data sets are public domain with no access or use constraints, though a boilerplate liability disclaimer inserted by EDAC encourages the citation of data sources.⁴² However, because the preservation model in GSToRE is centered on exporting archive-ready packages to external systems, such as IRs, by implication mirroring collections is an anticipated and generally approved use. Importantly, prior to a harvest and ingest, IR managers or data librarians may refer to GSToRE documentation of service-level and other agreements provided to data depositors. As above, this information can be used to develop boilerplate statements for inclusion within data set metadata for any items mirrored within an IR.

Once stakeholder roles and any conditions for access and use have been addressed, the harvest and ingest process can be further broken down into defining and fulfilling requirements around metadata and content modeling. These requirements will often amount to technical compromises negotiated between the IR and DR. Therefore, it is useful to have access to a development server for prototyping. Testing the ingest procedures within a development environment will additionally allow IR managers to assess what, if any, impact a batch data set ingest may have on IR storage capacity and performance.

Harvesting and Metadata Processing

Common scenarios for metadata harvest include automated retrieval via an API or more manual processes using a web crawler such as Wget.⁴³ Between the two, APIs are the preferred means of access where available; DRs may publish custom APIs or make use of standard APIs including the Simple Web-service Offering Repository Deposit (SWORD) protocol.⁴⁴ Many repository architectures likewise support the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), a metadata-specific API that facilitates discovery and federation.⁴⁵


For example, a popular turnkey data repository application that makes use of SWORD as well as OAI-PMH is the Dataverse Network (DVN). While the provision and maintenance of a Dataverse may not be within a given institution's capabilities, the system's growing adoption⁴⁶ together with its open and interoperable design may result in an increasing use of existing Dataverse Networks by faculty and researchers external to the hosting institution. In such cases, a potential service model for IR managers would be the aggregation of researchers' externally published data. As a means of previewing or analyzing corresponding metadata ahead of transferring data sets from a Dataverse to an IR, the DVN OAI-PMH interface can be accessed via a web browser or scripted using Wget, cURL, or other HTTP interfaces.⁴⁷ In cases where it's preferable to mirror just the metadata and maintain the external DVN as the canonical source of data files, DSpace and other IR applications include OAI-PMH utilities that allow metadata from external repositories to be harvested and published in minutes.

Harvesting metadata via custom APIs will likely be more complex, as IR managers or data librarians will need to develop the necessary software or scripts. For example a set of Python scripts have been developed for a planned harvest of GSToRE data for archiving in UNM's IR that will access canonical DR metadata via JSON and XML through the specifically developed repository and metadata access API functions published by GSToRE.⁴⁸

Once the metadata has been harvested, then mapping or cross-walking activities to align the DR-provided metadata with the IR metadata schema can take place. Some schemas will be easy to map and will include descriptors that are synonymous with the IR metadata such as the simple Dublin Core elements "title," "description," and "publisher." However, even when fields from the source schema map to identically named fields in the destination schema, some analysis is necessary to determine if the fields are used in the same way. Especially when the source metadata schema includes a deeply nested hierarchy, IR managers will need to determine how best to represent multiple source fields that map to a single destination field. For example, the Data Documentation Initiative (DDI) Codebook standard defines unique fields for topic classes, keywords, study concepts, and coverage.⁴⁹ Many of these might

be cross-walked to the Dublin Core “subject” field, but concatenation of multiple source metadata fields could be noisy and negatively impact web displays or usability. Decisions about metadata mapping will be decided by IR capabilities and the preferences of the DR stakeholders. In some cases, documentation of best practices and recommended cross-walks will be available. Specific to the example given here, a DDI-to-Dublin Core cross-walk is provided by the DDI Alliance.⁵⁰

If the IR capabilities include metadata extension or customization, another strategy is to map the source metadata to an alternative schema or use multiple schemas. For example, DSpace versions 1.5 and above support the registration of multiple “flat” schemas, which enable IR managers to combine fields from different schemas when describing items.⁵¹ In practice, while a complex schema such as the Ecological Metadata Language (EML) cannot be fully cross-walked to the DSpace Dublin Core profile,⁵² the standard can be mapped to Darwin Core in a semantically meaningful way.⁵³ Without utilizing a nested hierarchy of “coverage” fields as in EML, a single qualified term set in Darwin Core nonetheless includes categories of domain-specific terms such as “GeologicalContext” and “Taxon.” In support of mirroring data from repositories that use EML metadata, extending the DSpace metadata registry to implement Darwin Core is a simple process of registering the namespace URI and adding desired fields (figure 7.1).

Jonathan Wheeler ▾

🏠 LoboVault Home / Metadata registry

UNM LIBRARIES

University Libraries

Law Library

Health Sciences Library

Search 🔍

BROWSE

All of LoboVault

Communities & Collections

Date

Authors

Titles

Subjects

MY ACCOUNT

My Exports

Logout

Profile

Submissions

ADMINISTRATIVE

Metadata registry

The metadata registry maintains a list of all metadata fields available in the repository. These fields may be divided amongst multiple schemas. However, LoboVault requires the qualified Dublin Core schema. You may extend the Dublin Core schema with additional fields or add new schemas to the registry.

ID	Namespace	Name	
1	http://dublincore.org/documents/dcmi-terms/	dc	
<input type="checkbox"/>	2	http://elibrary.unm.edu/embargo-terms/	emb
<input type="checkbox"/>	4	http://www.ndtd.org/standards/metadata/eldms/1.0/xml.xsd	eldms
<input type="checkbox"/>	5	http://elibrary.unm.edu/data	data
<input type="checkbox"/>	6	http://purl.org/dc/terms/	dcterms
<input type="checkbox"/>	7	http://dspace.org/eperson	eperson
<input type="checkbox"/>	8	http://rs.tdwg.org/dwc/terms/index.htm	dwc

☐ Delete schema

Add a new schema

Namespace: *

Namespace should be an established URI location for the new schema.

Name: *

FIGURE 7.1
The DSpace administrator’s view of the metadata registry. The Darwin Core namespace is highlighted.

The benefits of this approach are demonstrated by a map visualization feature within UNM's IR that was developed in support of the Sevilleta LTER data-archiving project. Because of the important geographical context of the data, it was desirable to reproduce the maps drawn by the network portal for items with coordinate metadata. Using the qualified Dublin Core "spatial coverage" element was impractical because existing items already used that field to provide place names, and mixing coordinate and text data types would have broken the JavaScript/XSL mapping template developed for UNM's DSpace instance. By extending the metadata registry to include the Darwin Core "decimalLatitude" and "decimalLongitude" elements, librarians were able to enforce a coordinate data constraint within the mapping template (figure 7.2).

UNM LIBRARIES

- University Libraries
- Law Library
- Health Sciences Library

Search

BROWSE

- All of LoboVault
- Communities & Collections
- Date
- Authors
- Titles
- Subjects

MY ACCOUNT

- My Exports
- Logout
- Profile
- Submissions

ADMINISTRATIVE

- Control Panel
- Statistics
- Curation Tasks
- Access Control
- People
- Groups
- Authorizations
- Content Administration

Metadata Schema: "dwc"

This is the metadata schema for "http://rs.tdwg.org/dwc/terms/index.html". You may add new or update existing metadata fields to this schema. Fields may also be selected for deletion or be moved to another schema.

Add new metadata field

Field Name:

Scope Note:

Additional notes about this metadata field.

Schema metadata fields

ID	Field	Scope Note
<input type="checkbox"/> 160	dwc:decimalLatitude	The geographic latitude (in decimal degrees, using the spatial reference system given in geodeticDatum) of the geographic center of a Location. Positive values are north of the Equator, negative values are south of it. Legal values lie between -90 and 90, inclusive.
<input type="checkbox"/> 161	dwc:decimalLongitude	The geographic longitude (in decimal degrees, using the spatial reference system given in geodeticDatum) of the geographic center of a Location. Positive values are east of the Greenwich Meridian, negative values are west of it. Legal values lie between -180 and 180, inclusive.

FIGURE 7.2
Adding fields for metadata schema.

Once decisions about representing DR metadata within the IR have been made, the harvested metadata content must be cross-walked to the IR schema and saved in a file format accepted by the IR for batch ingest. Typically, this pro-

cess will be accomplished using XSL templates to transform XML metadata, but other options may exist. DSpace, for example, allows batch creation and editing of metadata via CSV upload through the web interface.⁵⁴

Content Curation and Packaging

Finally, together with its metadata schema, the destination IR will have specific requirements for associating content files with their respective metadata and packaging items for ingest. As with metadata, content files can be harvested through a variety of means, preferably via API but alternatively through batch HTTP requests via cURL or Wget. Whichever method is used, an important pre-harvest activity is to create an inventory of the DR assets to be acquired. This information, which may be published as site statistics or requested from DR administrators, minimally provides a quick overview of item and version counts that can be used later to verify the completeness of the harvest. In addition to an inventory, librarians managing a harvest must also identify the file validation scheme used within the DR. For example, checksums will often be made accessible via API and should be used to validate harvested files.

Wherever possible, IR managers and librarians should seek to curate the data for preservation and explore options for otherwise adding value to the data and metadata. Minimally, curation will involve documenting and exposing provenance information relevant to the mirroring process, such as the date of harvest and the outcomes of virus scans, file validation, and format identification routines. These processes may be readily incorporated into a collection-scale workflow through the use of existing batch utilities like the Digital Record Object Identification (DROID) software tool.⁵⁵

Further curation actions may include compiling any additional documentation necessary to support data discovery and use within the IR context. For example, an early and relatively small batch data ingest into the UNM IR involved mirroring a set of colonia population data published by the Bureau of Business and Economic Research (BBER).⁵⁶ In communication with the lead researcher, the content files were harvested from the BBER website using Wget, and the metadata and supporting documentation were compiled through discussion and by collating any corresponding presentations, reports, and so on. Additional curation activities performed on the data set included transforming files from proprietary formats to open formats and creating provenance and technical metadata using DROID and a locally developed METS utility.⁵⁷ These and other value-add activities resulted in the publication of an IR mirror of the BBER data set that was more than just a duplication of the original resource.⁵⁸ Also, the simple but scalable batch workflow was a prototype of the procedures used to curate and package the Sevilleta LTER data.

For the final ingest into the IR, item- and collection-level packaging requirements will be platform-dependent. Consequently, the role or involvement of the repository manager will vary according to whether the IR is hosted by a third party, locally maintained, or open source. While the IR manager's participation in batch ingest routines within proprietary systems may be limited, the necessary features should exist, and vendors are often interested in exploring innovative uses of their systems. As an example, Carlson and colleagues reported on a project in which materials from a large research center were curated within a bepress Digital Commons repository at Purdue.⁵⁹

Alternatively, managers of locally hosted, open-source platforms such as DSpace may capitalize on available documentation and utilities developed by the user community. Specifically, DSpace supports batch ingest of items packaged according to a Simple Archive Format (SAF) specification.⁶⁰ Similar to the Bagit digital content transfer utility developed by the Library of Congress,⁶¹ SAF describes a per-item file structure and automated ingest process for DSpace repositories. The available documentation is comprehensive, but in summary a collection packaged for ingest using SAF will consist of a directory or zip archive containing individual, item-level subdirectories. The subdirectories will contain the item's associated content files, one or more XML metadata files, a text file manifest describing the content file types, and, optionally, a text file designating the collection or collections to which the item belongs. Ingest is completed by submitting SAF packages to the repository via a command line utility or, alternatively, using a web-based batch import feature introduced in DSpace version 5.⁶²

Following ingest, some post-processing for quality assurance purposes is recommended. In addition to verifying that the process concluded without errors, quality checks can range in scope and depth and can be implemented through various manual or automated processes. For example, following ingest of the BBER colonia data, the relatively small size of the data set enabled librarians to perform manual quality checks. These checks included downloading the individual files to identify file formats and validate checksums using a second run through DROID. In the case of the LTER data ingest, a percentage of the collection was manually checked for format and checksum validation, but automated processes were run against the full collection using available DSpace curation tools. These tools include file format identification and checksum validation processes that may be run on demand against an item, collection, or community. None of the quality checks performed on either the BBER or LTER collections identified any errors. However, because batch processing can result in the propagation of errors across an entire collection, such follow-up checks are an important element of a harvest and ingest workflow.

Conclusion

As researcher and institutional data management needs evolve to encompass federal public access planning and DMP compliance requirements, the demand for library data management services may be expected to grow accordingly. In addition to well-established activities such as DMP consultation and data reference, technical support for asset management and data preservation represent additional niche services that academic libraries are well-situated to provide based on existing professional skill sets, established IR infrastructures, and corresponding digital preservation workflows. While near-term sharing and timely publication of data via DRs remains a researcher-preferred strategy, the migration or mirroring of previously published data within IRs may provide capabilities in support of archiving and reuse that are complementary or supplementary to DR publication features. Although such mirroring represents a promising service model for libraries, the potential for incorporating a routine collection-scale ingest activity requires the corresponding development of batch harvest, packaging, and ingest processes.

While acknowledging that the workflows presented here are desktop-based and thus do not fully address scalability issues, there are some advantages to maintaining desktop workflows, such as quality control. Further, the curation and packaging of collections is similar to curating individual data sets in that it is a high-touch activity and requirements will vary from case to case. Because of this, the need to customize processes will inevitably impact scalability. However, bandwidth issues and storage constraints will present themselves, and a future development of flexible utilities for data transfer between DRs and IRs is needed. In particular, as initiatives such as the Digital Preservation Network (DPN) grow,⁶³ a near-term focus for IR managers should be the development of processes that automatically generate archival information packages for DR data on harvest. Because not all IRs are maintained as archival or preservation platforms, such a feature would enable a parallel transfer of DR data collections to alternative preservation services such as DPN and DuraCloud.⁶⁴ Through development of these and other services to better position IRs within the web of repositories, the collective contribution of libraries to data preservation will further demonstrate their value as memory institutions and partners within a global data infrastructure.

Acknowledgments

The author would like to thank and acknowledge the following for their feedback and contributions during the Sevilleta LTER data ingest into UNM's institutional repository: Kristin Vanderbilt (Sevilleta LTER Program), Mark Servilla (LTER Network Office), Jacob Nash (UNM Health Sciences Library and Informatics Center), and UNM Libraries Information Technology Services.

The Python and XSLT scripts used to harvest and package the Sevilleta LTER data for ingest into a DSpace repository are available at <https://lobogit.unm.edu/jwheel01/lter-collection-harvest>.

Notes

1. Karen Antell, Jody Bales Foote, Jaymie Turner, and Brian Shults, "Dealing with Data: Science Librarians' Participation in Data Management at Association of Research Libraries Institutions," *College and Research Libraries* 75, no. 4 (July 2014): 557–74, doi:10.5860/crl.75.4.557. Regarding planning, implementation, and perceived value of RDS services among ACRL libraries, an interesting corollary discussion of how perceptions align between library administrators and librarians is provided in Carol Tenopir, Robert J. Sandusky, Suzie Allard, and Ben Birch, "Research Data Management Services in Academic Research Libraries and Perceptions of Librarians," *Library and Information Science Research* 36, no. 2 (April 2014): 84–90, doi:10.1016/j.lisr.2013.11.003.
2. Antell et al., "Dealing with Data"; Hans Jørn Nielsen and Birger Hjørland, "Curating Research Data: The Potential Roles of Libraries and Information Professionals," *Journal of Documentation* 70, no. 2 (2014): 221–240, doi:10.1108/JD-03-2013-0034.
3. Don MacMillan, "Data Sharing and Discovery: What Librarians Need to Know," *Journal of Academic Librarianship* 40, no. 5 (September 2014): 546, doi:10.1016/j.acalib.2014.06.011.
4. Nancy Y. McGovern and Aprille C. McKay, "Leveraging Short-Term Opportunities to Address Long-Term Obligations: A Perspective on Institutional Repositories and Digital Preservation Programs," *Library Trends* 57, no. 2 (2008): 262–79, <https://muse.jhu.edu/article/262030>; Priti Jain, "New Trends and Future Applications/Directions of Institutional Repositories in Academic Institutions," *Library Review* 60, no. 2 (March 2011): 125–41, doi:10.1108/00242531111113078.
5. Karen S. Baker and Lynn Yarmey, "Data Stewardship: Environmental Data Curation and a Web-of-Repositories," *International Journal of Digital Curation* 4, no. 2 (October 15, 2009): 12–27, doi:10.2218/ijdc.v4i2.90; Karen S. Baker and Florence Millerand, "Infrastructuring Ecology: Challenges in Achieving Data Sharing," in *Collaboration in the New Life Sciences*, ed. John N. Parker, Niki Vermeulen, and Bart Penders (Burlington, VT: Ashgate, 2010), 111–38.
6. Tyler Walters, "Assimilating Digital Repositories into the Active Research Process," in *Research Data Management: Practical Strategies for Information Professionals*, ed. Joyce M. Ray (West Lafayette, IN: Purdue University Press, 2014), eBook Collection, EBSCO-host, ISBN 9781461956815. Accessed February 19, 2016.
7. Greg Janée, Justin Mathena, and James Frew, "A Data Model and Architecture for Long-Term Preservation," in *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries* (New York: ACM, 2008), 134–44. doi:10.1145/1378889.1378912.
8. National Science Foundation, Grant Proposal Guide, Chapter II, Proposal Preparation Instructions, Section C.2.j, last modified January 25, 2016, accessed March 23, 2016, http://www.nsf.gov/pubs/policydocs/pappguide/nsf16001/gpg_2.jsp#IIC2j. The DMP requirement described within the Grant Proposal Guide (GPG) includes separate recommendations covering "policies for access and sharing" and "plans for archiving data, samples, and other research products, and for preservation of access to them"; US

- Department of Energy, Office of Science “Statement on Digital Data Management,” last modified July 28, 2014, accessed March 22, 2016, <http://science.energy.gov/funding-opportunities/digital-data-management>.
9. G. Sayeed Choudhury, “Case Study 1: Johns Hopkins University Data Management Services,” in *Delivering Research Data Management Services*, ed. Graham Pryor, Sarah Jones, and Angus Whyte (London: Facet Publishing, 2014), 118.
 10. Paul F. Uhlig, “Information Gulags, Intellectual Straightjackets, and Memory Holes: Three Principles to Guide the Preservation of Scientific Data,” *Data Science Journal* 9 (2010): ES5. https://www.jstage.jst.go.jp/article/dsj/9/0/9_Essay-001-Uhlig/_article.
 11. Key Perspectives Ltd., *Data Dimensions: Disciplinary Differences in Research Data Sharing, Reuse and Long Term Viability. SCARP Synthesis Study* (Digital Curation Centre, 2010), <http://hdl.handle.net/1842/3364>.
 12. Donatella Castelli, Paolo Manghi, and Costantino Thanos, “A Vision towards Scientific Communication Infrastructures: On Bridging the Realms of Research Digital Libraries and Scientific Data Centers,” *International Journal on Digital Libraries* 13, no. 3–4 (September 2013): 155–69, doi:10.1007/s00799-013-0106-7.
 13. *Ibid.*, 162.
 14. *Ibid.*, 162–163.
 15. Key Perspectives, *Data Dimensions*.
 16. Michael Witt, “Co-designing, Co-developing, and Co-implementing an Institutional Data Repository Service,” *Journal of Library Administration* 52, no. 2 (2012): 172–88; Purdue University Research Repository, accessed March 23, 2016, <https://purr.purdue.edu/>.
 17. Witt, “Co-designing,” 176.
 18. William C. Block, Eric Chen, Jim Cordes, Dianne Dietrich, Dean B Krafft, Stefan Kramer, David Lifka, Janet McCue, and Gail Steinhart, *Meeting Funders’ Data Policies: Blueprint for a Research Data Management Service Group (RDMSG)*, project report (Ithaca, NY: Cornell University, 2010), <http://hdl.handle.net/1813/28570>.
 19. Carol Tenopir, Ben Birch, and Suzie Allard, *Academic Libraries and Research Data Services*, an ACRL white paper (Chicago: Association of College and Research Libraries, 2012), 37–39.
 20. Choudhury, “Case Study 1,” 128.
 21. Johns Hopkins Data Archive Dataverse Network, accessed March 23, 2016, <https://archive.data.jhu.edu/dvn/>.
 22. Tenopir et al., “Research Data Management,” 87, provides an example of a distinction between informational (or consulting) RDS and technical RDS.
 23. Janice Jaguszewski and Karen Williams, *New Roles for New Times*, report (Washington, DC: Association of Research Libraries, August 2013), 13, <http://hdl.handle.net/11299/169867>.
 24. Mark P. Newton, C. C. Miller, and Marianne Stowell Bracke, “Librarian Roles in Institutional Repository Data Set Collecting: Outcomes of a Research Library Task Force,” *Collection Management* 36, no. 1 (2010): 53–67, doi:10.1080/01462679.2011.530546.
 25. Tenopir et al., “Research Data Management,” 87; Ashley E. Sands, Christine L. Borgman, Sharon Traweck, and Laura A. Wynholds, “We’re Working on It: Transferring the Sloan Digital Sky Survey from Laboratory to Library,” *International Journal of Digital Curation* 9, no. 2 (October 30, 2014), doi:10.2218/ijdc.v9i2.336; Liz Lyon, “The Informatics Transform: Re-engineering Libraries for the Data Decade,” *International Journal of Digital Curation* 7, no. 1 (March 12, 2012): 131–32, doi:10.2218/ijdc.v7i1.220.

26. Jeonghyun Kim, "Data Sharing and Its Implications for Academic Libraries," *New Library World* 114, no. 11/12 (November 18, 2013): 503, doi:10.1108/NLW-06-2013-0051.
27. Jared Lyle, George Alter, and Ann Green, "Partnering to Curate and Archive Social Science Data," in *Research Data Management: Practical Strategies for Information Professionals*, ed. Joyce M. Ray (West Lafayette, IN: Purdue University Press, 2014) eBook Collection, EBSCOhost, ISBN 9781461956815 accessed February 19, 2016; Inter-university Consortium for Political and Social Research homepage, accessed March 23, 2016, <https://www.icpsr.umich.edu/icpsrweb/landing.jsp>.
28. Lyle, Alter, and Green, "Partnering to Curate and Archive Social Science Data," under the heading "Need for Support."
29. Ibid., under the heading "Find."
30. Antell et al., "Dealing with Data," 567.
31. Newton, Miller, and Bracke, "Librarian Roles," 58, 61.
32. Marianne Stowell Bracke, "Emerging Data Curation Roles for Librarians: A Case Study of Agricultural Data," *Journal of Agricultural and Food Information* 12, no. 1 (2011): 65–74, doi:10.1080/10496505.2011.539158.
33. For example information about DSpace's default metadata schema see DSpace, "Functional Overview," DSpace 5.x Documentation, accessed March 23, 2016, <https://wiki.duraspace.org/display/DSDOC5x/Functional+Overview#FunctionalOverview-MetadataManagement>. Similar information for bepress's Digital Commons is available at bepress, "Metadata Options in Digital Commons," Digital Commons Reference Material and User Guides, last modified January 2016, accessed March 23, 2016, <http://digital-commons.bepress.com/cgi/viewcontent.cgi?article=1095&context=reference>.
34. Jake Carlson, Alexis E. Ramsey, and J. David Kotterman, "Using an Institutional Repository to Address Local-scale Needs: A Case Study at Purdue University," *Library Hi Tech* 28, no. 1 (March 9, 2010): 152–73, doi:10.1108/07378831011026751; Lisa R. Johnston, *A Workflow Model for Curating Research Data in the University of Minnesota Libraries: Report from the 2013 Data Curation Pilot* (University of Minnesota Digital Conservancy, 2014), <http://hdl.handle.net/11299/162338>.
35. Digital Commons homepage, accessed March 23, 2016, <http://digitalcommons.bepress.com/>; DSpace homepage, accessed March 23, 2016, <http://dspace.org/>.
36. Melissa H. Cragin, Carole L. Palmer, Jacob R. Carlson, and Michael Witt, "Data Sharing, Small Science and Institutional Repositories," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 368, no. 1926 (September 13, 2010): 4023–38, doi:10.1098/rsta.2010.0165; Merinda McLure, Allison V. Level, Catherine L. Cranston, Beth Oehlerts, and Mike Culbertson, "Data Curation: A Study of Researcher Practices and Needs," *portal: Libraries and the Academy* 14, no. 2 (2014): 139–64, doi:10.1353/pla.2014.0009.
37. Cragin et al., "Data Sharing," 4035–36.
38. McLure et al., "Data Curation," 154.
39. Sevilleta Long Term Ecological Research Program, accessed March 23, 2016, <http://sev.lternet.edu/>; LTER Network Data Portal, accessed March 23, 2016, <https://portal.lternet.edu/nis/home.jsp>.
40. Long Term Ecological Research Network, "LTER Network Data Access Policy, Data Access Requirements, and General Data Use Agreement," accessed March 23, 2016, <http://www.lternet.edu/policies/data-access>.

41. GSToRE (Geographic Storage, Transformation and Retrieval Engine), version 3, homepage, accessed March 23, 2016, <https://gstore.unm.edu/>; EDAC (Earth Data Analysis Center) homepage, accessed March 23, 2016, <http://edac.unm.edu/>.
42. See, for example, the section on “Distribution Liability” in GSToRE “Wildfire Risk Main Model,” accessed March 22, 2016, <http://gstore.unm.edu/apps/rgis/data-sets/71be383b-ad19-4252-9c01-cfad3216a0ca/metadata/FGDC-STD-001-1998.html>.
43. “GNU Wget 1.18 Manual,” last modified December 11, 2015, accessed March 23, 2016, <https://www.gnu.org/software/wget/>.
44. SWORD homepage, accessed March 23, 2016, <http://swordapp.org/>.
45. Open Archives Initiative, “Protocol for Metadata Harvesting,” accessed March 23, 2016, <https://www.openarchives.org/pmh/>.
46. Dataverse Project, “Dataverse Repositories,” accessed March 22, 2016, <http://dataverse.org/>. This overview on the Dataverse Project website shows fourteen Dataverse repositories worldwide as of March 11, 2016. It should be noted, however, that an individual repository may host Dataverses for other institutions. For example, the Harvard repository includes over 1,400 “sub” Dataverses, many of which are sponsored by external universities and organizations. In Europe, the Utrecht University’s DataverseNL likewise hosts Dataverses sponsored by institutions throughout Central and Eastern Europe.
47. cURL homepage, accessed March 23, 2016, <https://curl.haxx.se/>.
48. python homepage, accessed March 23, 2016, <https://www.python.org/>.
49. Data Documentation Initiative, “DDI-Codebook 2.5,” accessed March 23, 2016, <http://www.ddialliance.org/Specification/DDI-Codebook/2.5/>.
50. Data Documentation Initiative, “Mapping to Dublin Core (DDI Version 2),” accessed March 23, 2016, <http://www.ddialliance.org/resources/ddi-profiles/dc>.
51. Available since at least version 1.5, documentation for DSpace version 5 is available at DSpace, “Functional Overview,” DSpace 5.x Documentation, accessed March 23, 2016, <https://wiki.duraspace.org/display/DSDOC5x/Functional+Overview#FunctionalOverview-MetadataManagement>.
52. The Knowledge Network for Biocomplexity, accessed March 24, 2016, <https://knb.ecoinformatics.org/>.
53. Darwin Core Task Group, “Darwin Core,” issued February 12, 2009, last updated June 5, 2015, accessed March 23, 2016, <http://rs.tdwg.org/dwc/>.
54. Available since DSpace version 1.6, documentation for the current version is available at DSpace, “Batch Metadata Editing,” DSpace 5.x Documentation, accessed March 23, 2016, <https://wiki.duraspace.org/display/DSDOC5x/Batch+Metadata+Editing>.
55. National Archives, “File Profiling Tool (DROID),” accessed March 23, 2016, <http://www.nationalarchives.gov.uk/information-management/manage-information/policy-process/digital-continuity/file-profiling-tool-droid/>.
56. Bureau of Business and Economic Research homepage, accessed March 23, 2016, <http://bber.unm.edu/>.
57. Library of Congress, “METS: Metadata Encoding and Transmission Standard,” last modified February 9, 2016, accessed March 23, 2016, <http://www.loc.gov/standards/mets/>.
58. Daren Ruiz, “Colonia Population and Socioeconomic and Housing Characteristic Estimates, Maps and Shape File Update: November 2012 [data set],” University of New Mexico (2012), <http://hdl.handle.net/1928/22547>. Additional content and metadata available at <http://repository.unm.edu/archive/Projects/22547/>.

59. Carlson, Ramsey, and Kotterman, "Using an Institutional Repository."
60. DSpace, "Importing and Exporting Items via Simple Archive Format," DSpace 5.x Documentation, accessed March 23, 2016, <https://wiki.duraspace.org/display/DSDOC5x/Importing+and+Exporting+Items+via+Simple+Archive+Format>.
61. Library of Congress, "Bagit: Transferring Content for Digital Preservation," video, 3:14, posted June 24, 2009, accessed March 23, 2016, <http://www.digitalpreservation.gov/multimedia/videos/bagit0609.html>.
62. DSpace, "Latest Release," accessed March 23, 2016, <http://www.dspace.org/latest-release>.
63. Digital Preservation Network homepage, accessed March 23, 2016, <http://dpn.org/>.
64. Duracloud homepage, accessed March 23, 2016, <http://www.duracloud.org/>.

Bibliography

- Antell, Karen, Jody Bales Foote, Jaymie Turner, and Brian Shults. "Dealing with Data: Science Librarians' Participation in Data Management at Association of Research Libraries Institutions." *College and Research Libraries* 75, no. 4 (July 2014): 557–74. doi:10.5860/crl.75.4.557.
- Baker, Karen S., and Florence Millerand. "Infrastructuring Ecology: Challenges in Achieving Data Sharing." In *Collaboration in the New Life Sciences*. Edited by John N. Parker, Niki Vermeulen, and Bart Penders, 111–38. Burlington, VT: Ashgate, 2010.
- Baker, Karen S., and Lynn Yarmey. "Data Stewardship: Environmental Data Curation and a Web-of-Repositories." *International Journal of Digital Curation* 4, no. 2 (October 15, 2009): 12–27. doi:10.2218/ijdc.v4i2.90.
- bepress, "Metadata Options in Digital Commons," Digital Commons Reference Material and User Guides, last modified January 2016, accessed March 23, 2016, <http://digitalcommons.bepress.com/cgi/viewcontent.cgi?article=1095&context=reference>.
- Block, William C., Eric Chen, Jim Cordes, Dianne Dietrich, Dean B. Krafft, Stefan Kramer, David Lifka, Janet McCue, and Gail Steinhart. *Meeting Funders' Data Policies: Blueprint for a Research Data Management Service Group (RDMSG)*. Project report. Ithaca, NY: Cornell University, 2010. <http://hdl.handle.net/1813/28570>.
- Bracke, Marianne Stowell. "Emerging Data Curation Roles for Librarians: A Case Study of Agricultural Data." *Journal of Agricultural and Food Information* 12, no. 1 (2011): 65–74. doi:10.1080/10496505.2011.539158.
- Bureau of Business and Economic Research, accessed March 23, 2016, <http://bber.unm.edu/>.
- Carlson, Jake, Alexis E. Ramsey, and J. David Kotterman. "Using an Institutional Repository to Address Local-Scale Needs: A Case Study at Purdue University." *Library Hi Tech* 28, no. 1 (March 9, 2010): 152–73. doi:10.1108/07378831011026751.
- Castelli, Donatella, Paolo Manghi, and Costantino Thanos. "A Vision towards Scientific Communication Infrastructures: On Bridging the Realms of Research Digital Libraries and Scientific Data Centers." *International Journal on Digital Libraries* 13, no. 3–4 (September 2013): 155–69. doi:10.1007/s00799-013-0106-7.
- Choudhury, G. Sayeed. "Case Study 1: Johns Hopkins University Data Management Services." In *Delivering Research Data Management Services*. Edited by Graham Pryor, Sarah Jones, and Angus Whyte, 115–33. London: Facet Publishing, 2014.

- Cragin, Melissa H., Carole L. Palmer, Jacob R. Carlson, and Michael Witt. "Data Sharing, Small Science and Institutional Repositories." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 368, no. 1926 (September 13, 2010): 4023–38. doi:10.1098/rsta.2010.0165.
- cURL, accessed March 23, 2016, <https://curl.haxx.se/>.
- Darwin Core Task Group, "Darwin Core," issued February 12, 2009, last updated June 5, 2015, accessed March 23, 2016, <http://rs.tdwg.org/dwc/>.
- Data Documentation Initiative, "DDI-Codebook 2.5," accessed March 23, 2016, <http://www.ddialliance.org/Specification/DDI-Codebook/2.5/>.
- Data Documentation Initiative, "Mapping to Dublin Core (DDI Version 2)," accessed March 23, 2016, <http://www.ddialliance.org/resources/ddi-profiles/dc>.
- Dataverse Project, "Dataverse Repositories," accessed March 22, 2016, <http://dataverse.org/>.
- Digital Commons, accessed March 23, 2016, <http://digitalcommons.bepress.com/>.
- Digital Preservation Network, accessed March 23, 2016, <http://dpn.org/>.
- DSpace, accessed March 23, 2016, <http://dspace.org/>.
- DSpace, "Batch Metadata Editing," DSpace 5.x Documentation, accessed March 23, 2016, <https://wiki.duraspace.org/display/DSDOC5x/Batch+Metadata+Editing>.
- DSpace, "Functional Overview," DSpace 5.x Documentation, accessed March 23, 2016, <https://wiki.duraspace.org/display/DSDOC5x/Functional+Overview#FunctionalOverview-MetadataManagement>.
- DSpace, "Importing and Exporting Items via Simple Archive Format," DSpace 5.x Documentation, accessed March 23, 2016, <https://wiki.duraspace.org/display/DSDOC5x/Importing+and+Exporting+Items+via+Simple+Archive+Format>.
- DSpace, "Latest Release," accessed March 23, 2016, <http://www.dspace.org/latest-release>.
- Duracloud, accessed March 23, 2016, <http://www.duracloud.org/>.
- EDAC (Earth Data Analysis Center) homepage, accessed March 23, 2016, <http://edac.unm.edu/>.
- GNU Wget 1.18 Manual," last modified December 11, 2015, accessed March 23, 2016, <https://www.gnu.org/software/wget/>.
- GSToRE (Geographic Storage, Transformation and Retrieval Engine), version 3, homepage, accessed March 23, 2016, <https://gstore.unm.edu/>.
- GSToRE "Wildfire Risk Main Model," accessed March 22, 2016, <http://gstore.unm.edu/apps/rgis/datasets/71be383b-ad19-4252-9c01-cfad3216a0ca/metadata/FG-DC-STD-001-1998.html>.
- Inter-university Consortium for Political and Social Research, accessed March 23, 2016, <https://www.icpsr.umich.edu/icpsrweb/landing.jsp>.
- Jaguszewski, Janice, and Karen Williams. *New Roles for New Times: Transforming Liaison Roles in Research Libraries*. Report. Washington, DC: Association of Research Libraries, August 2013. <http://hdl.handle.net/11299/169867>.
- Jain, Priti. "New Trends and Future Applications/Directions of Institutional Repositories in Academic Institutions." *Library Review* 60, no. 2 (March 2011): 125–41. doi:10.1108/00242531111113078.
- Janée, Greg, Justin Mathena, and James Frew. "A Data Model and Architecture for Long-Term Preservation." In *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries*, 134–44. New York: ACM, 2008. doi:10.1145/1378889.1378912.
- Johns Hopkins Data Archive Dataverse Network, accessed March 23, 2016, <https://archive.data.jhu.edu/dvn/>.

- Johnston, Lisa R. *A Workflow Model for Curating Research Data in the University of Minnesota Libraries: Report from the 2013 Data Curation Pilot*. University of Minnesota Digital Conservancy, 2014. <http://hdl.handle.net/11299/162338>.
- Key Perspectives Ltd. *Data Dimensions: Disciplinary Differences in Research Data Sharing, Reuse and Long Term Viability. SCARP Synthesis Study*. Digital Curation Centre, 2010. <http://hdl.handle.net/1842/3364>.
- Kim, Jeonghyun. "Data Sharing and Its Implications for Academic Libraries." *New Library World* 114, no. 11/12 (November 18, 2013): 494–506. doi:10.1108/NLW-06-2013-0051.
- Knowledge Network for Biocomplexity, accessed March 24, 2016, <https://knb.ecoinformatics.org/>.
- Library of Congress, "Bagit: Transferring Content for Digital Preservation," video, 3:14, posted June 24, 2009, accessed March 23, 2016, <http://www.digitalpreservation.gov/multimedia/videos/bagit0609.html>.
- Library of Congress, "METS: Metadata Encoding and Transmission Standard," last modified February 9, 2016, accessed March 23, 2016, <http://www.loc.gov/standards/mets/>.
- Long Term Ecological Research Network, "LTER Network Data Access Policy, Data Access Requirements, and General Data Use Agreement," accessed March 23, 2016, <http://www.lternet.edu/policies/data-access>.
- LTER Network Data Portal, accessed March 23, 2016, <https://portal.lternet.edu/nis/home.jsp>.
- Lyle, Jared, George Alter, and Ann Green. "Partnering to Curate and Archive Social Science Data." in *Research Data Management: Practical Strategies for Information Professionals*. Edited by Joyce M. Ray. West Lafayette, IN: Purdue University Press, 2014. eBook Collection, EBSCOhost, ISBN 9781461956815. Accessed February 19, 2016.
- Lyon, Liz. "The Informatics Transform: Re-engineering Libraries for the Data Decade." *International Journal of Digital Curation* 7, no. 1 (March 12, 2012): 126–38. doi:10.2218/ijdc.v7i1.220.
- MacMillan, Don. "Data Sharing and Discovery: What Librarians Need to Know." *Journal of Academic Librarianship* 40, no. 5 (September 2014): 541–49. doi:10.1016/j.acalib.2014.06.011.
- McGovern, Nancy Y., and Aprille C. McKay. "Leveraging Short-Term Opportunities to Address Long-Term Obligations: A Perspective on Institutional Repositories and Digital Preservation Programs." *Library Trends* 57, no. 2 (2008): 262–79. <https://muse.jhu.edu/article/262030>.
- McLure, Merinda, Allison V. Level, Catherine L. Cranston, Beth Oehlerts, and Mike Culbertson. "Data Curation: A Study of Researcher Practices and Needs." *portal: Libraries and the Academy* 14, no. 2 (2014): 139–64. doi:10.1353/pla.2014.0009.
- National Archives, "File Profiling Tool (DROID)," accessed March 23, 2016, <http://www.nationalarchives.gov.uk/information-management/manage-information/policy-process/digital-continuity/file-profiling-tool-droid/>.
- National Science Foundation, Grant Proposal Guide, Chapter II, Proposal Preparation Instructions, Section C.2.j, last modified January 25, 2016, accessed March 23, 2016, http://www.nsf.gov/pubs/policydocs/pappguide/nsf16001/gpg_2.jsp#IIC2j.
- Newton, Mark P., C. C. Miller, and Marianne Stowell Bracke. "Librarian Roles in Institutional Repository Data Set Collecting: Outcomes of a Research Library Task Force." *Collection Management* 36, no. 1 (2010): 53–67. doi:10.1080/01462679.2011.530546.

- Nielsen, Hans Jørn, and Birger Hjørland, "Curating Research Data: The Potential Roles of Libraries and Information Professionals," *Journal of Documentation* 70, no. 2 (2014): 221–40, doi:10.1108/JD-03-2013-0034.
- Open Archives Initiative, "Protocol for Metadata Harvesting," accessed March 23, 2016, <https://www.openarchives.org/pmh/>.
- Purdue University Research Repository, accessed March 23, 2016, <https://purr.purdue.edu/python>, accessed March 23, 2016, <https://www.python.org/>.
- Ruiz, Daren. "Colonia Population and Socioeconomic and Housing Characteristic Estimates, Maps and Shape File Update: November 2012 [data set]." University of New Mexico (2012). <http://hdl.handle.net/1928/22547>.
- Sands, Ashley E., Christine L. Borgman, Sharon Traweek, and Laura A. Wynholds. "We're Working on It: Transferring the Sloan Digital Sky Survey from Laboratory to Library." *International Journal of Digital Curation* 9, no. 2 (October 30, 2014). doi:10.2218/ijdc.v9i2.336.
- Sevilleta Long Term Ecological Research Program, accessed March 23, 2016, <http://sevlternet.edu/>.
- SWORD, accessed March 23, 2016, <http://swordapp.org/>.
- Tenopir, Carol, Ben Birch, and Suzie Allard. *Academic Libraries and Research Data Services: Current Practices and Plans for the Future*. An ACRL white paper. Chicago: Association of College and Research Libraries, 2012.
- Tenopir, Carol, Robert J. Sandusky, Suzie Allard, and Ben Birch. "Research Data Management Services in Academic Research Libraries and Perceptions of Librarians." *Library and Information Science Research* 36, no. 2 (April 2014): 84–90. doi:10.1016/j.lisr.2013.11.003.
- Uhlir, Paul F. "Information Gulags, Intellectual Straightjackets, and Memory Holes: Three Principles to Guide the Preservation of Scientific Data" *Data Science Journal* 9 (2010): ES1–5. https://www.jstage.jst.go.jp/article/dsj/9/0/9_Essay-001-Uhlir/_article.US
- Department of Energy, Office of Science. "Statement on Digital Data Management." Last modified July 28, 2014, accessed March 22, 2016. <http://science.energy.gov/funding-opportunities/digital-data-management>.
- Walters, Tyler. "Assimilating Digital Repositories into the Active Research Process." In *Research Data Management: Practical Strategies for Information Professionals*. Edited by Joyce M. Ray. West Lafayette, IN: Purdue University Press, 2014. eBook Collection, EBSCOhost, ISBN 9781461956815. Accessed February 19, 2016.
- Witt, Michael. "Co-designing, Co-developing, and Co-implementing an Institutional Data Repository Service." *Journal of Library Administration* 52, no. 2 (2012): 172–88.