Data Curation Primers: Expanding the Data Curation Toolkit

Presented to the DataOne community 9-10-2019

Presented by

Lisa R. Johnston
Research Data Management Curation Lead
Principal Investigator, DCN
University of Minnesota

Cynthia Hudson Vitale
Head, Research Informatics & Publishing
Principal Investigator, DCN Education
Pennsylvania State University

Hannah Hadley
Project Manager
DCN Workshops 2018-2020
Research Informatics & Publishing
Pennsylvania State University

The Data Curation Network: Radical Collaboration

Lisa Johnston

Data Curation Network All Hands Meeting 2019

DATA CURATION NETWORK

Alfred P. Sloan
FOUNDATION

DATA
CURATION
NETWORK

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

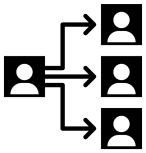
- We generate so much data!
- Data are messy (lack context!)
- Digital file formats are constantly at risk
- Most data never leaves their author's laptop ⇒ benign neglect
- Sharing is good (funders, publishers, disciplines)

Curation makes data sharing betterrrrr...

- Reuse (FAIR)
- Reproducibility
- Retractions (avoids them)
- Reputation (trust, transparency)
- References (citations)

DATA CURATION NETWORK

Vision for the Data Curation Network



- Provide expert data curation services for network partners

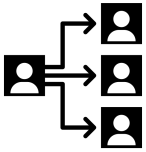
- Offer professional development opportunities for an emerging data curator professional community

- Create and openly share data curation best practices

- Demonstrate that curated datasets are measurably of greater reuse value than non-curated data

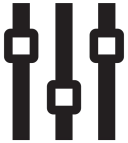
- Expand into a sustainable entity that grows beyond our initial partner institutions

DATA CURATION NETWORK



- Provide expert data curation services for network partners → **DCN Curation**

- Offer professional development opportunities for an emerging data curator professional community → **DCN Education**



- Create and openly share data curation best practices → **DCN Primers**

- Demonstrate that curated datasets are measurably of greater reuse value than non-curated data → **DCN R&D**

- Expand into a sustainable entity that grows beyond our initial partner institutions → **DCN Sustainability**

DATA CURATION NETWORK

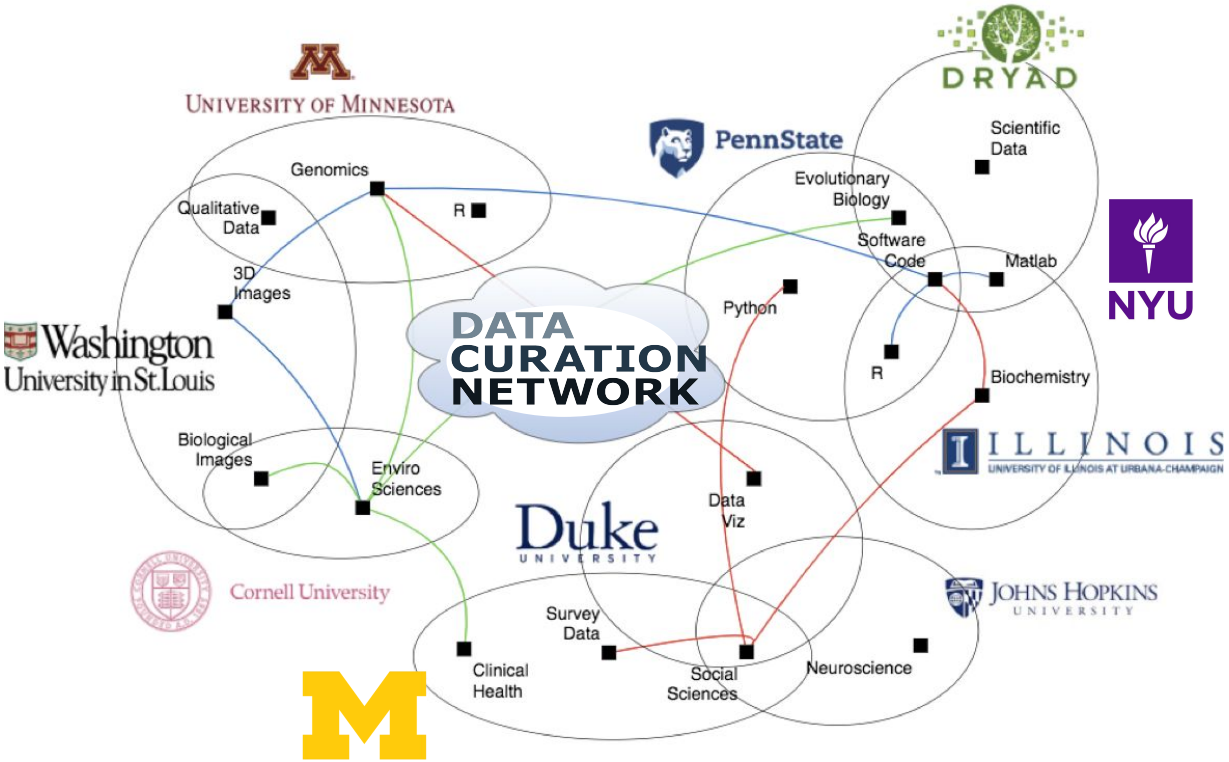
Value of Curation

- 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 ir

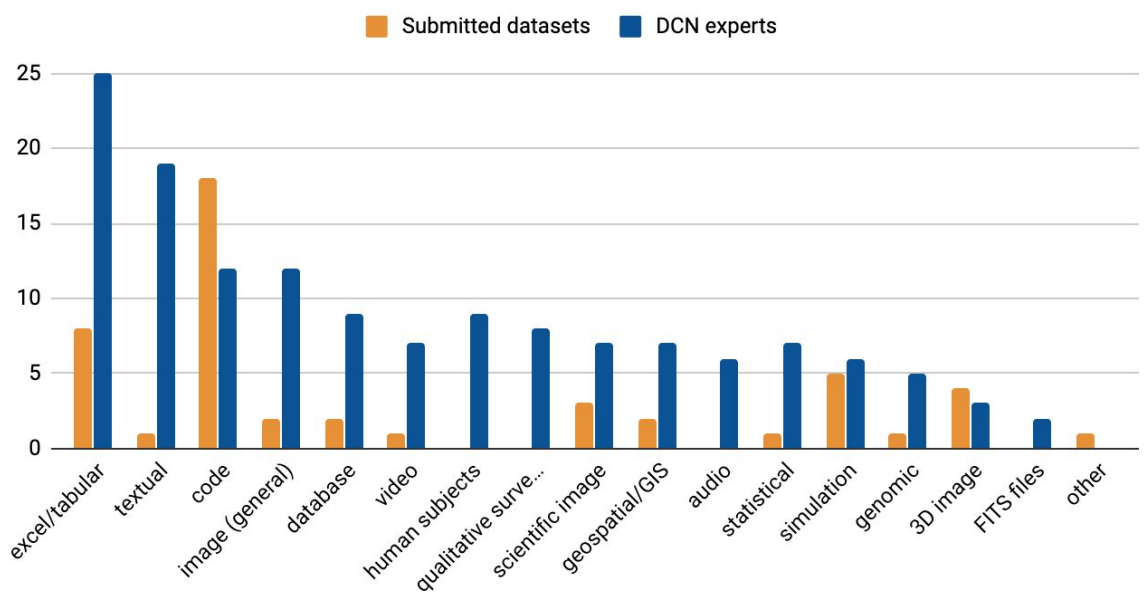
- **C** **Check** data files and read documentation
- **U** **Understand** the data (try to), if not…
- **R** **Request** missing information or changes
- **A** **Augment** the submission with metadata for findability
- **T** **Transform** file formats for reuse and long-term preservation
- **E** **Evaluate** and rate the overall submission for FAIRness.

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

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

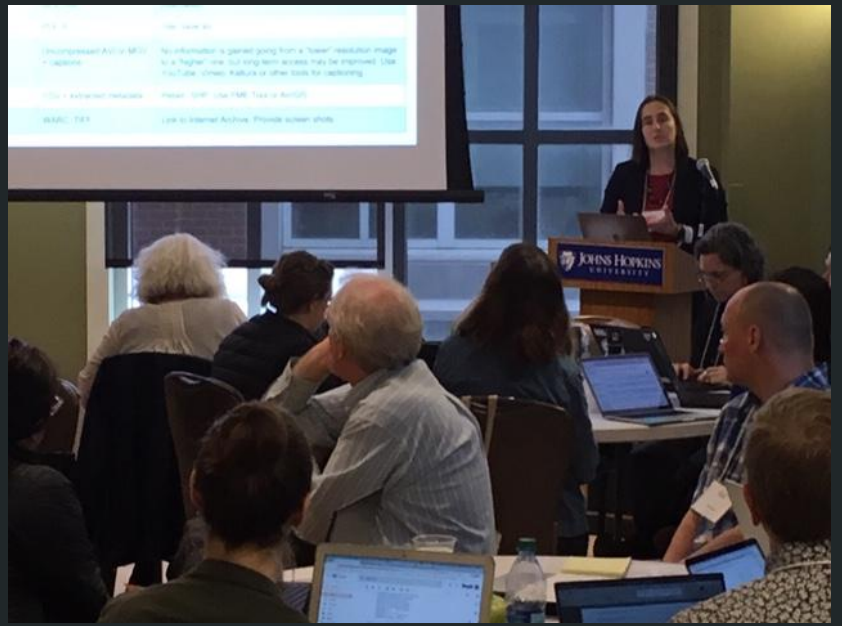
DCN Growth and Sustainability

- 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



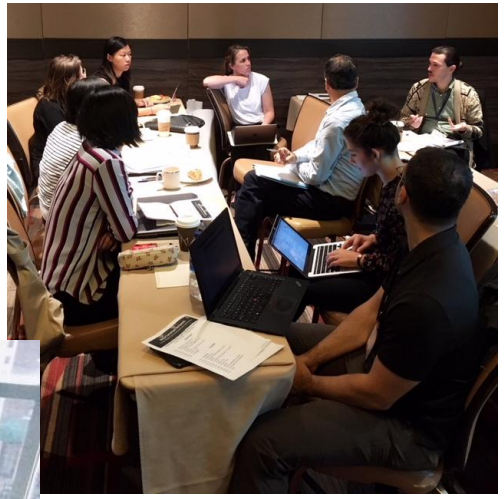
Enhancing Expertise throughout the Broader Community

Cynthia Hudson Vitale



Specialized Data Curation Workshop @JHU 2019

DCN Education



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

Data Curation Network IMLS Subgroup

Cornell University

Duke UNIVERSITY

University of Minnesota

PennState

UNIVERSITY OF MICHIGAN

JOHNS HOPKINS UNIVERSITY

Washington University in St. Louis

Learning Outcomes

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.

INSTITUTE of Museum and Library SERVICES

DATA CURATION NETWORK

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
10:15	Review Day 1
10:30	Curation Deep Dive #3: T Step
11:30	Lunch
12:15	Curation Deep Dive #3: E & D Step
1:15	Primer Time 2
2:00	Group feedback on primers
2:15	Wrap up
2:30	Everyone Disperses

Check files
Understand or try to
Request missing information
Augment the submission
Transform the format
Evaluate for FAIRness
Document throughout

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

www.datacurationnetwork.org

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

Hands-on data curation activities

- Survey Data
- Tabular Data
- Code
- Image Data
- Geospatial Data

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

- Survey Data
- Tabular Data
- Code
- Image Data
- Geospatial Data

EVALUATE Step

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

* 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

DATA CURATION NETWORK

Total # of WS2 Applicants	56
Total # of Canadian Applicants	3
Total # of African Applicants	4
Total # of US Applicants	49
Total # of US student applicants	2
Total # of US government applicants	4
Total # of US archives applicants	2
Total # of US museum applicants	2
Total # of other US applicants	1
Total # of US library applicants	38
Total # of R1 (US)	32
Total # of R2 (US)	3
Total # of SF (US)	2
Total # of B/A&S (US)	1

International Interest

Interest by other agency types and/or professions

Idea of the environment

DATA CURATION NETWORK

Total # of WS3 Applicants	59
Total # of Canadian Applicants	1
Total # of African Applicants	3
Total # of African non-profit applicants	2
Total # of US Applicants	55
Total # of US student applicants	1
Total # of US non-profit applicants	1
Total # of US grant funded applicants	1
Total # of US archives applicants	1
Total # of US library applicants	51
Total # of R1 (US)	35
Total # of R2 (US)	4
Total # of D/PU (US)	2
Total # of M (US)	5
Total # of B (US)	4
Total # of SF (US)	1

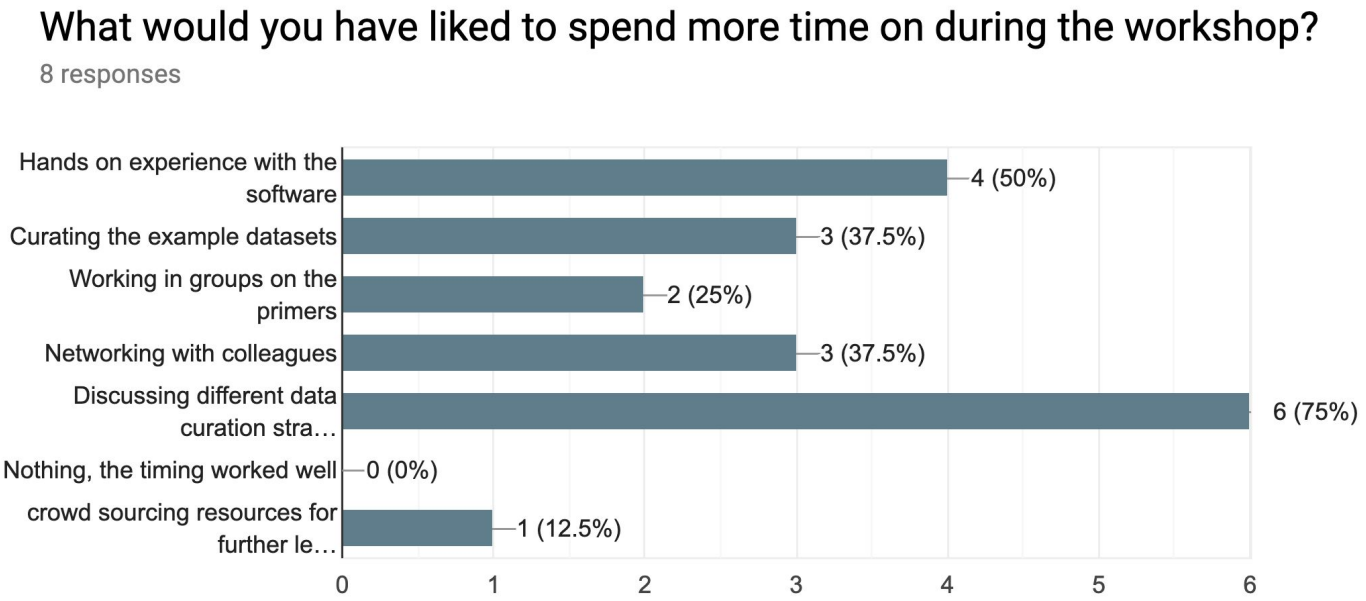
International Interest

Interest by other agency types and/or professions

Idea of the environment

DATA CURATION NETWORK

Attendee Feedback



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 NETWORK

Data Curation Primers are concise, actionable resources meant to assist data curation in adding value to a dataset.

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

DataCurationNetwork / data-primers

Unwatch 2 | Star 7 | Fork 2

Code | Issues 1 | Pull requests 0 | Projects 0 | Wiki | Security | Insights | Settings

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

Find file | Copy path

 cynhudson Add files via upload 566c52a on May 30

1 contributor

253 lines (186 sloc) | 17.2 KB

Raw | Blame | History

DATA CURATION NETWORK

Jupyter Notebooks: A Primer for Data Curators

Participants:

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

Example of Table Components

Primer Template
Overview

Topic	Description
File Extension	
MIME Type	
Structure	
Versions	
Primary fields or areas of use	
Source and affiliation	
Metadata standards	
Key questions for curation review	
Tools for curation review	
Date Created	
Created by	
Date updated and summary of changes made	

Primer Template

Format overview

Topic	Description
File Extension	.gdb
MIME type	
Structure	
Versions	
Primary fields or areas of use	Any field that makes use of geographic information systems (GIS) in which the primary GIS program used is ESRI's ArcGIS. Example fields include archaeology, ecology, geology, urban planning, etc.
Source and affiliation	Geodatabases are a proprietary file format developed and managed by ESRI.
Metadata standards	ISO19115, ISO19110, FGDC CSDGM content standards; .xml format (ISO19139, FGDC, Geoblacklight schema)
Tools for curation review	ArcGIS Desktop (ArcMap, ArcCatalog), ArcGIS Pro, QGIS
Date created	February 4, 2019
Created by	Andrew Battista, Tom Brittnacher, Zenobie Garrett, Jennifer Moore, Carrie Pirmann
Date updated and summary of changes made	February 4, 2019

Geodatabases Primer

Example Contents

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

Primer Template

Table of Contents

Geodatabases Primer

Our Peer Review Process

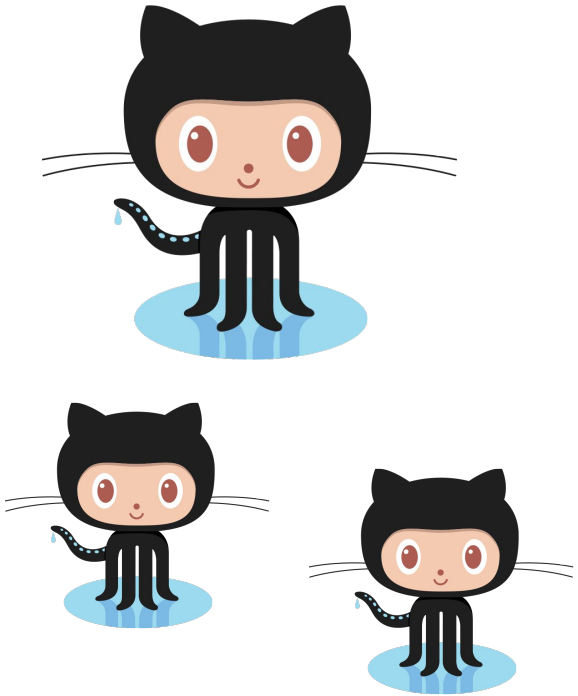
Primers are formally peer reviewed half-way through the six month process, or additionally if needed.

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

UNIVERSITY OF MINNESOTA

DRUM

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

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

DATA CURATION NETWORK

DataCurationNetwork / data-primers

Unwatch 2 | Star 7 | Fork 2

Code | Issues 1 | Pull requests 0 | Projects 0 | Wiki | Security | Insights | Settings

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

Find file | Copy path

 cynhudson Add files via upload 566c52a on May 30

1 contributor

253 lines (186 sloc) | 17.2 KB

Raw | Blame | History

DATA CURATION NETWORK

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

DATA CURATION NETWORK

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

Data Curation Network. (2019). RDAP Primerpalooza: Introducing Data Curation Primers [Geodatabases Primer slides by Andrew Battista, Tom Brittnacher, Zenobie Garrett, Jennifer Moore & Carrie Pirmann]. Retrieved from: https://vimeo.com/350235467

netCDF

Author: Sophie Hou

DCN Mentors: Jake Carlson & Susan Borda

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

DATA CURATION NETWORK

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

Data Curation Network. (2019). RDAP Primerpalooza: Introducing Data Curation Primers [netCDF Primer slides by Sophie Hou]. Retrieved from: https://vimeo.com/350235467

SPSS

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

DCN Mentor: Sophia Lafferty-Hess

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

DATA CURATION NETWORK

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

Data Curation Network. (2019). RDAP Primerpalooza: Introducing Data Curation Primers [SPSS Primer slides by Joshua Dull, Sai Deng, Shahira Khair & Jeanine Finn]. Retrieved from: https://vimeo.com/350235467

Microsoft Excel

Authors: Ho Jung Yoo, Sandra Sawchuk & Greg Janée

DCN Mentor: Wendy Kozlowski

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

DATA CURATION NETWORK

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

Data Curation Network. (2019). RDAP Primerpalooza: Introducing Data Curation Primers [Microsoft Excel Primer slides by Ho Jung Yoo, Sandra Sawchuk & Greg Janée]. Retrieved from: https://vimeo.com/350235467

Microsoft Access

Author: Fernando Rios & Dave Fearon

DCN Mentor: Dave Fearon

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

DATA CURATION NETWORK

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

Data Curation Network. (2019). RDAP Primerpalooza: Introducing Data Curation Primers [Microsoft Access Primer slides by Fernando Rios & Dave Fearon]. Retrieved from: https://vimeo.com/350235467

Jupyter Notebooks

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

DCN Mentor: Susan Borda

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

DATA CURATION NETWORK

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

Data Curation Network. (2019). RDAP Primerpalooza: Introducing Data Curation Primers [Jupyter Notebooks Primer slides by Daina Bouquin, Matthew Benzing, Sophie Hou & Lee Wilson]. Retrieved from: https://vimeo.com/350235467

Wordpress

Author: Heather James

DCN Mentor: Lisa Johnston

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

DATA CURATION NETWORK

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?

Data Curation Network. (2019). RDAP Primerpalooza: Introducing Data Curation Primers [Wordpress Primer slides by Heather James]. Retrieved from: https://vimeo.com/350235467

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)

DATA CURATION NETWORK

Thanks again to everyone involved in this project!

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.

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

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

DATA CURATION NETWORK

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

DATA CURATION NETWORK

We are pleased to answer questions

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

https://datacurationnetwork.org