Data Curation Primers: Expanding the Data Curation Toolkit

Presented to the DataOne community 9-10-2019
The Data Curation Network: Radical Collaboration

Lisa Johnston
Mission

“The Data Curation Network will enable researchers that are faced with a growing number of requirements to ethically share their research data in ways that make it findable, accessible, interoperable and reusable (FAIR).”
What is data curation?

Data curators enrich research data publications and ensure the data are FAIR, by:

- Finding and adding missing files and documentation
- Screening for privacy disclosure risk
- Detecting and fixing code and other quality assurance issues
- Transforming file formats for long term access
- Arranging and describing files
- Reviewing and augmenting metadata
Value of Curation

- We generate so much data!
- Data are messy (lack context!)
- Digital file formats are constantly at risk
- Most data never leaves their author’s laptop ⇒ benign neglect
- Sharing is good (funders, publishers, disciplines)

Curation makes data sharing betterrrrrr...

- Reuse (FAIR)
- Reproducibility
- Retractions (avoids them)
- Reputation (trust, transparency)
- References (citations)
Vision for the Data Curation Network

- Provide expert data curation services for network partners
- Offer professional development opportunities for an emerging data curator professional community
- Create and openly share data curation best practices
- Demonstrate that curated datasets are measurably of greater reuse value than non-curated data
- Expand into a sustainable entity that grows beyond our initial partner institutions
- Provide expert data curation services for network partners
- Offer professional development opportunities for an emerging data curator professional community
- Create and openly share data curation best practices
- Demonstrate that curated datasets are measurably of greater reuse value than non-curated data
- Expand into a sustainable entity that grows beyond our initial partner institutions
Value of Curation

- Better data

Value of Data Curation Network

- Better curation services
- Better best practices
- Better experiences
- Stronger relationships
- Community-led infrastructure
- Peer-to-peer learning
- ...and better data
DCN Expert Network
DCN CURATE Steps

DCN Curators will take CURATE steps for each data set, that is:

**Check** data files and read documentation

**Understand** the data (try to), if not...

**Request** missing information or changes

**Augment** the submission with metadata for findability

**Transform** file formats for reuse and long-term preservation

**Evaluate** and rate the overall submission for FAIRnness.

---

Table A1. Draft checklist of DCN CURATE steps and FAIRnness scorecard

<table>
<thead>
<tr>
<th>CURATE Actions</th>
<th>Curation Checklist</th>
</tr>
</thead>
<tbody>
<tr>
<td>Check data files and read documentation</td>
<td>□ Files open as expected</td>
</tr>
<tr>
<td>▪ Review the content of the data files (e.g., open and run the files or code).</td>
<td>□ Issues □</td>
</tr>
<tr>
<td>▪ Verify all metadata provided by the author and review the available documentation.</td>
<td>□ Code runs as expected</td>
</tr>
<tr>
<td></td>
<td>□ Produces minor errors</td>
</tr>
<tr>
<td></td>
<td>□ Does not run and/or produces many errors</td>
</tr>
<tr>
<td></td>
<td>□ Metadata quality is rich, accurate, and complete</td>
</tr>
<tr>
<td></td>
<td>□ Metadata has issues □</td>
</tr>
<tr>
<td></td>
<td>□ Documentation Type (circle)</td>
</tr>
<tr>
<td></td>
<td>Readme / Codebook / Data Dictionary / Other: □</td>
</tr>
<tr>
<td></td>
<td>□ Missing/None</td>
</tr>
<tr>
<td></td>
<td>□ Needs work</td>
</tr>
</tbody>
</table>

Understand the data (or try to)

- Check for quality assurance and usability issues such as missing

*Varies based on file formats and subject domain. For example...*
DCN Growth and Sustainability

- Curated 50 data sets since Jan 1, 2019!
- 2 new members in Year 2
- Aim to add two more in Year 3
- Canada and Dutch groups planning stages to launch their own network
- Exploring fiscal and administrative models to support beyond grant
Enhancing Expertise throughout the Broader Community

Cynthia Hudson Vitale

Specialized Data Curation Workshop @JHU 2019
DCN Education

https://sites.psu.edu/dcnworkshops/
1. Increase understanding of data curation practices and tools in various disciplines, data types, and formats.
2. Share expertise and enhance curation capacity for curation nationwide.
3. Meet like-minded colleagues who are interested in building and extending curation practices.
Specialized Data Curation Workshop Agenda
April 17th & 18th  •  Johns Hopkins University  •  Baltimore, Maryland

Wednesday

9:00  Welcome & Breakfast
9:30  The Value of Curation
10:00  Curation Deep Dive #1: C Step
10:30  Break
10:45  Curation Deep Dive #1: U Step
12pm  Lunch
1:00  Primer Timer → pitch idea of primer topics
1:30  Curation Deep Dive #2: R & A Steps
2:30  Break
3:00  Curation Deep Dive #2: R & A Steps continued
4:00  End of Day One
5:30  Reception

Thursday

9:00  Breakfast
9:30  Coffee with Data
10:00  Review Day 1
10:15  Curation Deep Dive #3: T Step
10:30  Break
12pm  Lunch
12:15  Curation Deep Dive #3: E & D Step
1:15  Primer Time 2
1:30  Group feedback on primers
2:00  Wrap up
2:15  Everyone Disperses
2:30  Everyone Disperses
Check files
Understand or try to
Request missing information
Augment the submission
Transform the format
Evaluate for FAIRness
Document throughout

Our curriculum engages attendees with lectures, group activities and demonstrations.

Pictured: Group activity at the DCN Specialized Data Curation Workshop, co-located at the DLF Forum on October 17-18, 2018.

www.datacurationnetwork.org
Hands-on data curation activities

Data Curation Assignment: Images (Penn State)

Title: S‘Urachi Site-Based Archaeological Survey 2015

Author: Victor T. Hail

Discipline: Archeology

Date: 2015

Access: Public

Reason for deposit: Connect to published article and report

Data Curation Network
Hands-on data curation activities

EVALUATE Step

<table>
<thead>
<tr>
<th>CURATE Action</th>
<th>Curator Checklist</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Evaluate</strong> and rate the overall data record for FAIRness.*</td>
<td>Findable -</td>
</tr>
<tr>
<td>• Score the dataset and recommend ways to increase the FAIRness of the data and become “DCN approved.”</td>
<td>- Metadata exceeds author/title/date,</td>
</tr>
<tr>
<td></td>
<td>- Unique PID (DOI, Handle, PURL, etc.).</td>
</tr>
<tr>
<td></td>
<td>- Discoverable via web search engines.</td>
</tr>
<tr>
<td>Accessible -</td>
<td>Retrieveable via a standard protocol (e.g., HTTP).</td>
</tr>
<tr>
<td></td>
<td>Free, open (e.g., download link).</td>
</tr>
<tr>
<td>Interoperable -</td>
<td>Metadata formatted in a standard schema (e.g., Dublin Core).</td>
</tr>
<tr>
<td></td>
<td>Metadata provided in machine-readable format (OAI feed).</td>
</tr>
<tr>
<td>Reusable -</td>
<td>Data include sufficient metadata about the data characteristics to reuse</td>
</tr>
<tr>
<td></td>
<td>Contact info displayed if the direct assistance of the author needed.</td>
</tr>
<tr>
<td></td>
<td>Clear indicators of who created, owns, and stewards the data.</td>
</tr>
<tr>
<td></td>
<td>Data are released with clear data usage terms (e.g., a CC License).</td>
</tr>
</tbody>
</table>

* Rubric evaluating the FAIR principles are based on the scoring matrix by Dunning, de Smaele, & Böhmer (2017).

Notes:
DCN Workshops by the numbers

Workshop @DLF

- 40 Primer topics pitched
- 10 Primer groups formed at the workshop
- 7 Completed primers

Workshop @JHU

- 26 Primer topics pitched
- 13 Primer groups formed

Applicants

- 44 - Workshop @DLF
- 56 - Workshop @JHU
- 59 - Workshop @WUSTL

Attendees

- 22 - Workshop @DLF
- 27 - Workshop @JHU
- 31 - Workshop @WUSTL
<table>
<thead>
<tr>
<th>Category</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total # of WS2 Applicants</td>
<td>56</td>
</tr>
<tr>
<td>Total # of Canadian Applicants</td>
<td>3</td>
</tr>
<tr>
<td>Total # of African Applicants</td>
<td>4</td>
</tr>
<tr>
<td>Total # of US Applicants</td>
<td>49</td>
</tr>
<tr>
<td>Total # of US student applicants</td>
<td>2</td>
</tr>
<tr>
<td>Total # of US government applicants</td>
<td>4</td>
</tr>
<tr>
<td>Total # of US archives applicants</td>
<td>2</td>
</tr>
<tr>
<td>Total # of US museum applicants</td>
<td>2</td>
</tr>
<tr>
<td>Total # of other US applicants</td>
<td>1</td>
</tr>
<tr>
<td>Total # of US library applicants</td>
<td>38</td>
</tr>
<tr>
<td>Total # of R1 (US)</td>
<td>32</td>
</tr>
<tr>
<td>Total # of R2 (US)</td>
<td>3</td>
</tr>
<tr>
<td>Total # of SF (US)</td>
<td>2</td>
</tr>
<tr>
<td>Total # of B/A&amp;S (US)</td>
<td>1</td>
</tr>
<tr>
<td>Total # of WS3 Applicants</td>
<td>59</td>
</tr>
<tr>
<td>--------------------------</td>
<td>----</td>
</tr>
<tr>
<td>Total # of Canadian Applicants</td>
<td>1</td>
</tr>
<tr>
<td>Total # of African Applicants</td>
<td>3</td>
</tr>
<tr>
<td>Total # of African non-profit applicants</td>
<td>2</td>
</tr>
<tr>
<td>Total # of US Applicants</td>
<td>55</td>
</tr>
<tr>
<td>Total # of US student applicants</td>
<td>1</td>
</tr>
<tr>
<td>Total # of US non-profit applicants</td>
<td>1</td>
</tr>
<tr>
<td>Total # of US grant funded applicants</td>
<td>1</td>
</tr>
<tr>
<td>Total # of US archives applicants</td>
<td>1</td>
</tr>
<tr>
<td>Total # of US library applicants</td>
<td>51</td>
</tr>
<tr>
<td>Total # of R1 (US)</td>
<td>35</td>
</tr>
<tr>
<td>Total # of R2 (US)</td>
<td>4</td>
</tr>
<tr>
<td>Total # of D/PU (US)</td>
<td>2</td>
</tr>
<tr>
<td>Total # of M (US)</td>
<td>5</td>
</tr>
<tr>
<td>Total # of B (US)</td>
<td>4</td>
</tr>
<tr>
<td>Total # of SF (US)</td>
<td>1</td>
</tr>
</tbody>
</table>
Attendee Feedback

What would you have liked to spend more time on during the workshop?
8 responses

- Hands on experience with the software: 4 (50%)
- Curating the example datasets: 3 (37.5%)
- Working in groups on the primers: 2 (25%)
- Networking with colleagues: 3 (37.5%)
- Discussing different data curation strategies: 6 (75%)
- Nothing, the timing worked well: 0 (0%)
- Crowd sourcing resources for further learning: 1 (12.5%)
Community Built Actionable Resources

Hannah Hadley

Together We Can Make Research Better
Primer creation process:

- Primer topics are selected at each workshop
- Each group receives a “roadmap” and primer template
- DCN mentors are assigned to assist each group
- Groups meet each month for six months
- Primer drafts are submitted about half-way through the process for peer review; co-occurring with a webinar
- Revisions are made based on review recommendations
- Final submissions are published to Github (archival copies are published to University of Minnesota’s DRUM repository)
Data Curation Primers are concise, actionable resources meant to assist data curation in adding value to a dataset.

https://github.com/DataCurationNetwork/data-primers
**Data Curation Network / data-primers**

**Branch:** master  
**data-primers** / Jupyter Notebook Data Curation Primer / Jupyter Notebooks Data Curation Primer.md

- **cynhudson** Add files via upload  
  566c52a on May 30

1 contributor

253 lines (186 sloc)  
17.2 KB

---

**DATA CURATION NETWORK**

**Jupyter Notebooks: A Primer for Data Curators**

**Participants:**

- Daina Bouquin, Center for Astrophysics. Harvard & Smithsonian. ([daina.bouquin@cfa.harvard.edu](mailto:daina.bouquin@cfa.harvard.edu))
### Primer Template

**Overview**

<table>
<thead>
<tr>
<th>Topic</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>File Extension</td>
<td></td>
</tr>
<tr>
<td>MIME Type</td>
<td></td>
</tr>
<tr>
<td>Structure</td>
<td></td>
</tr>
<tr>
<td>Versions</td>
<td></td>
</tr>
<tr>
<td>Primary fields or areas of use</td>
<td></td>
</tr>
<tr>
<td>Source and affiliation</td>
<td></td>
</tr>
<tr>
<td>Metadata standards</td>
<td></td>
</tr>
<tr>
<td>Key questions for curation review</td>
<td></td>
</tr>
<tr>
<td>Tools for curation review</td>
<td></td>
</tr>
<tr>
<td>Date Created</td>
<td></td>
</tr>
<tr>
<td>Created by</td>
<td></td>
</tr>
<tr>
<td>Date updated and summary of changes made</td>
<td></td>
</tr>
</tbody>
</table>

### Format overview

<table>
<thead>
<tr>
<th>Topic</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>File Extension</td>
<td>.gdb</td>
</tr>
<tr>
<td>MIME type</td>
<td></td>
</tr>
<tr>
<td>Structure</td>
<td></td>
</tr>
<tr>
<td>Versions</td>
<td></td>
</tr>
<tr>
<td>Primary fields or areas of use</td>
<td></td>
</tr>
<tr>
<td>Source and affiliation</td>
<td></td>
</tr>
<tr>
<td>Metadata standards</td>
<td>ISO19115, ISO19110, FGDC CSDGM content standards; .xml format (ISO19139, FGDC, Geospatial metadata)</td>
</tr>
<tr>
<td>Tools for curation review</td>
<td>ArcGIS Desktop (ArcMap, ArcCatalog), ArcGIS Pro, QGIS</td>
</tr>
<tr>
<td>Date created</td>
<td>February 4, 2019</td>
</tr>
<tr>
<td>Created by</td>
<td>Andrew Battista, Tom Brittnacher, Zenobie Garrett, Jennifer Moore, Carrie Pirmann</td>
</tr>
<tr>
<td>Date updated and summary of changes made</td>
<td>February 4, 2019</td>
</tr>
</tbody>
</table>
**Example Contents**

**Table of Contents**

- Format overview
- Description of format
- File geodatabases
- Personal geodatabase
- ArcSDE geodatabase
- Exploring geodatabases
- Examples of geodatabase datasets
- Public .gdb data collections
- Sample data set used in this document
- Key questions
- Instructions for resources to use in the curation review of geodatabase files
- Metadata
- Geospatial Metadata Standards
- Viewing and Exporting Metadata
- Metadata Completeness
- Other Metadata Schemas
- Preservation actions
- Bibliography

**Example Table of Contents [Optional Components]**

1. Description of format
2. Examples
3. Sample data set citations
4. Key questions to ask yourself
5. Key clarifications to get from researcher
6. Applicable metadata standard, core elements and readme requirements
7. Resources for reviewing data
8. Software for viewing or analyzing data
9. Preservation actions
10. What to look for to make sure this file meets FAIR principles
11. Ways in which fields may use this format
12. Unresolved Issues/Further Questions [for example: tracking provenance of data creation, level of detail in dataset]
13. Documentation of curation process: What do capture from curation process
14. Appendix A - filetype CURATED checklist
Our Peer Review Process

Primers are formally peer reviewed half-way through the six month process, or additionally if needed.

Reviewers include:

- Present and former DCN workshop/primer participants
- All Data Curation Network Members
- DCN Workshop Instructors

1. What steps in the curation process were you able to complete using this primer?
2. What sections of the primer did you find most useful?
3. Do you have suggestions for how content may be revised or enhanced?
Publication

Github

- Primers are expected to grow from their original version
- The community may suggest revisions

DRUM

- Contains the archived primer drafts from the IMLS supported workshops
- Version 1.0 only

https://github.com/DataCurationNetwork/data-primers
Jupyter Notebooks: A Primer for Data Curators

Participants:

- Daina Bouquin, Center for Astrophysics. Harvard & Smithsonian. (daina.bouquin@cfa.harvard.edu)
Geodatabases

Authors: Andrew Battista, Tom Brittnacher, Zenobie Garrett, Jennifer Moore & Carrie Pirmann

DCN Mentor: Mara Blake

https://github.com/DataCurationNetwork/data-primers

Major Benefits of the Primer:

- Clear steps on converting portions or all of geodatabases into shapefiles
- Advice on opening geodatabases with multiple formats
- Guidance on generating discipline standard metadata
- Insight on emerging discovery metadata standards (e.g., GeoBlacklight)
- Thoughts about long-term preservation and the ongoing support for these files
netCDF

Author: Sophie Hou

DCN Mentors: Jake Carlson & Susan Borda

https://github.com/DataCurationNetwork/data-primers

Key Curation Components:

- Intro to the Research Data Archive
- Instructions for the sample dataset
- Sample visualizations
- Answers to “Key Questions to Answer” in the main netCDF Primer using Panoply as the curation review tool
- Instructions for using the Integrated Data Viewer (IDV) for providing curation review

SPSS

Authors: Joshua Dull, Sai Deng, Shahira Khair & Jeanine Finn

DCN Mentor: Sophia Lafferty-Hess

https://github.com/DataCurationNetwork/data-primers

Key Curation Considerations:

● Preservation actions
  ○ Save as .por? To ASCII or not to ASCII?
  ○ Preservation recommendations
    ■ ICPSR, LOC and others
  ○ Suggested software for converting & reviewing SPSS files

● Further considerations
  ○ SPSS Version
  ○ Researcher feedback
    ■ Which files do researchers save?

● Other highlights
  ○ SPSS Tutorials
  ○ Bibliography for more curation resources
Microsoft Excel

Authors: Ho Jung Yoo, Sandra Sawchuk & Greg Janée

DCN Mentor: Wendy Kozlowski

https://github.com/DataCurationNetwork/data-primers

Key Curatorial Considerations:

There are no metadata standards for Microsoft Excel, so detailed documentation from the depositor is encouraged. Documentation should contain info about:

- Context of the original study
- Description of each file
- Description of each worksheet (ideally one table per worksheet)
- Revisions of the data
- Description of each variable in the files
Microsoft Access

Author: Fernando Rios & Dave Fearon
DCN Mentor: Dave Fearon

https://github.com/DataCurationNetwork/data-primers

Key Considerations:

What is the complexity of the database?
- Simple DBs (few tables, no forms, queries, macros) could be curated like a spreadsheet

As a base level for preservation:
- Keep original files + export tables to flat CSVs
- Screenshot the Relationships Diagram
- Run the Database Documenter and save the report alongside the DB
- Check for linked tables
- Other objects (SQL, forms, VB)?

Need help from creator
- Table relations, meaning of column names, how data is to be queried

Jupyter Notebooks

Authors: Daina Bouquin, Matthew Benzing, Sophie Hou & Lee Wilson

DCN Mentor: Susan Borda

https://github.com/DataCurationNetwork/data-primers

Code is Not Data

Jupyter notebooks contain code, incorporate data, and require different considerations

Different metadata for different situations

- Minimal deposit
- Runnable deposit
- Comprehensive deposit

Consider repository suitability

Wordpress

Author: Heather James
DCN Mentor: Lisa Johnston

https://github.com/DataCurationNetwork/data-primers

Check/Understand:

❖ Who runs the export? [Media files exported separately.]
❖ Will there be screen captures or archive-it.org scans to accompany?
❖ Mutual expectations for functionality of the site after deposit (both deposited version and live site)?
   ➢ Export All as XML doc; Export media library as .tar
➢ What metadata is there for media files?
Primer Topic Preview - Workshop at Johns Hopkins University

- Atlas.ti
- Confocal microscopy
- GeoJSON
- Google Docs
- Lidar Point Clouds
- NVivo (Note: Internal/DCN authored)
- PDF
- R
- .STL files
- Tableau
- Text/character encoding (Note: Internal/DCN authored)
Thanks again to everyone involved in this project!

Ho Jung Yoo - University of California San Diego
Sandra Sawchuk - Mount Saint Vincent University
Greg Janée - University of California Santa Barbara
Fernando Rios - University of Arizona
Daina Bouquin - Harvard University
Matthew Benzing - Miami University
Sophie Hou - University of Michigan
Lee Wilson - Portage Network
Andrew Battista - New York University
Tom Brittnacher - University of California Santa Barbara
Zenobie Garrett - University of Oklahoma
Carrie Pirmann - Bucknell University
Joshua Dull - Yale University
Sai Deng - University of Central Florida
Shahira Khair - University of Victoria
Jeanine Finn - Claremont Colleges
Heather James - Marquette University
Amanda Wittmire - Stanford University

The first Specialized Data Curation Workshop was co-located at the DLF Forum on October 17-18, 2018. Data curation primers created by this group are published to Github.
Thanks again to everyone involved in this project!

The second Specialized Data Curation Workshop was located at Johns Hopkins University on April 17-18, 2019. Data curation primers are in progress for this group.

- Susan Ivey - North Carolina State University
- Amy Koshoffer - University of Cincinnati
- Gretchen Sneff - Temple University
- Huajin Wang - Carnegie Mellon University
- Reina Murray - Johns Hopkins University
- Rachel Starry - University at Buffalo
- Nadia Dixon - City of Somerville Archives
- Genevieve Milliken - Pratt Institute
- Keshav Mukunda - Simon Fraser University
- Doug Joubert - National Institutes of Health
- Elizabeth Blackwood - Hillwood Estate Museum
- James Sobczak - University of Miami
- Tim Norris - University of Miami
- Kat Koziar - University of California, Riverside
- Lynda Kellam - Cornell University
- Standa Pejša - Purdue University
- Gin Corden - ICPSR
- Peace Ossom-Williamson - University of Texas, Arlington
- Nicole Contaxis - New York University
- Margaret Lam - George Mason University
- Adam Kriesberg - University of Maryland
- Seth Erickson - Pennsylvania State University
- Margarita Corral - Brandeis University
Thanks again to everyone involved in this project!

The Data Curation Network - IMLS Subgroup

Cynthia Hudson Vitale - Pennsylvania State University
Hannah Hadley - Pennsylvania State University
Lisa Johnston - University of Minnesota
Wendy Kozlowski - Cornell University
Dave Fearon - Johns Hopkins University
Mara Blake - Johns Hopkins University
Susan Borda - University of Michigan
Jake Carlson - University of Michigan
Jennifer Moore - Washington University in St.Louis
Sophia Lafferty-Hess - Duke University
Joel Herndon - Duke University
Jenn Darragh - Duke University
Share your expertise

Community Authored Data Curation Primers

https://github.com/DataCurationNetwork/data-primers

Get involved!

Contribute to these community resources via Github

Thanks to the DataOne Community for making this presentation possible
We are pleased to answer questions

https://sites.psu.edu/dcnworkshops/

https://datacurationnetwork.org