Introductions

Heidi Imker, University of Illinois
Cynthia Hudson-Vitale, Washington University in St. Louis
Rob Olendorf, Pennsylvania State University
Lisa Johnston, University of Minnesota
Jake Carlson, University of Michigan
Claire Stewart, University of Minnesota
Wendy Kozlowski, Cornell University

2 Association of Research Libraries

#ARLSPECKit354
What do we mean by Data Curation?

Data curation may be broadly defined as the active and on-going management of data through its lifecycle of interest and usefulness to scholarly and educational activities.

Demographics

Survey sent to 124 ARL Institutions
Open: Jan 3, 2017—Jan 30, 2017
80 survey responses completed (65% response rate)

Citation: [http://old.arl.org/arl/membership/members.shtml](http://old.arl.org/arl/membership/members.shtml)
Goal of the Survey

Our research was intended to understand:

• current staffing and infrastructure (policy and technical) at ARL member institutions for data curation,

• current level of demand for data curation services, and

• any challenges that institutions are currently facing regarding providing data curation services.
Secondary Goal

Begin to establish a community of practice for data curators as part of our work on the Data Curation Network project—a cross-institutional staffing model for data curation.

https://sites.google.com/site/datacurationnetwork/
Does your institution currently provide research data curation services?

Most institutions were already or in the process of providing data curation services.

- Yes: 51
- In Process: 13
- No: 16
Please enter the year your institution begin providing data curation services.

More than half of the institutions currently providing services (35 out of 51) started doing so in 2010 or later.
Which subject domains represent the greatest demand for your data curation services?

Demand from the arts & humanities edged out both engineering and applied sciences and the physical sciences (20 and 19 responses, respectively).
Please indicate how many staff members’ work responsibilities focus exclusively/partially on providing data curation services.

N = 49

Many libraries spread responsibility for services across partial staff.
Finding: Data curation services often includes repository services

- 90% that provide data curation services also provide repository services for data.
- 22% are self-deposit
  30% are mediated deposit
  48% are a combination of both
- The majority of data repositories (78%) limit the size of file uploads with an average reported at around 2.5 GB per file.
- 65% of the current providers also help researchers prepare their data for deposit to external repositories.
- The external data repositories they support most often are ICPSR, Figshare, and the Open Science Framework.
Does your library currently provide local repository services for research data (institutional repository, data repository, other)?

Most data curation providers (46) also provide repository services for data.

- Yes: 90%
- No: 10%
- A stand-alone data repository: 15%
- A disciplinary repository that accepts data: 16%
- An institutional repository that accepts data: 57%
- Other service, please briefly describe: 2%

N = 51
Which of the following platforms are you using for your data repository? Check all that apply.

DSpace is the most common repository platform and is used by 22 of the reporting institutions.
How many new data sets does your data repository service receive and curate each month, on average?

The majority of institutions curate 1 or fewer datasets per month.

<table>
<thead>
<tr>
<th>Number of data sets received</th>
<th>Number of data sets curated</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>&lt;1</td>
<td>8</td>
</tr>
<tr>
<td>1</td>
<td>14</td>
</tr>
<tr>
<td>2–10</td>
<td>6</td>
</tr>
<tr>
<td>&gt;10</td>
<td>4</td>
</tr>
</tbody>
</table>

N = 41
Please enter the total number of data sets in your repository.

N = 43

Median number of datasets is 39.
What metadata schema are you primarily using for discovery of data?

Dublin Core is the most common metadata schema used.
Finding: Data curation policies and tools vary considerably across institutions

- Fewer than half support sensitive data
- Only 17 institutions require documentation or readme files. But 32 institutions reported that they provide support in creating them.
- The most commonly used tools:
  - BagIt: 13
  - Fixity: 12
  - Bitcurator: 9
  - FITS: 9
  - JHOVE: 9
- The most commonly employed persistent identifiers:
  - Handles: 26
  - DataCite DOI: 25
  - CrossRef DOI: 9
  - PURLS: 5
  - ARKS: 4
Finding: Data preservation platforms are less common

- 68% provide preservation services for curated data.
- Data preservation commitment
  At least 10 years: 14
  12–25 years: 4
  Indefinitely: 10
- Preservation platforms for data vary widely and one participant responded: “We presently steer clear of the word preservation, relying instead on long-term stewardship as our nomenclature.”
Please indicate your institution's level of support for...

<table>
<thead>
<tr>
<th>Curation Step</th>
<th>Data Curation Activities (47)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ingest</td>
<td>authentication; chain of custody; deposit agreement; documentation; file validation; metadata</td>
</tr>
<tr>
<td>Appraisal</td>
<td>rights management; risk management; selection</td>
</tr>
<tr>
<td>Processing &amp; Review</td>
<td>arrangement and description; code review; contextualize; conversion; curation log; data cleaning; de-identification; file format transformations; file inventory; file renaming; indexing; interoperability; peer-review; persistent identifier; quality assurance; restructure; software registry; transcoding</td>
</tr>
<tr>
<td>Access</td>
<td>contact information; data citation; data visualization; discovery services; embargo; file download; full-text indexing; metadata brokerage; restricted access; terms of use; use analytics</td>
</tr>
<tr>
<td>Preservation</td>
<td>cease data curation; emulation; file audit; migration; repository certification; secure storage; succession planning; technology monitoring and refresh; versioning</td>
</tr>
</tbody>
</table>
92% of libraries currently provide one or more of these services.

N = 49
Support for Access activities

These curation activities are frequently a function of the repository technology.
Support for Processing and Review activities (Part 1)

Comment:

“These activities require a high degree of both technical training and disciplinary knowledge.”
Support for Processing and Review activities (Part 2)

Comment:

“These activities require a high degree of both technical training and disciplinary knowledge.”

N = 49
Support for Preservation activities

Comment:

“Some of these activities are dependent on infrastructures provided by departments outside the Libraries but within the university.”

N = 49
Support for Appraisal activities

Risk management was commonly viewed as the responsibility of the depositor.

N = 49

#ARLSPECKit354
Finding: Aspirational vs. Not the Libraries’ role

“We believe all this is important, just not things the LIBRARY needs to do or should do.”

Data curation activities that librarians would like to perform but are unable to:
- Repository Certification: 30
- Software Registry: 23
- Interoperability: 28

No interest in providing:
- Code Review: 10
- Emulation: 14
- Peer Review: 20
- Software Registry: 12
- Deidentification: 11
Finding: Challenges to providing data curation services

Challenges for providing data curation services

- Training library staff: 6 Not challenging, 14 Neutral, 30 Very challenging
- Recruiting curation staff: 12 Not challenging, 10 Neutral, 28 Very challenging
- Outreach/Marketing: 10 Not challenging, 10 Neutral, 30 Very challenging
- Changing requirements: 9 Not challenging, 16 Neutral, 25 Very challenging
- Expertise in domain data: 3 Not challenging, 11 Neutral, 36 Very challenging
- Keeping up technology: 8 Not challenging, 15 Neutral, 27 Very challenging
- Scaling, increased demand: 6 Not challenging, 11 Neutral, 33 Very challenging

N = 50

27 Association of Research Libraries

#ARLSPECKit354
Conclusions: Growth But Not Yet Maturity in Data Curation Services?

- A few institutions reported operation and maintenance of long-standing, established repositories with a high level of sophistication across the majority of curation activities.
- A larger subset of respondents recently took steps to develop and launch more robust curation services, such as curating data in an established IR or developing a standalone data repository.
- A final group of survey respondents have established core research data services, namely researcher training, data management plan reviews, and may accept datasets into library collections, but have yet to embark on the larger suite of possible curation activities.
Questions & Discussion

Join the conversation by typing questions in the chat box in the lower left corner of your screen.
Thank you!
Hello, I am Lee Anne George, coordinator of the SPEC Survey Program at the Association of Research Libraries, and I would like to thank you for joining us for this SPEC Survey Webcast. Today we will hear about the results of the survey on Data Curation. These results have been published in SPEC Kit 354.

Before we begin there are a few announcements:

Everyone but the presenters has been muted to cut down on background noise. So, if you are part a group today, feel free to speak among yourselves.

We do want you to join the conversation by typing questions in the chat box in the lower left corner of your screen. I will read the questions aloud before the presenters answer them.

This webcast is being recorded and we will send registrants the slides and a link to the recording in the next week.

Slide 2—Introductions

Now let me introduce today’s presenters:

Cynthia Hudson-Vitale is the Data Services Coordinator in Data and GIS Services at Washington University in St. Louis Libraries

Heidi Imker is the director of the Research Data Service at the University of Illinois at Urbana-Champaign

Lisa R. Johnston is the Research Data Management/Curation Lead at the University of Minnesota Twin Cities Libraries

Jake Carlson is the Research Data Services Manager at the University of Michigan Library

Wendy Kozlowski is Data Curation Specialist at Cornell University

Robert Olendorf is Science Data Librarian at Pennsylvania State University

and Claire Stewart, Associate University Librarian for Research and Learning at the University of Minnesota.

Use the hashtag #ARLSPECKit354 to continue the conversation with them on Twitter.

Now, let me turn the presentation over to Lisa.

Slide 3—What do we mean by Data Curation?

Hi everyone. I’m going to take the lead for today’s presentation and my co-authors are on the line ready to jump in with the Q&A.
With this survey we focused on Data curation which can be broadly defined as the active and on-going management of data through its lifecycle of interest and usefulness to scholarly and educational activities. Curatorial actions may include quality assurance, file integrity checks, documentation review, metadata creation, file transformations, and rights management.

Important to note that Data curation services may be provided with or without a local data repository (e.g., library may support local researchers prepare their data for deposit to an external data repository).

You might be asking, what is the difference between RDM and data curation. This distinction is admittedly murky.

A number of studies and surveys have recently assessed library engagement with the broader concept of research data management (RDM) services– such as DMP support or training researchers in DM best practices and other consultative roles. We specifically wanted to understand if and how libraries are taking a more hands-on approach to curate research data.

**Slide 4—Demographics**

Our online survey was open to 124 ARL institutions between Jan 3 – Jan 30 earlier this year. We received responses on behalf of 80 institutions, or a 65% response rate.

**Slide 5—Goal of the Survey**

The questions in the survey were targeted to address the following: What is the...

**Slide 6—Secondary Goal**

The authors of this survey are participating in the Data Curation Network project, a Sloan-funded grant that is developing a cross-institutional staffing model to account for the wide range of data types, formats and disciplinary aspects for curating research data. Therefore the results of this survey would help us better understand the landscape of data curation activities and staff currently doing this work. Our other research and materials on this topic are openly available on our website linked here.

**Slide 7—Does your institution currently provide research data curation services?**

In our first question we branched the survey between those that said “Yes” they were providing data curation services. And those that responded with “in process” or “no” were asked to rank the importance of various curation activities. Interestingly, only 20% of the sample, or 16 libraries, indicated that they do not provide nor are actively developing data curation services.

Today we will focus on the responses from just the “Current providers”

**Slide 8—Please enter the year your institution began providing data curation services.**

We asked:

And found that Data curation services appear to be a relatively recent initiative; more than half of the libraries that currently provide services (35 of 51) started doing so in 2010 or later.
Slide 9—Which subject domains represent the greatest demand for your data curation services?

The 51 responses to a question on the source of greatest demand for data curation services shows interest from researchers across subject domains. Life sciences and social sciences are most likely to ask for these services (33 responses each or 65%).

Perhaps somewhat surprisingly given the focus STEM disciplines often receive in discussing data, arts & humanities edged out both engineering and applied sciences and the physical sciences (21, 20, and 19 responses respectively)

Slide 10—Please indicate how many staff members’ work responsibilities focus exclusively/partially on providing data curation services

Interest in data curation services does not yet appear to have translated into strong staff levels to provide these services however. The survey asked how many staff focus 100% of their time and how many spend part of their time on data curation services. The responses show that the majority of libraries place responsibility for data curation services on a few individuals who have other duties to carry out.

Slide 11—Finding: Data curation services often includes repository services

Looking closer at the 51 institutions that provide data curation services, most (46 or 90%) also provide repository services for data. These repositories can be self deposit or mediated or both. Many limit upload file sizes of datasets, with the average reported at 2.5GB per file, more than half also assist with deposits to external data repositories (ICPSR, FigShare, OSF).

Slide 12—Does your library currently provide local repository services for research data (institutional repository, data repository, other)?

Here is a breakdown of the type of repository service. The majority (29) have an institutional repository that accepts data. A smaller number (8) have a stand-alone repository specific for data.

Slide 13—Which of the following platforms are you using for your data repository? Check all that apply

DSpace is the most common repository platform and is used by 22 of the reporting institutions. 11 use Dataverse (as either a hosted or a local installation), 10 use Fedora/Hydra, and 7 use Islandora.

Other platforms or custom solutions included

- Digital Commons, CKAN, RStar is our preservation repository,
- DataBrary
- Ruby on Rails app that integrates directly with our preservation system
- http://hubzero.org
- "SobekCM
- Hybrid DSpace and Apache platform.
- Maria-based, CSS Front-end
Slide 14—How many data sets does your data repository service receive and curate each month, on average?

The nascent nature of data curation services and treatments across the ARL institutional landscape is evident in a number of results from this survey. Although the Office of Science and Technology Policy memo on access to federally funded scientific data was released in 2013, library technical and human infrastructure are just now reaching the point of accepting and curating data. Of the 46 libraries that accept data, the receiving approximately one new dataset a month, and three receiving more than 10 a month.

Slide 15—Please enter the total number of data sets in your repository

Consequently most institutions (26 or 61%) have fewer than 50 data sets in their entire collection. Ten libraries have between 51 and 200 data sets but only 7 report having over 200 in their repository.

Slide 16—What metadata schema are you primarily using for discovery of data?

Describing data sets using standard metadata schemas is of significant importance for data discovery, dissemination, and reuse. Yet, there are many schemas to choose from, including discipline-specific, and institution specific. The current provider subset indicated six major metadata schemas are in use: DublinCore, MODS, DDI, DataCite, and Dataverse (which is based on a number of standards). A number of institutions also employ others, such as ISO19115, Geoblacklight, MARC, and VRACore4, or custom metadata schemas.

Additionally, many organizations use more than one schema for different purposes, and some institutions reported they use up to four.

Slide 17—Finding: Data curation policies and tools vary considerably across institutions

Curating sensitive data is a topic debated among data repository managers and librarians. Fewer than half of the respondents to a question on private or sensitive data (21 or 42%) reported their service supports sensitive data.

One who does explained how the process for curating such data is not insignificant: “We collaborated with compliance officers on our campus to establish workflows for sensitive and restricted data, addressing IRB, HIPPA, FERPA, and government and export controlled data. Our service is currently undergoing a formal RQA (research quality assurance) review to ensure regulatory compliance.”

Slide 18—Finding: Data preservation platforms are less common

One key component of the data curation lifecycle is data preservation. Preservation services (such as emulation, file audits, migration, secure storage, and succession planning) help ensure that the data and technology is reusable and stable over the long term.

The most common preservation-compliant metadata standards used are MODS and PREMIS (12 of 28 responses each or 43%). There is little standardization across institutions in backup services.
Many are employing tape systems and cloud services to ensure redundant copies of the data remain available.

**Slide 19—Please indicate your institution’s level of support for ...**

Data curation services comprise a variety of different types of activities. The survey asked respondents to indicate whether their service provides any of 47 different activities grouped into five different aspects of data curation: ingest, appraisal, processing and review, access, and preservation. If an activity is not currently included as a part of the service, we asked if they plan or aspire to include the activity in the future.

**Slide 20—Support for Ingest activities**

The most universally provided data curation services are ingest activities, which include metadata, deposit agreements, authentication, documentation, file validation, and chain of custody. Forty-five libraries (92%) currently provide one or more of these services and all but chain of custody are offered by more than two-thirds of the libraries.

**Slide 21—Support for Access activities**

The access category covers 11 activities that are likewise commonly supported. These curation activities with noticeably uniform levels of support for datasets are frequently a function of the repository technology.

Forty-three libraries currently provide one or more of these services. More than two-thirds provide file download, terms of use, discovery services, embargo, use analytics, metadata brokerage, and data citation. Only 14 provide data visualization.

**Slides 22–23—Support for Processing and Review activities**

Most of the responding libraries provide some of the 18 processing and review activities. However, this category shows an interesting bimodal distribution of results between activities that are currently supported and those the respondents would like to provide, but are unable to at this time.

As one respondent commented:

“These ten activities are the most difficult to implement because they are the most time consuming and resource intensive. These activities also require a high degree of both technical training and disciplinary knowledge. We are slowly working towards supporting these activities, however some, like peer-review, are and will continue to be out of reach. If depositors/users supply us with this metadata, and/or ask us for assistance, then we will provide this support where possible. However, we cannot currently provide large-scale support across all datasets deposited in our repository.”

This bifurcation is also seen for the nine activities in the preservation category and the three appraisal activities.

**Slide 24—Support for Preservation activities**

These curation activities with noticeably uniform levels of support for datasets are frequently a function of the repository technology.
Slide 29—Questions & Discussion

We welcome your questions. Please join the conversation by typing questions in the chat box in the lower left corner of your screen. I will read the questions aloud before the presenters answer them.

Slide 30—Thank you!

Thank you all for joining us today to discuss the results of the data curation SPEC survey. You will receive the slides and a link to the recording in the next week.