NDS/MBDH 2018

Launching the Data Curation Network

Lisa Johnston University of Minnesota
Jake Carlson University of Michigan
Cynthia Hudson-Vitale Penn State Univ.
Heidi Imker University of Illinois
Wendy Kozlowski Cornell University
Robert Olendorf Penn State University
Claire Stewart University of Minnesota
Mara Blake Johns Hopkins University
Joel Herndon Duke University
Elizabeth Hull Dryad Data Repository
Timothy M. McGeary Duke University

7-11-2018

Data Curation Network
http://DataCurationNetwork.org
Rise of the Data Sharing Culture

Researchers are increasingly required/incentivised to share data
● Funder data sharing mandates
● Journal data sharing policies
● Disciplinary practices → emphasis on transparency and reproducibility

But! It’s not enough to just share the files, **well-curated data** are more valuable!

**Goal of data curation** ⇒ Ingest and maintain (trusted digital repositories) in ways that make it findable, accessible, interoperable and reusable.
The Data Repository for University of Minnesota (DRUM)

DRUM is a publicly available collection of digital research data generated by U of M researchers, students, and staff. Anyone can search and download the data housed in the repository, instantly or by request.

The Data Repository accepts submissions from University affiliates for digital archiving and access. Learn more about depositing to the Data Repository and other services to manage your data.

How to Upload
1. Prepare Data
   Data should be free of identifying or sensitive information and include adequate documentation. Not sure? Contact us for help!
2. Upload

Features
- Flexible Access Options
  Choose to make your data immediately accessible to everyone, or moderate access to your data upon request.
- Meet Grant Requirements

Our Services
Data Management Plan Assistance
We offer personalized assistance for drafting your next grant’s Data Management Plan. Contact us for assistance during your planning process.

Metadata Consultation
Link Lists for Websites Reporting Information on Hurricane Sandy from 2003 to 2012


Published Date
2018-06-20

Author Contact
Weber, Matthew S. (msw@umn.edu)

Type
Dataset

Observational Data
Other Dataset

Abstract
Data contains hyperlinks that existed between websites reporting information on Superstorm Sandy from 2003 – 2012. The data tracks 20,013,455 unique URLs.

Referenced by

License
Attribution-NonCommercial-ShareAlike 3.0 United States

Suggested Citation

Persistent link to this item
http://hdl.handle.net/11299/197957

File View/Open
View Usage Statistics

View/Download file

NSFIA_SANDY_2003_2012-all.tar
Hyperlink Data from Superstorm Sandy Websites
4.009Gb
application/x-tar

By using these files, users agree to the Terms of Use. Content distributed via the University of Minnesota's Digital Conservancy may be subject to additional license and use restrictions applied by the depositor.
<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.296</td>
<td>-0.55</td>
<td>-0.243</td>
<td>-3.04</td>
<td>-2.885</td>
<td>0.071</td>
<td>1.847</td>
<td>-0.521</td>
<td>0.107</td>
<td>-3.528</td>
<td>-3.089</td>
<td>2.45</td>
<td></td>
</tr>
<tr>
<td>0.354</td>
<td>-0.487</td>
<td>-0.193</td>
<td>-2.974</td>
<td>-2.933</td>
<td>-0.029</td>
<td>1.951</td>
<td>-0.506</td>
<td>0.061</td>
<td>-3.643</td>
<td>-3.162</td>
<td>2.131</td>
<td>-1</td>
</tr>
<tr>
<td>0.438</td>
<td>-0.449</td>
<td>-0.089</td>
<td>-2.979</td>
<td>-2.901</td>
<td>-0.09</td>
<td>2.051</td>
<td>-0.501</td>
<td>0.005</td>
<td>-3.647</td>
<td>-3.34</td>
<td>1.713</td>
<td>-1</td>
</tr>
<tr>
<td>0.519</td>
<td>-0.431</td>
<td>-0.018</td>
<td>-3.042</td>
<td>-2.831</td>
<td>-0.13</td>
<td>2.107</td>
<td>-0.452</td>
<td>0.027</td>
<td>-3.616</td>
<td>-3.42</td>
<td>1.322</td>
<td>-1</td>
</tr>
<tr>
<td>0.696</td>
<td>-0.37</td>
<td>0.023</td>
<td>-3.065</td>
<td>-2.832</td>
<td>-0.187</td>
<td>2.202</td>
<td>-0.415</td>
<td>0.072</td>
<td>-3.648</td>
<td>-3.504</td>
<td>0.831</td>
<td>-1</td>
</tr>
<tr>
<td>0.999</td>
<td>-0.332</td>
<td>0.083</td>
<td>-3.089</td>
<td>-2.815</td>
<td>-0.233</td>
<td>2.314</td>
<td>-0.342</td>
<td>0.119</td>
<td>-3.603</td>
<td>-3.648</td>
<td>0.38</td>
<td>-1</td>
</tr>
<tr>
<td>1.188</td>
<td>-0.295</td>
<td>0.171</td>
<td>-3.08</td>
<td>-2.753</td>
<td>-0.295</td>
<td>2.431</td>
<td>-0.197</td>
<td>0.186</td>
<td>-3.598</td>
<td>-3.619</td>
<td>-0.07</td>
<td>-1</td>
</tr>
<tr>
<td>1.503</td>
<td>-0.284</td>
<td>0.279</td>
<td>-3.129</td>
<td>-2.746</td>
<td>-0.363</td>
<td>2.52</td>
<td>-0.116</td>
<td>-0.298</td>
<td>-3.487</td>
<td>-3.5</td>
<td>-0.608</td>
<td>-1</td>
</tr>
<tr>
<td>1.826</td>
<td>-0.288</td>
<td>0.36</td>
<td>-3.183</td>
<td>-2.743</td>
<td>-0.496</td>
<td>2.59</td>
<td>-0.012</td>
<td>-0.316</td>
<td>-3.318</td>
<td>-3.456</td>
<td>-0.989</td>
<td>-1</td>
</tr>
<tr>
<td>2.153</td>
<td>-0.289</td>
<td>0.369</td>
<td>-3.162</td>
<td>-2.632</td>
<td>-0.615</td>
<td>2.653</td>
<td>0.197</td>
<td>-0.345</td>
<td>-3.249</td>
<td>-3.388</td>
<td>-1.33</td>
<td>0</td>
</tr>
<tr>
<td>2.59</td>
<td>-0.244</td>
<td>0.359</td>
<td>-3.205</td>
<td>-2.51</td>
<td>-0.761</td>
<td>2.761</td>
<td>0.412</td>
<td>-0.416</td>
<td>-3.204</td>
<td>-3.296</td>
<td>-1.58</td>
<td>0</td>
</tr>
<tr>
<td>2.97</td>
<td>-0.196</td>
<td>0.319</td>
<td>-3.218</td>
<td>-2.463</td>
<td>-0.944</td>
<td>2.933</td>
<td>0.643</td>
<td>-0.421</td>
<td>-3.143</td>
<td>-3.089</td>
<td>-1.746</td>
<td>0</td>
</tr>
<tr>
<td>3.269</td>
<td>-0.222</td>
<td>0.297</td>
<td>-3.148</td>
<td>-2.454</td>
<td>-1.045</td>
<td>3.051</td>
<td>0.904</td>
<td>-0.356</td>
<td>-2.983</td>
<td>-2.829</td>
<td>-1.813</td>
<td>0</td>
</tr>
<tr>
<td>3.512</td>
<td>-0.266</td>
<td>0.274</td>
<td>-3.157</td>
<td>-2.429</td>
<td>-1.147</td>
<td>3.119</td>
<td>1.116</td>
<td>-0.286</td>
<td>-2.783</td>
<td>-2.595</td>
<td>-1.927</td>
<td>0</td>
</tr>
<tr>
<td>3.684</td>
<td>-0.271</td>
<td>0.289</td>
<td>-3.214</td>
<td>-2.396</td>
<td>-1.255</td>
<td>3.052</td>
<td>1.222</td>
<td>-0.227</td>
<td>-2.627</td>
<td>-2.292</td>
<td>-2.081</td>
<td>0</td>
</tr>
<tr>
<td>3.824</td>
<td>-0.275</td>
<td>0.233</td>
<td>-3.289</td>
<td>-2.4</td>
<td>-1.626</td>
<td>2.996</td>
<td>1.39</td>
<td>-0.16</td>
<td>-2.475</td>
<td>-2.019</td>
<td>-2.286</td>
<td>0</td>
</tr>
<tr>
<td>3.889</td>
<td>-0.294</td>
<td>0.186</td>
<td>-3.295</td>
<td>-2.303</td>
<td>-1.306</td>
<td>2.961</td>
<td>1.545</td>
<td>-0.083</td>
<td>-2.293</td>
<td>-1.825</td>
<td>-2.461</td>
<td>0</td>
</tr>
<tr>
<td>3.896</td>
<td>-0.295</td>
<td>0.158</td>
<td>-3.289</td>
<td>-2.266</td>
<td>-1.383</td>
<td>2.93</td>
<td>1.645</td>
<td>-0.095</td>
<td>-2.195</td>
<td>-1.606</td>
<td>-2.573</td>
<td>0</td>
</tr>
<tr>
<td>3.838</td>
<td>-0.283</td>
<td>0.152</td>
<td>-3.286</td>
<td>-2.273</td>
<td>-1.352</td>
<td>2.876</td>
<td>1.615</td>
<td>-0.074</td>
<td>-2.086</td>
<td>-1.385</td>
<td>-2.672</td>
<td>0</td>
</tr>
<tr>
<td>3.712</td>
<td>-0.338</td>
<td>0.139</td>
<td>-3.328</td>
<td>-2.23</td>
<td>-1.302</td>
<td>2.778</td>
<td>1.637</td>
<td>-0.007</td>
<td>-1.971</td>
<td>-1.328</td>
<td>-2.759</td>
<td>0</td>
</tr>
<tr>
<td>3.526</td>
<td>-0.363</td>
<td>0.125</td>
<td>-3.387</td>
<td>-2.198</td>
<td>-1.275</td>
<td>2.604</td>
<td>1.624</td>
<td>-0.019</td>
<td>-1.814</td>
<td>-1.35</td>
<td>-2.817</td>
<td>0</td>
</tr>
</tbody>
</table>
Challenges for IR Data Curation Services

- How to scale local data curation services across all disciplines?
- How many data curation experts are needed?
  - Types: GIS, spreadsheet/tabular, statistical/survey, software code, video/audio…
  - Disciplines: genomic sequence, chemical spectra, bioinformatics…
- Are there ways to more efficiently curate rare or infrequently generated data types?
- Might our institutions specialize in curation skills? Represent our academic expertise?
The Data Curation Network (DCN) addresses this challenge by collaboratively sharing data curation staff across a network of partner institutions and data repositories.
The Data Curation Network (DCN) 3-year implementation phase launched June 2018 with funding from the Alfred P Sloan Foundation.
The Data Curation Network (DCN) serves as the “human layer” in the data repository stack that provides specialized dataset curation and professional development training for an emerging data curator community.
Katie Wilson
Scientific Data Curator

Valerie Collins
DRUM coordinator

Shanda Hunt
Public Health Data Curator

Ben Wiggins
Digital Arts and Humanities Curator

Melinda Kernik
GIS/Spatial Data Curator

Alicia Hofelich Mohr
College of Liberal Arts Data Management Specialist

Katie Wilson
Scientific Data Curator

Valerie Collins
DRUM coordinator

Shanda Hunt
Public Health Data Curator

Ben Wiggins
Digital Arts and Humanities Curator

Melinda Kernik
GIS/Spatial Data Curator

Alicia Hofelich Mohr
College of Liberal Arts Data Management Specialist

Katie Wilson
Scientific Data Curator

Valerie Collins
DRUM coordinator

Shanda Hunt
Public Health Data Curator

Ben Wiggins
Digital Arts and Humanities Curator

Melinda Kernik
GIS/Spatial Data Curator

Alicia Hofelich Mohr
College of Liberal Arts Data Management Specialist

Katie Wilson
Scientific Data Curator

Valerie Collins
DRUM coordinator

Shanda Hunt
Public Health Data Curator

Ben Wiggins
Digital Arts and Humanities Curator

Melinda Kernik
GIS/Spatial Data Curator

Alicia Hofelich Mohr
College of Liberal Arts Data Management Specialist

Katie Wilson
Scientific Data Curator

Valerie Collins
DRUM coordinator

Shanda Hunt
Public Health Data Curator

Ben Wiggins
Digital Arts and Humanities Curator

Melinda Kernik
GIS/Spatial Data Curator

Alicia Hofelich Mohr
College of Liberal Arts Data Management Specialist

Katie Wilson
Scientific Data Curator

Valerie Collins
DRUM coordinator

Shanda Hunt
Public Health Data Curator

Ben Wiggins
Digital Arts and Humanities Curator

Melinda Kernik
GIS/Spatial Data Curator

Alicia Hofelich Mohr
College of Liberal Arts Data Management Specialist
Jake Carlson
Research Data Services Manager
University of Michigan

Susan Borda
Data Workflows Specialist

Rachel Woodbrook
Data Curation Librarian
Robert Olendorf, PhD
Science Data Librarian
Pennsylvania State University

Cynthia Hudson-Vitale
Head, Digital Scholarship and Data Services
Pennsylvania State University Libraries

John Russell
Associate Director Center for Humanities and Information

Nathan Piekielek
Geospatial Services Librarian
Elizabeth Hull
Operations Manager
Dryad Digital Repository

Erin Clary
Senior Curator

Debra Fagan
Curation and Technical Specialist

Data Curation Network
Heidi Imker
Director, Research Data Service
University of Illinois at Urbana-Champaign

Hoa Luong
Research Data Specialist
University of Illinois at Urbana-Champaign

Ashley Hetrick
Research Data Specialist
University of Illinois at Urbana-Champaign
DCN Workflow

Uncurated Data
Presenting scale and expertise challenges to individual institutions

Ingest → Appraise and Select → DCN → Facilitate Access → Preserve Long-Term

Curated Data
at scale and with great efficiency through shared Data Curation Network
DCN Workflow

- Researchers deposit like normal
- Researchers deposit like normal
- DCN functions as a microservice layer (the “human layer in your repository stack”)
DCN Workflow

- Researchers deposit like normal
- DCN functions as a microservice layer (the “human layer in your repository stack”)
- Local institution maintain full responsibility for all technical functionality (eg. storage) and authority for local decision-making (what to ingest, how long to retain, etc.)
DCN Workflow

- Researchers deposit like normal
- DCN functions as a microservice layer (the “human layer in your repository stack”)
- Local institution maintain full responsibility for all technical functionality (eg. storage) and authority for local decision-making (what to ingest, how long to retain, etc.)
- Seamlessly integrates into all repository systems (Samvera, Fedora, DSpace, Dataverse, etc.)
DCN Workflow

Uncurated Data
Presenting scale and expertise challenges to individual institutions

Ingest → Appraise and Select → DCN → Facilitate Access → Preserve Long-Term → Curated Data
at scale and with great efficiency through shared Data Curation Network

Data Curation Network

Review → Assign → CURATE → Mediate → Approve

DCN Coordinator Workflow

Check files and metadata → Understand and run files → Request missing information → Augment metadata → Transform file formats → Evaluate for FAIRness

DCN Curator Workflow
CURATE Steps in DCN Workflow

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

- **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 FAIRness.
DCN CURATE Steps

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

<table>
<thead>
<tr>
<th>CURATE Actions</th>
<th>Curation Checklist</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Check</strong> 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 many errors</td>
</tr>
<tr>
<td></td>
<td>□ Metadata quality is rich, complete</td>
</tr>
<tr>
<td></td>
<td>□ Metadata has issues</td>
</tr>
<tr>
<td></td>
<td>□ Documentation Type (e.g., Readme / Codebook / Data Dictionary)</td>
</tr>
<tr>
<td></td>
<td>□ Description / Details</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) | **Evaluate** and rate the overall data record for FAIRness.²

- □ Check for quality assurance and usability issues such as missing...

- Varies based on file formats and an example.

- □ 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 like Google.

- Accessible -
  □ Retrieveable 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 (OA1 feed).

- Reusable -
  □ Data include sufficient metadata about the data characteristics to reuse without the direct assistance of the author,
  □ Clear indicators of who created, owns, and stewards the data,
  □ Data are released with clear data usage terms (e.g., a CC License).

¹ Format Recommendations, [http://guides.library.cornell.edu/eecommons/formats](http://guides.library.cornell.edu/eecommons/formats)
² Rubric evaluating the FAIR principles are based on the scoring matrix by Dunning, de Smaele, & Böhmer (2017).
Specialized Curation Training (2018-2020)

DATA CURATION WORKSHOP
SLIDES AND HANDOUTS

DATE: DECEMBER 11 & 12, 2017
TWEET: #DCW2017
LOCATION: WASHINGTON UNIVERSITY IN ST. LOUIS, MCMILLAN HALL, ST. LOUIS, MISSOURI

DESCRIPTION:
This free, 1.5 day workshop is open to all library staff and data professionals who are interested in data curation.

Participants will learn practical, hands-on treatments for data curation based on the Data Curation Network CURATE model.

- 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.

ATTENDEES WILL COME AWAY WITH:
1. A customized, implementable plan to enhance data curation activities at your local institution or organization,
2. Stakeholder focused tailoring points related to the value of data curation activities,
3. An in-depth understanding of specialized data curation practices in various disciplines, data types, and formats.
DCN Implementation (2018-2021)

Assessment Plan (two-prong)

<table>
<thead>
<tr>
<th>Is a networked approach to curating research data more efficient?</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Number of datasets</td>
</tr>
<tr>
<td>- Frequency (high-volume time periods, etc.)</td>
</tr>
<tr>
<td>- Variety (data file formats; range of disciplines)</td>
</tr>
<tr>
<td>- Efficiency (time, costs)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Are curated data are more valuable?</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Track reuse indicators (download counts, citations, alt-metrics)</td>
</tr>
<tr>
<td>- Implement a DCN registry</td>
</tr>
<tr>
<td>- Apply badges and metadata to signal that data sets curated by the DCN are FAIR.</td>
</tr>
</tbody>
</table>
In Year 3, the DCN will begin transitioning to a self-sustaining service model where institutional and disciplinary partners contribute data curation staff and central operations costs are offset by users of the Network.
Data Curation-as-Service

Alliance Model
Institutional Partners contribute staff and fund central coordinator

Data Curation Network

Institutional Partner
Local Representative
Local Curator(s)
roles distributed among 1 or many people at the institution as needed

DCN Coordinator
matches expertise from the pool of local Curators
DCN Curator
roles distributed among 1 or many people at the institution as needed

Individual and Institutional Users gain access to the Network on a fee-for-service basis
## Sustainability and Expansion (2020-)

<table>
<thead>
<tr>
<th>Stakeholder</th>
<th>Benefits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academic libraries with existing data curation services</td>
<td>Gain access to data curation expertise in more disciplines/formats than locally available</td>
</tr>
<tr>
<td>Academic libraries with limited to no resources for data curation services</td>
<td>Are able to provide critical new data curation services when local resources are limited (without needing to hire);</td>
</tr>
<tr>
<td>Disciplinary- and general-subject data repositories</td>
<td>Receive better, more valuable data submissions from DCN partner institutions and customers;</td>
</tr>
<tr>
<td></td>
<td>Have potential to partner with the DCN to expand the scope of curation support for new and/or less frequently encountered data types</td>
</tr>
</tbody>
</table>
# 6-year Roadmap toward Sustainability

*Transition from planning phase to sustaining phase*

<table>
<thead>
<tr>
<th>Year</th>
<th>Support</th>
<th>Timing</th>
<th>Phase</th>
<th>Partners</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Sloan Grant</td>
<td>2016-17</td>
<td>Planning</td>
<td>6 academic institutions</td>
</tr>
<tr>
<td>1</td>
<td>Grant Funded (Y1-Y2) transition to partnership model (Y3)</td>
<td>2018-19</td>
<td>Implementation</td>
<td>8 academic institutions and 2 disciplinary partners</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>2020-21</td>
<td>Transition</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td>Sustaining</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Curation-as-service (Y4-6)</td>
<td></td>
<td></td>
<td>Recruit new partners as use and demand dictate</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Mission: With a proven and appealing value-proposition, the Data Curation Network will expand into a sustainable entity that grows beyond our initial partner institutions.
Thanks!

https://DataCurationNetwork.org

Twitter #DataCurationNetwork