ASSOCIATION OF RESEARCH LIBRARIES

SPEC Survey on Data Curation

1. Introduction

Researchers are required by many federal and private funders and publishers to make the digital data underlying their research openly available for sharing and reuse. Merely making data available, though, is not enough to ensure its on-going viability and re-usability—the data must be curated to ensure/facilitate optimal discovery and re-use.

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. Curatorial actions may include quality assurance, file integrity checks, documentation review, metadata creation for discoverability, file transformations into archival formats, and suitable license/copyright. Data curation services may be provided with or without a local data repository (e.g., allowing deposit of data into the institutional repository or helping local researchers prepare their data for deposit to an external data repository).

Although a number of studies and surveys have recently been published on data services provided by libraries, they have focused more on the broader concept of research data management (RDM) or services, without detailing curation policies, staffing, and treatment actions described above. Although these reports have all been useful, the library community would benefit from a more thorough and comprehensive understanding of needs and services focused specifically on data curation.

The purpose of this survey is to uncover the current infrastructure (policy and technical) at ARL member institutions for data curation, explore the current level of demand for data curation services, and discover any challenges that institutions are currently facing regarding providing these services.

This survey was co-designed by **Cynthia Hudson-Vitale**, the Data Services Coordinator in Data and GIS Services at Washington University in St. Louis Libraries and **Heidi Imker**, the director of the Research Data Service at the University of Illinois at Urbana-Champaign in collaboration with the Data Curation Network project team, which also includes (lead) **Lisa R. Johnston**, the Research Data Management/Curation Lead at the University of Minnesota Twin Cities Libraries; **Jake Carlson**, the Research Data Services Manager at the University of

Michigan Library; **Wendy Kozlowski**, Data Curation Specialist at Cornell University; **Robert Olendorf**, Science Data Librarian at Pennsylvania State University, and **Claire Stewart**, Associate University Librarian for Research and Learning at the University of Minnesota. For more information on the Data Curation Network project, which is funded by the Alfred P. Sloan foundation, see https://sites.google.com/site/datacurationnetwork/.

Please complete this survey and send the requested documentation by **January 30, 2017**. If you are not able to complete the survey in one sitting, you may return to the survey and resume where you left off.

NB: You will need to use the same computer and Web browser each time you access the survey and have cookies enabled.

An * indicates a required response.

As always, individual responses to the survey will be treated confidentially and responses will only be reported in the aggregate.

Questions can be directed to the SPEC survey staff.

*** Select your institution:**

*** Please provide the following contact information:**

Name:

Job title:

E-mail:

ASSOCIATION OF RESEARCH LIBRARIES

SPEC Survey on Data Curation

2. Background

*** Does your institution currently provide research data curation services?**

- ◯ Yes
- ◯ No
- ◯ In process

If you answered "Yes" above, you will be directed to the section "Data Curation Service Demographics." If you answered "No" or "In process" above, you will be directed to the section "Importance of Data Curation Services."

ASSOCIATION OF RESEARCH LIBRARIES

SPEC Survey on Data Curation

3. Data Curation Service Demographics

Please enter the year your institution begin providing data curation services.

Who may take advantage of your data curation services?

- ◯ Only researchers affiliated with our institution
- ◯ Any researcher regardless of affiliation

Comments

Please indicate how many staff members' work responsibilities focus exclusively (100%) on providing data curation services and how many staff focus partially (less than 100%) on providing data curation services.

Exclusively:

Partially:

For staff who focus partially on data curation, please briefly describe about how much time they spend on these services, for example, "2 staff members at 50% time each."

Which subject domains represent the greatest demand for your data curation services? Check all that apply.

- [ ] Physical Sciences
- [ ] Life Sciences
- [ ] Engineering and Applied Sciences
- [ ] Agricultural and Natural Sciences
- [ ] Social Sciences
- [ ] Arts & Humanities
- [ ] Health Sciences
- [ ] Multi-disciplinary
- [ ] Library Science
- [ ] Other subject, please specify

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

- ( ) Yes
- ( ) No

ASSOCIATION OF RESEARCH LIBRARIES

SPEC Survey on Data Curation

4. Local Repository Services

Please enter the year your library began providing data repository services.

Which of the following statements best describes your repository service for data?

- A stand-alone data repository
- An institutional repository that accepts data
- A disciplinary repository that accepts data
- Other service, please briefly describe

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

- [ ] DSpace
- [ ] Digital Commons/BePress
- [ ] Fedora/Hydra
- [ ] iRODS
- [ ] Islandora
- [ ] Dataverse (hosted)
- [ ] Dataverse (local installation)
- [ ] Ckan/Dkan
- [ ] Custom solution
- [ ] Other platform, please specify

If you selected Custom solution above, please briefly describe it.

How many new data sets does your data repository service receive each month, on average?

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

Please enter the total number of data sets that have received curation treatments (reviewed/enhanced/processed) by library staff.

How many new data sets receive data curation services each month, on average?

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

In which of the following ways do researchers deposit data into your data repository?

- ◯ Self-deposit
- ◯ Mediated
- ◯ Both self-deposit and mediated
- ◯ Other process, please briefly describe

Are there individual file size upload limits for your data repository platform?

- ◯ Yes
- ◯ No

If yes, please specify the file size limit.

ASSOCIATION OF RESEARCH LIBRARIES

SPEC Survey on Data Curation

5. Support for External Repositories

Does library staff help researchers prepare or curate their data for deposit to external data repositories outside of your institution?

- ◯ Yes
- ◯ No

If yes, which external data repositories do you support most often? Check all that apply.

- ☐ Figshare
- ☐ Dryad
- ☐ Zenodo
- ☐ Open Science Framework
- ☐ ICPSR
- ☐ Harvard's Dataverse
- ☐ Genbank
- ☐ Other external data repository, please specify

Please enter any additional comments you have about external data repositories.

ASSOCIATION OF RESEARCH LIBRARIES

SPEC Survey on Data Curation

6. Curation Policies

Does your data curation service support private or sensitive data?

- Yes
- No

Comments

Does your data curation service support embargoes and/or restricted access conditions?

- Yes
- No

Comments

Please indicate if your data curation service requires any of the following documentation from depositors and if your service helps create any of the documentation for depositors. Check all that apply.

	Requires	Helps create
Code books	☐	☐
Metadata	☐	☐
Readme files	☐	☐
Methodology	☐	☐
Scripts or software used to analyze the data	☐	☐
Other documentation	☐	☐

If you selected Requires Other documentation above, please briefly describe what type of documentation.

If you selected Helps create Other documentation above, please briefly describe what type of documentation.

Which of the following tools are you using in your curation treatments and/or activities? Check all that apply.

- ☐ Bitcurator
- ☐ Data Accessionner
- ☐ Fixity
- ☐ FITS
- ☐ JHOVE
- ☐ BagIt
- ☐ Identity Finder
- ☐ Bulk Extractor
- ☐ Other tool, please briefly describe

Comments

How does your service provide persistent identifiers for data? Check all that apply.

- [ ] Datacite DOIs
- [ ] Crossref DOIs
- [ ] ARKs
- [ ] Handles
- [ ] PURLs
- [ ] Other identifier, please specify

Comments

ASSOCIATION OF RESEARCH LIBRARIES

SPEC Survey on Data Curation

7. Preservation Services

Does your data curation service provide preservation services for data?

- ◯ Yes
- ◯ No

Comments

If yes, please answer the following questions.
If no, please continue to the next screen.

Please enter the number of years your service will preserve the curated data.

Comments

Which of the following platforms are you using for your archiving/preservation solution/management? Check all that apply.

- [ ] Rosetta
- [ ] Archivematica
- [ ] Preservica
- [ ] Duraspace
- [ ] Custom solution
- [ ] Other platform, please specify

What metadata schema are you using for the preservation of data? Check all that apply.

- [ ] MODS
- [ ] METS
- [ ] PREMIS
- [ ] Other schema, please specify

How are you backing up the data sets currently curated? Check all that apply.

- [ ] Cloud Services (AWS, DropBox, Box, Duraspace, etc.)
- [ ] Local LOCKSS
- [ ] CLOCKSS
- [ ] Portico
- [ ] DPN
- [ ] Other service, please briefly describe

ASSOCIATION OF RESEARCH LIBRARIES

SPEC Survey on Data Curation

8. Support for Ingest Activities

Here are descriptions of six data curation ingest activities.

Authentication: The process of confirming the identity of a person, generally the depositor, who is contributing data to the data repository. (e.g., password authentication or authorization via digital signature). Used for tracking provenance of the data files.

Chain of Custody: Intentional recording of provenance metadata of the files (e.g., metadata about who created the file, when it was last edited, etc.) in order to preserve file authenticity when data are transferred to third parties.

Deposit Agreement: The certification by the data author (or depositor) that the data conform to all policies and conditions (e.g., do not violate any legal restrictions placed on the data) and are fit for deposit into the repository. A deposit agreement may also include rights transfer to the repository for ongoing stewardship.

Documentation: Information describing any necessary information to use and understand the data. Documentation may be structured (e.g., a code book) or unstructured (e.g., a plain text "Readme" file).

File Validation: A computational process to ensure that the intended data transfer to a repository was perfect and complete using means such as generating and validating file checksums (e.g., test if a digital file has changed at the bit level) and format validation to ensure that file types match their extensions.

Metadata: Information about a data set that is structured (often in machine-readable format) for purposes of search and retrieval. Metadata elements may include basic information (e.g. title, author, date created, etc.) and/or specific elements inherent to datasets (e.g., spatial coverage, time periods).

Please indicate your institution's level of support for these data curation ingest activities on a scale of 1 to 5 where 1=currently providing; 2=will provide in the near future; 3=would like to provide, but unable to at this time; 4=no interest/desire to provide; 5=unsure.

	1	2	3	4	5
Authentication	○	○	○	○	○
Chain of custody	○	○	○	○	○
Deposit agreement	○	○	○	○	○
Documentation	○	○	○	○	○
File validation	○	○	○	○	○
Metadata	○	○	○	○	○

Comments

ASSOCIATION OF RESEARCH LIBRARIES

SPEC Survey on Data Curation

9. Support for Appraisal Activities

Here are descriptions of three data curation appraisal activities.

Rights Management: The process of tracking and managing ownership and copyright inherent to a data set as well as monitoring conditions and policies for access and reuse (e.g., licenses and data use agreements).

Risk Management: The process of reviewing data for known risks such as confidentiality issues inherent to human subjects data, sensitive information (e.g., sexual histories, credit card information) or data regulated by law (e.g. HIPAA, FERPA) and taking actions to reject or facilitate remediation (e.g., de-identification services) when necessary.

Selection: The result of a successful appraisal. The data are determined appropriate for acceptance and ingest into the repository according to local collection policy and practice.

Please indicate your institution's level of support for these data curation appraisal activities on a scale of 1 to 5 where 1=currently providing; 2=will provide in the near future; 3=would like to provide, but unable to at this time; 4=no interest/desire to provide; 5=unsure.

	1	2	3	4	5
Rights Management	○	○	○	○	○
Risk Management	○	○	○	○	○
Selection	○	○	○	○	○

Comments

ASSOCIATION OF RESEARCH LIBRARIES

SPEC Survey on Data Curation

10. Support for Processing and Review Activities part 1

Here are descriptions of eight data curation processing and review activities.

Arrangement and Description: The re-organization of files (e.g., new folder directory structure) in a dataset that may also involve the creation of new file names, file descriptions, and the recording of technical metadata inherent to the files (e.g., date last modified).

Code Review: Run and validate computer code (e.g., look for missing files and/or errors) in order to find mistakes overlooked in the initial development phase, improving the overall quality of software.

Contextualize: Use metadata to link the data set to related publications, dissertations, and/or projects that provide added context to how the data were generated and why.

Conversion (Analog): In effort to increase the usability of a data set, the information is transferred into digital file formats (e.g., analog data keyed into a database). Note: digital conversion is also used to convert "fixed" data (e.g., PDF formats) into machine-readable formats.

Curation Log: A written record of any changes made to the data during the curation process and by whom. File is often preserved as part of the overall record.

Data Cleaning: A process used to improve data quality by detecting and correcting (or removing) defects & errors in data.

Deidentification: Redacting or removing personally identifiable or protected information (e.g., sensitive geographic locations) from a dataset prior to sharing with third parties.

File Format Transformations: Transform files into open, non-proprietary file formats that broaden the potential for long-term reuse and ensure that additional preservation actions might be taken in the future. Note: Retention of the original file formats may be necessary if data transfer is not perfect.

Please indicate your institution's level of support for these data curation processing and review activities on a scale of 1 to 5 where 1=currently providing; 2=will provide in the near future; 3=would like to provide, but unable to at this time; 4=no interest/desire to provide; 5=unsure.

	1	2	3	4	5
Arrangement and Description	○	○	○	○	○
Code review	○	○	○	○	○
Contextualize	○	○	○	○	○
Conversion (Analog)	○	○	○	○	○
Curation Log	○	○	○	○	○
Data Cleaning	○	○	○	○	○
Deidentification	○	○	○	○	○
File Format Transformations	○	○	○	○	○

Comments

ASSOCIATION OF RESEARCH LIBRARIES

SPEC Survey on Data Curation

11. Support for Processing and Review Activities part 2

Here are descriptions of ten more data curation processing and review activities.

File Inventory or Manifest: The data files are inspected periodically and the number, file types (extensions), and file sizes of the data are understood and documented. Any missing, duplicate, or corrupt (e.g., unable to open) files are discovered.

File Renaming: To rename files in a dataset, often to standardize and/or reflect important metadata.

Indexing: Verify all metadata provided by the author and crosswalk to descriptive and administrative metadata compliant with a standard format for repository interoperability.

Interoperability: Formatting the data using a disciplinary standard for better integration with other datasets and/or systems.

Peer-review: The review of a data set by an expert with similar credentials and subject knowledge as the data creator for the purposes of validating the soundness and trustworthiness of the file contents.

Persistent Identifier: A URL (or Uniform Resource Locator) that is monitored by an authority to ensure a stable web location for consistent citation and long-term discoverability. Provides redirection when necessary (e.g., a Digital Object Identifier or DOI).

Quality Assurance: Ensure that all documentation and metadata are comprehensive and complete. Example actions might include: open and run the data files; inspect the contents in order to validate, clean, and/or enhance data for future use; look for missing documentation about codes used, the significance of "null" and "blank" values, or unclear acronyms.

Restructure: Organize and/or reformate poorly structured data files to clarify their meaning and importance.

Software Registry: Maintain copies of modern and obsolete versions of software (and any relevant code libraries) so that data may be opened/used overtime.

Transcoding: With audio and video files, detect technical metadata (min resolution, audio/video codec) and encode files in ways that optimize reuse and long-term preservation actions (e.g., Convert QuickTime files to MPEG4).

Please indicate your institution's level of support for these data curation processing and review activities on a scale of 1 to 5 where 1=currently providing; 2=will provide in the near future; 3=would like to provide, but unable to at this time; 4=no interest/desire to provide; 5=unsure.

	1	2	3	4	5
File Inventory or Manifest	○	○	○	○	○
File renaming	○	○	○	○	○
Indexing	○	○	○	○	○
Interoperability	○	○	○	○	○
Peer-review	○	○	○	○	○
Persistent Identifier	○	○	○	○	○
Quality Assurance	○	○	○	○	○
Restructure	○	○	○	○	○
Software Registry	○	○	○	○	○
Transcoding	○	○	○	○	○

Comments

ASSOCIATION OF RESEARCH LIBRARIES

SPEC Survey on Data Curation

12. Support for Access Activities

Here are descriptions of eleven data curation access activities.

Contact Information: Keep up-to-date contact information for the data authors and/or the contact persons in order to facilitate connection with third-party users. Often involves managing ephemeral information that will change over time.

Data Citation: Display of a recommended bibliographic citation for a dataset to enable appropriate attribution by third-party users in order to formally incorporate data reuse as part of the scholarly ecosystem.

Data Visualization: The presentation of pictorial and/or graphical representations of a data set used to identify patterns, detect errors, and/or demonstrate the extent of a data set to third party users.

Discovery Services: Services that incorporate machine-based search and retrieval functionality that help users identify what data exist, where the data are located, and how can they be accessed (e.g., full-text indexing or web optimization).

Embargo: To restrict or mediate access to a data set, usually for a set period of time. In some cases an embargo may be used to protect not only access, but any knowledge that the data exist.

File Download: Allow access to the data materials by authorized third parties.

Full-Text Indexing: Enhance the data for discovery purposes by generating search-engine-optimized formats of the text inherent to the data.

Metadata Brokerage: Active dissemination of a data set's metadata to search and discovery services (e.g., article databases, catalogs, web-based indexes) for federated search and discovery.

Restricted Access: In order to maintain the privacy of research subjects without losing integral components of the data, some data access will be protected and/or mediated to individuals that meet predefined criteria.

Terms of Use: Information provided to end users of a data set that outline the requirements or conditions for use (e.g., a Creative Commons License).

Use Analytics: Monitor and record how often data are viewed, requested, and/or downloaded. Track and report reuse metrics, such as data citations and impact measures for the data over time.

Please indicate your institution's level of support for these data curation access activities on a scale of 1 to 5 where 1=currently providing; 2=will provide in the near future; 3=would like to provide, but unable to at this time; 4=no interest/desire to provide; 5=unsure.

	1	2	3	4	5
Contact Information	○	○	○	○	○
Data Citation	○	○	○	○	○
Data Visualization	○	○	○	○	○
Discovery Services	○	○	○	○	○
Embargo	○	○	○	○	○
File download	○	○	○	○	○
Full-Text Indexing	○	○	○	○	○
Metadata Brokerage	○	○	○	○	○
Restricted Access	○	○	○	○	○
Terms of Use	○	○	○	○	○
Use Analytics	○	○	○	○	○

Comments

ASSOCIATION OF RESEARCH LIBRARIES

SPEC Survey on Data Curation

13. Support for Preservation Activities

Here are descriptions of nine data curation preservation activities.

Cease Data Curation: Plan for any contingencies that will ultimately terminate access to the data. For example, providing tombstones or metadata records for data that have been deselected and removed from stewardship.

Emulation: Provide legacy system configurations in modern equipment in order to ensure long-term usability of data (e.g., arcade games emulated on modern web-browsers)

File Audit: Periodic review of the digital integrity of the data files and taking action when needed to protect data from digital erosion (e.g., bitrot) and/or hardware failure.

Migration: Monitor and anticipate file format obsolescence and, as needed, transform obsolete file formats to new formats as standards and use dictate.

Repository Certification: The technical and administrative capacities of the repository undergo review through a transparent and well-documented process by a trusted third-party accreditation body (e.g., TRAC, or Data Seal of Approval).

Secure Storage: Data files are properly stored in a well-configured (in terms of hardware and software) storage environment that is routinely backed-up and physically protected. Perform routine fixity checks (to detect degradation or loss) and provide recovery services as needed.

Succession Planning: Planning for contingency, and/or escrow arrangements, in the case that the repository (or other entity responsible) ceases to operate or the institution substantially changes its scope.

Technology Monitoring and Refresh: Formal, periodic review and assessment to ensure responsiveness to technological developments and evolving requirements of the digital infrastructure and hardware storing the data.

Versioning: Provide mechanisms to ingest new versions of the data overtime that includes metadata describing the version history and any changes made for each version.

Please indicate your institution's level of support for these data curation preservation activities on a scale of 1 to 5 where 1=currently providing; 2=will provide in the near future; 3=would like to provide, but unable to at this time; 4=no interest/desire to provide; 5=unsure.

	1	2	3	4	5
Cease Data Curation	○	○	○	○	○
Emulation	○	○	○	○	○
File Audit	○	○	○	○	○
Migration	○	○	○	○	○
Repository Certification	○	○	○	○	○
Secure Storage	○	○	○	○	○
Succession Planning	○	○	○	○	○
Technology Monitoring and Refresh	○	○	○	○	○
Versioning	○	○	○	○	○

Comments

ASSOCIATION OF RESEARCH LIBRARIES

SPEC Survey on Data Curation

14. Challenges

Please indicate how challenging you expect the following aspects of data curation to be in the next 3 to 5 years on a scale of 1 to 5 where 1=Not challenging and 5=Very challenging.

	1 Not challenging	2	3	4	5 Very challenging
Training and retooling library staff to support data curation services	○	○	○	○	○
Recruiting and retaining data curation staff	○	○	○	○	○
Outreach/Marketing of services	○	○	○	○	○
Changing journal/funder/domain requirements for data sharing	○	○	○	○	○
Expertise in curating certain domain data	○	○	○	○	○
Keeping up with technology changes	○	○	○	○	○
Scaling curation services with increased demand	○	○	○	○	○

Please enter any additional comments you have about data curation challenges.

ASSOCIATION OF RESEARCH LIBRARIES

SPEC Survey on Data Curation

15. Importance of Data Curation Services

While your library may not currently provide data curation services and treatments, the project team is interested in understanding which curation treatments you and your institution find important. The following sections will provide a list of treatments and definitions for five categories of data curation services. Please indicate the importance of these treatments along the specified spectrum.

Ingest Activities
Here are descriptions of six data curation ingest activities.

Authentication: The process of confirming the identity of a person, generally the depositor, who is contributing data to the data repository. (e.g., password authentication or authorization via digital signature). Used for tracking provenance of the data files.

Chain of Custody: Intentional recording of provenance metadata of the files (e.g., metadata about who created the file, when it was last edited, etc.) in order to preserve file authenticity when data are transferred to third parties.

Deposit Agreement: The certification by the data author (or depositor) that the data conform to all policies and conditions (e.g., do not violate any legal restrictions placed on the data) and are fit for deposit into the repository. A deposit agreement may also include rights transfer to the repository for ongoing stewardship.

Documentation: Information describing any necessary information to use and understand the data. Documentation may be structured (e.g., a code book) or unstructured (e.g., a plain text "Readme" file).

File Validation: A computational process to ensure that the intended data transfer to a repository was perfect and complete using means such as generating and validating file checksums (e.g., test if a digital file has changed at the bit level) and format validation to ensure that file types match their extensions.

Metadata: Information about a data set that is structured (often in machine-readable format) for purposes of search and retrieval. Metadata elements may include basic information (e.g. title, author, date created, etc.) and/or specific elements inherent to datasets (e.g., spatial coverage, time periods).

Please indicate the importance of these data curation ingest activities on a scale of 1 to 5 where 1=essential; 2=very important; 3=moderately important; 4=less important; 5=not important.

	1	2	3	4	5
Authentication	○	○	○	○	○
Chain of custody	○	○	○	○	○
Deposit agreement	○	○	○	○	○
Documentation	○	○	○	○	○
File validation	○	○	○	○	○
Metadata	○	○	○	○	○

Comments

ASSOCIATION OF RESEARCH LIBRARIES

SPEC Survey on Data Curation

16. Importance of Appraisal Activities

Here are descriptions of three data curation appraisal activities.

Rights Management: The process of tracking and managing ownership and copyright inherent to a data set as well as monitoring conditions and policies for access and reuse (e.g., licenses and data use agreements).

Risk Management: The process of reviewing data for known risks such as confidentiality issues inherent to human subjects data, sensitive information (e.g., sexual histories, credit card information) or data regulated by law (e.g. HIPAA, FERPA) and taking actions to reject or facilitate remediation (e.g., de-identification services) when necessary.

Selection: The result of a successful appraisal. The data are determined appropriate for acceptance and ingest into the repository according to local collection policy and practice.

Please indicate the importance of these data curation appraisal activities on a scale of 1 to 5 where 1=essential; 2=very important; 3=moderately important; 4=less important; 5=not important.

	1	2	3	4	5
Rights Management	○	○	○	○	○
Risk Management	○	○	○	○	○
Selection	○	○	○	○	○

Comments

ASSOCIATION OF RESEARCH LIBRARIES

SPEC Survey on Data Curation

17. Importance of Processing and Review Activities part 1

Here are descriptions of eight data curation processing and review activities.

Arrangement and Description: The re-organization of files (e.g., new folder directory structure) in a dataset that may also involve the creation of new file names, file descriptions, and the recording of technical metadata inherent to the files (e.g., date last modified).

Code Review: Run and validate computer code (e.g., look for missing files and/or errors) in order to find mistakes overlooked in the initial development phase, improving the overall quality of software.

Contextualize: Use metadata to link the data set to related publications, dissertations, and/or projects that provide added context to how the data were generated and why.

Conversion (Analog): In effort to increase the usability of a data set, the information is transferred into digital file formats (e.g., analog data keyed into a database). Note: digital conversion is also used to convert "fixed" data (e.g., PDF formats) into machine-readable formats.

Curation Log: A written record of any changes made to the data during the curation process and by whom. File is often preserved as part of the overall record.

Data Cleaning: A process used to improve data quality by detecting and correcting (or removing) defects & errors in data.

Deidentification: Redacting or removing personally identifiable or protected information (e.g., sensitive geographic locations) from a dataset prior to sharing with third parties.

File Format Transformations: Transform files into open, non-proprietary file formats that broaden the potential for long-term reuse and ensure that additional preservation actions might be taken in the future. Note: Retention of the original file formats may be necessary if data transfer is not perfect.

Please indicate the importance of these data curation processing and reveiw activities on a scale of 1 to 5 where 1=essential; 2=very important; 3=moderately important; 4=less important; 5=not important.

	1	2	3	4	5
Arrangement and Description	○	○	○	○	○
Code review	○	○	○	○	○
Contextualize	○	○	○	○	○
Conversion (Analog	○	○	○	○	○
Curation Log	○	○	○	○	○
Data Cleaning	○	○	○	○	○
Deidentification	○	○	○	○	○
File Format Transformations	○	○	○	○	○

Comments

ASSOCIATION OF RESEARCH LIBRARIES

SPEC Survey on Data Curation

18. Importance of Processing and Review Activities part 2

Here are descriptions of ten more data curation processing and review activities.

File Inventory or Manifest: The data files are inspected periodically and the number, file types (extensions), and file sizes of the data are understood and documented. Any missing, duplicate, or corrupt (e.g., unable to open) files are discovered.

File Renaming: To rename files in a dataset, often to standardize and/or reflect important metadata.

Indexing: Verify all metadata provided by the author and crosswalk to descriptive and administrative metadata compliant with a standard format for repository interoperability.

Interoperability: Formatting the data using a disciplinary standard for better integration with other datasets and/or systems.

Peer-review: The review of a data set by an expert with similar credentials and subject knowledge as the data creator for the purposes of validating the soundness and trustworthiness of the file contents.

Persistent Identifier: A URL (or Uniform Resource Locator) that is monitored by an authority to ensure a stable web location for consistent citation and long-term discoverability. Provides redirection when necessary (e.g., a Digital Object Identifier or DOI).

Quality Assurance: Ensure that all documentation and metadata are comprehensive and complete. Example actions might include: open and run the data files; inspect the contents in order to validate, clean, and/or enhance data for future use; look for missing documentation about codes used, the significance of "null" and "blank" values, or unclear acronyms.

Restructure: Organize and/or reformate poorly structured data files to clarify their meaning and importance.

Software Registry: Maintain copies of modern and obsolete versions of software (and any relevant code libraries) so that data may be opened/used overtime.

Transcoding: With audio and video files, detect technical metadata (min resolution, audio/video codec) and encode files in ways that optimize reuse and long-term preservation actions (e.g., Convert QuickTime files to MPEG4).

Please indicate the importance of these data curation processing and reveiw activities on a scale of 1 to 5 where 1=essential; 2=very important; 3=moderately important; 4=less important; 5=not important.

	1	2	3	4	5
File Inventory or Manifest	○	○	○	○	○
File renaming	○	○	○	○	○
Indexing	○	○	○	○	○
Interoperability	○	○	○	○	○
Peer-review	○	○	○	○	○
Persistent Identifier	○	○	○	○	○
Quality Assurance	○	○	○	○	○
Restructure	○	○	○	○	○
Software Registry	○	○	○	○	○
Transcoding	○	○	○	○	○

Comments

ASSOCIATION OF RESEARCH LIBRARIES

SPEC Survey on Data Curation

19. Importance of Access Activities

Here are descriptions of eleven data curation access activities.

Contact Information: Keep up-to-date contact information for the data authors and/or the contact persons in order to facilitate connection with third-party users. Often involves managing ephemeral information that will change over time.

Data Citation: Display of a recommended bibliographic citation for a dataset to enable appropriate attribution by third-party users in order to formally incorporate data reuse as part of the scholarly ecosystem.

Data Visualization: The presentation of pictorial and/or graphical representations of a data set used to identify patterns, detect errors, and/or demonstrate the extent of a data set to third party users.

Discovery Services: Services that incorporate machine-based search and retrieval functionality that help users identify what data exist, where the data are located, and how can they be accessed (e.g., full-text indexing or web optimization).

Embargo: To restrict or mediate access to a data set, usually for a set period of time. In some cases an embargo may be used to protect not only access, but any knowledge that the data exist.

File Download: Allow access to the data materials by authorized third parties.

Full-Text Indexing: Enhance the data for discovery purposes by generating search-engine-optimized formats of the text inherent to the data.

Metadata Brokerage: Active dissemination of a data set's metadata to search and discovery services (e.g., article databases, catalogs, web-based indexes) for federated search and discovery.

Restricted Access: In order to maintain the privacy of research subjects without losing integral components of the data, some data access will be protected and/or mediated to individuals that meet predefined criteria.

Terms of Use: Information provided to end users of a data set that outline the requirements or conditions for use (e.g., a Creative Commons License).

Use Analytics: Monitor and record how often data are viewed, requested, and/or downloaded. Track and report reuse metrics, such as data citations and impact measures for the data over time.

Please indicate the importance of these data curation access activities on a scale of 1 to 5 where 1=essential; 2=very important; 3=moderately important; 4=less important; 5=not important.

	1	2	3	4	5
Contact Information	○	○	○	○	○
Data Citation	○	○	○	○	○
Data Visualization	○	○	○	○	○
Discovery Services	○	○	○	○	○
Embargo	○	○	○	○	○
File download	○	○	○	○	○
Full-Text Indexing	○	○	○	○	○
Metadata Brokerage	○	○	○	○	○
Restricted Access	○	○	○	○	○
Terms of Use	○	○	○	○	○
Use Analytics	○	○	○	○	○

Comments

ASSOCIATION OF RESEARCH LIBRARIES

SPEC Survey on Data Curation

20. Importance of Preservation Activities

Here are descriptions of nine data curation preservation activities.

Cease Data Curation: Plan for any contingencies that will ultimately terminate access to the data. For example, providing tombstones or metadata records for data that have been deselected and removed from stewardship.

Emulation: Provide legacy system configurations in modern equipment in order to ensure long-term usability of data (e.g., arcade games emulated on modern web-browsers)

File Audit: Periodic review of the digital integrity of the data files and taking action when needed to protect data from digital erosion (e.g., bitrot) and/or hardware failure.

Migration: Monitor and anticipate file format obsolescence and, as needed, transform obsolete file formats to new formats as standards and use dictate.

Repository Certification: The technical and administrative capacities of the repository undergo review through a transparent and well-documented process by a trusted third-party accreditation body (e.g., TRAC, or Data Seal of Approval).

Secure Storage: Data files are properly stored in a well-configured (in terms of hardware and software) storage environment that is routinely backed-up and physically protected. Perform routine fixity checks (to detect degradation or loss) and provide recovery services as needed.

Succession Planning: Planning for contingency, and/or escrow arrangements, in the case that the repository (or other entity responsible) ceases to operate or the institution substantially changes its scope.

Technology Monitoring and Refresh: Formal, periodic review and assessment to ensure responsiveness to technological developments and evolving requirements of the digital infrastructure and hardware storing the data.

Versioning: Provide mechanisms to ingest new versions of the data overtime that includes metadata describing the version history and any changes made for each version.

Please indicate the importance of these data curation preservation activities on a scale of 1 to 5 where 1=essential; 2=very important; 3=moderately important; 4=less important; 5=not important.

	1	2	3	4	5
Cease Data Curation	○	○	○	○	○
Emulation	○	○	○	○	○
File Audit	○	○	○	○	○
Migration	○	○	○	○	○
Repository Certification	○	○	○	○	○
Secure Storage	○	○	○	○	○
Succession Planning	○	○	○	○	○
Technology Monitoring and Refresh	○	○	○	○	○
Versioning	○	○	○	○	○

Comments

ASSOCIATION OF RESEARCH LIBRARIES

SPEC Survey on Data Curation

21. Additional Comments

Please enter any additional information regarding data curation practices at your institution that may assist the authors in accurately analyzing the results of this survey.

ASSOCIATION OF RESEARCH LIBRARIES

SPEC Survey on Data Curation

22. Call for Documents

Please provide the URLs for the following documents/web pages relating to data curation in your library.

If these documents are not available on the Web or if the URL is for a page that is accessible only by the library staff, mail or email the document(s) by (or soon after) **January 30, 2017** to:

ARL SPEC Surveys
21 Dupont Circle NW
Suite 800
Washington, DC 20036

OR

spec@arl.org

NB: Submitted documents may be chosen for inclusion in the published SPEC Kit which will be distributed in both print and electronic formats on a variety of platforms, such as the ARL Digital Publications website, digital repositories, and HathiTrust. Please alert the SPEC survey staff if a submitted document should not be published.

Descriptions of data curation services

Descriptions of data curation infrastructure

Data repository webpage

Data curation workflows

Data curation staffing/organizational models

Data models/metadata schemas

Job descriptions for data curation services staff

Data deaccessioning policy

Check here if print documentation will be sent by mail or email.

- [ ] Will send document(s) by mail.
- [ ] Will send document(s) by email.

ASSOCIATION OF RESEARCH LIBRARIES

SPEC Survey on Data Curation

23. Thank you

Thank you for your contribution to this survey!

Questions about the survey, or a request for a PDF of your survey response, can be directed to the SPEC survey staff.