Developing & Scaling Research Data Management (RDM) & Curation Services

Tim McGeary
Associate University Librarian
Duke University

Sophia Lafferty-Hess
RDM Consultant
Duke University

Clare Stewart
Associate University Librarian
University of Minnesota

Jennifer Darragh
RDM Consultant
Duke University
The Duke Story
Reasons to Scale Up RDM and Curation

2011 - NSF requires Data Management Plans

2013 - OSTP memo requiring Federal agencies with more than $100M in R&D expenditures to develop plans to make the results of federally funded research freely available to the public—generally within one year of publication.

2016 - NSF and Dept of Energy begin requiring deposit of publications and require data to be made available at expense of research institutions
The History at Duke

2015 - NSF rejects Duke research proposal due to insufficient DMP, specifically due to lack of plans to deposit data

2015 - University Libraries, Trinity College of Arts & Sciences, and Office of Information Technology (OIT) each request significant funding increase for research data storage from Provost

2015 - Provost seeks white paper and charge for Faculty Working Group for Digital Research Data Services

2016 - Interdisciplinary faculty working group meets for 9 months and submits recommendation to Provost
Responsible Conduct of Research (RCR), students, staff, & faculty

Tools, Best Practices, Reproducibility, Sensitive Data, Funder Compliance

12 workshops, 435 attendees
57% Sciences and Engineering, 20% Social Sciences, 6% Health Sciences, 6% Humanities

Office of Research Support, Research Computing, School of Medicine, etc.
The data will be deposited into the Duke Digital Repository (DDR) - repository.duke.edu - an openly accessible preservation archive maintained by the Duke University Libraries. The DDR will assign appropriate metadata (Dublin Core) for discoverability and provide a Digital Object Identifier (DOI) for persistent access and unique identification of the data.
Lifecycle Services

Duke University
A research data service provided by Duke Libraries.

Data Workflow Design

<table>
<thead>
<tr>
<th>Name</th>
<th>Contributors</th>
<th>Modified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bergelson Lab</td>
<td>Dailey, Bergelson + 22</td>
<td>3 days ago</td>
</tr>
<tr>
<td>SEEDlings</td>
<td>Bergelson, Dailey + 4</td>
<td>3 days ago</td>
</tr>
<tr>
<td>Neural Subjective Value Representations across Age and Discount ... Seaman, Karrer + 7</td>
<td>5 days ago</td>
<td></td>
</tr>
<tr>
<td>A Walk-through of How to Set Up an OSF Project</td>
<td>Darragh, Lafferty-Hess + 1</td>
<td>24 days ago</td>
</tr>
<tr>
<td>OSF DEMO!</td>
<td>Tepper, Juarez + 1</td>
<td>25 days ago</td>
</tr>
<tr>
<td>OSF + TIER Demo - Public Version</td>
<td>Sedlins, Lafferty-Hess</td>
<td>a month ago</td>
</tr>
<tr>
<td>Dynamic FACES database</td>
<td>Holland, Ebner + 2</td>
<td>a month ago</td>
</tr>
</tbody>
</table>
Lifecycle Services

Sensitivity  Copyright/IP  Licensing  Formatting  Metadata

Data & Documentation Review

Icon by Christopher T. Howlett from Noun Project.
Lifecycle Services

Data Repository Support

https://library.duke.edu/ddr/research-data
Findable
Accessible
Interoperable
Reusable

Data Curation Pipeline @ Duke

Deposit

✓ Submit data, documentation, & metadata
✓ Select Creative Commons waiver/license
Data Curation Pipeline @ Duke

Deposit
- Submit data, documentation, & metadata
- Select Creative Commons waiver/license

Review
- Check documentation completeness
- Assess disclosure risks
- Assess formats
Data Curation Pipeline @ Duke

**Deposit**
- ✓ Submit data, documentation, & metadata
- ✓ Select Creative Commons waiver/license

**Review**
- ✓ Check documentation completeness
- ✓ Assess disclosure risks
- ✓ Assess formats

**Ingest**
- ✓ Transform files
- ✓ Arrange files
- ✓ Generate metadata
- ✓ Assign license
Data Curation Pipeline @ Duke

Deposit
- ✓ Submit data, documentation, & metadata
- ✓ Select Creative Commons waiver/license

Review
- ✓ Check documentation completeness
- ✓ Assess disclosure risks
- ✓ Assess formats

Ingest
- ✓ Transform files
- ✓ Arrange files
- ✓ Generate metadata
- ✓ Assign license

Publish
- ✓ Assign DOI
- ✓ Review & test
- ✓ Finalize administrative metadata
Lessons Learned

Value of looking towards peers for benchmarking and evaluation

Further enhancing the FAIRness of data will benefit from in-depth subject and data type expertise

Collaborating both internally and externally facilitated program development
Origins, campus policy, and engagement

Before Your Research Begins
- Schedule a data management plan (DMP) consultation (Request Form) or use one of the existing DMPs.
- Meet funding agency requirements for data and learn best practices for getting IRB approval for sharing data.
- See more tools for planning for data management.

During Your Research
- Attend workshops and explore online training resources on best practices for data management.
- Get help creating documentation and using metadata standards.
- Discover appropriate U of M services for data, such as data storage.

Methodology

The Data Curation Pilot project was completed over an eight-month period (May-December) in 2013. The project was implemented by the author through several phases, as illustrated in the visual roadmap in Figure 2. The actions taken in each of the five project phases are detailed in this section.

Figure 2: Visual Roadmap for the 2013 Data Curation Pilot Project
Origins, campus policy, and engagement

The Vice President for Research, the Vice President and Chief Information Officer, the Vice President for Health Sciences, and the University Librarian/Director of Libraries (depending on campus) are jointly responsible for periodically evaluating existing research data management solutions across the University and determining future research data management requirements, and sponsoring a Use Case Categorization Scheme (UCCS) Committee. The UCCS Committee is advisory and includes representation from the University research community (OVPF), the Office of the Vice President for System and Student Affairs, and the University's scholarly and professional organizations. The Committee will focus on the University’s data management needs, including compliance with existing and forthcoming regulations, and will make recommendations to the University for the adoption of data management policies and procedures. The Committee will develop and maintain a list of University data management requirements and recommendations, and will provide guidance on the implementation of data management policies and procedures. The Committee will also work with the Office of the Vice President for Research and the University's data management partners to develop and implement best practices for data management at the University.
Cross-Cultural Interaction and Migration: Retheorizing Greek Colonization in the 7th - 4th Centuries BC with the Exchange of Attic Figured Pottery

Faulkner-Gerity, Ivy (2018)

Published Date
2018-03-29

Author Contact
Faulkner-Gerity, Ivy (fau0078@umn.edu)

Abstract
This data is part of a doctoral dissertation study on the exchange of Attic figured pottery between Greek settlers and local populations in Western Europe and the Northern Black Sea in the 7th-4th centuries BC. This data comprises instances of Attic painted pottery in these regions with deposition context information as well as item descriptions and references. This data is also visualized geographically using ArcGIS Online (Esri), the results of which are shared via pdf and shapefiles. This data can be used to analyze inter- and intra-regional interaction and may form the foundation for expanded research on Greek migration in the Archaic period to other regions.

License
Attribution-NonCommercial-ShareAlike 3.0 United States

Predictors and benefits of microhabitat selection for offspring deposition in golden rocket frogs

Petito, Beth A; Bourne, Geoffrey P; Bee, Mark A (2018)

Published Date
2018-04-02

Author Contact
Bee, Mark A (mbee@umn.edu)

Abstract
This data set originates from a field study of habitat selection by parental frogs. In many tropical frogs, offspring development and survival potentially depend on microhabitat features associated with sites that parents select for oviposition and tadpole rearing. This study investigated the importance of microhabitat features in the selection of oviposition sites versus tadpole rearing sites, as well as in determining offspring survival, in the golden rocket frog, Aonomaloglossus beebii. Endemic to Guyana, this species exhibits biparental care and exclusively uses phytotelmata in bromeliads for oviposition and tadpole rearing. The data included here were used in model-based inference to evaluate evidence for the hypotheses that (1) parents prioritize different microhabitat features in selecting phytotelmata suitable for oviposition versus tadpole rearing and (2) microhabitat selection can adaptively promote offspring survival. The dataset includes description of breeding site, phytotelmata, habitat and mode of elevation of host plant.
Usage Statistics for Collection: Data Repository for U of M (DRUM)

Total Downloads (2014-Present)

Data Repository for U of M (DRUM)

85946
Data curation staff @ UMN

- Data Management and Curation initiative (now program) **lead**, 2013: Lisa Johnston
- **Humanities** data curator
  - 2013 CLIR two-year fellow digital humanities: Justin Schell
  - Continuous appointment librarian 2016: Director, Digital Arts Sciences and Humanities program (DASH), Ben Wiggins
- **Social sciences** data curator
  - 2015 CLIR two-year fellow social science data: Alice Motes
  - Continuous appointment librarian posting anticipated 2018
- **Health** data curator, new position 2015, Public Health Liaison and Data Curation Specialist: Shanda Hunt
- **Spatial data** analyst/curator, jointly funded with campus U-Spatial project, Melinda Kernik
- Digital Repositories **Archivist**, new position 2016: Valerie Collins
- **Scientific** data curator
  - Graduate student 50% time 2015-2016
  - Continuous appointment librarian 2017: Biosciences Librarian and Scientific Data Specialist, Katie Wilson

Also critical: repository developers/technologists, digital preservation librarian, University Archivist, Research Services Coordinators and liaisons, colleagues in collegiate technology units
Campus uptake, engagement, evolution

WELCOME!

Research Data Boot Camp
Spring Break 2018

Grab a name tag and breakfast!

Slides and materials are shared: z.umn.edu/rdcamp2018
New strategic plan for RDS in draft
Well curated data are more valuable.
The skills and expertise required to curate data cannot be fully automated nor reasonably be provided by a few experts siloed at single institutions.
The Data Curation Network (DCN) addresses this challenge by collaboratively sharing data curation staff across a network of partner institutions and data repositories.
Steps in building the Network

1. DCN planning phase research (2016-2017)

2. Implementation launch in Spring 2018
   a. DCN staffing model + Advisory panel
   b. DCN training/networking events
   c. DCN workflow and C-U-R-A-T-E steps
   d. Assessment Plan

3. Grow the DCN beyond our grant-funded phase to a sustainable entity
   a. Criteria for new partners
   b. Proposed financial model (alliance curation-as-service)
Planning Phase (2016-2017)

1. **Compared local policy**, technologies, and workflows across the 6 planning phase institutions;

2. **Held six focus groups** with researchers on what data curation activities were important;

3. **Ran controlled pilots** of data curation workflows with 17 data curators to ID issues;

4. **Surveyed the 124** ARL institutions to gauge support for data curation services;

5. **Researched cost recovery** models for sustainable data curation and repository services;

6. **Held information exchanges** with leaders of successful collaboration projects;

7. **Analyzed one-year** of data types, disciplines, frequency, and curation levels needed vs taken).
Planning Phase (2016-2017)

How would we deal with conflicting policy issues?

What do researchers actually need our help with? Will they care if curation is distributed?

Can I trust someone else to curate our data? What about quality control?

What skills do we need? What types of data sets are deposited into our data repositories? How long does curation take?

Baseline Assessment

<table>
<thead>
<tr>
<th>Workflow Steps by Institution</th>
<th>Pre-ingest Curation?</th>
<th>Mediated vs Self-deposit?</th>
<th>Post-ingest curation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Consult only</td>
<td>Staging Area for deposit</td>
<td>Mediated deposit</td>
</tr>
<tr>
<td>Minnesota</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Cornell</td>
<td>X</td>
<td>X*</td>
<td>X</td>
</tr>
<tr>
<td>Illinois</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Michigan</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Penn State</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Wash U</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

Planning Phase (2016-2017)

How would we deal with conflicting policy issues?

What do researchers actually need our help with? Will they care if curation is distributed?

Can I trust someone else to curate our data? What about quality control?

What skills do we need? What types of data sets are deposited into our data repositories? How long does curation take?

Researcher Focus Groups (n=91)

Ave Rating = 3.7 out of 5

- Secure Storage
- Quality Assurance
- Persistent IDs
- Software Registry
- Data Visualization
- Metadata
- Emulation
- Restricted Access
- Contact Information
- Full-Text Indexing


Data Curation Network datacurationnetwork.org
Planning Phase (2016-2017)

How would we deal with conflicting policy issues?

What do researchers actually need our help with? Will they care if curation is distributed?

Can I trust someone else to curate our data? What about quality control?

What skills do we need? What types of data sets are deposited into our data repositories? How long does curation take?

---

ARL Institutions Survey (n=80)

---

Data Curation Network  datacurationnetwork.org
Planning Phase (2016-2017)

What are the measure of success?

How can we grow and sustain the Network beyond the grant-funding period?
DCN Implementation

9 Institutions
- 8 Academic Libraries
- 1 General Data Repository

19 Data Curators

1 Project Coordinator (new hire)

1 Program Director (PI)

8 DCN Representatives (CO-PIs)

2 Admin Leads
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, 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

DCN Coordinator Workflow

Review → Assign → CURATE → Mediate → Approve

DCN Curator Workflow

C: Check files and metadata
U: Understand and run files
R: Request missing information
A: Augment metadata
T: Transform file formats
E: Evaluate for FAIRness

Check files and metadata
Understand and run files
Request missing information
Augment metadata
Transform file formats
Evaluate for FAIRness
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 Implementation (2018-2020)

Assessment Plan (two-prong)

Is a networked approach to curating research data more efficient?

- Number of datasets
- Frequency (high-volume time periods, etc.)
- Variety (data file formats; range of disciplines)
- Efficiency (time, costs)

Are curated data are more valuable?

- Track reuse indicators (download counts, citations, alt-metrics)
- Implement a DCN registry
- Apply badges and metadata to signal that data sets curated by the DCN are FAIR.

Data Curation Network datacurationnetwork.org
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
Local Representative
Local Curator(s)
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

Data Curation Network datacurationnetwork.org
## Value proposition

<table>
<thead>
<tr>
<th>Stakeholder</th>
<th>Benefits</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Academic libraries with existing</strong></td>
<td>Gain access to data curation expertise in more disciplines/formats than locally available</td>
</tr>
<tr>
<td><strong>data curation services</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Academic libraries with limited to no</strong></td>
<td>Are able to provide critical new data curation services when local resources are limited (without needing to hire);</td>
</tr>
<tr>
<td><strong>no resources for data curation services</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Disciplinary- and general-subject data</strong></td>
<td>Receive better, more valuable data submissions from DCN partner institutions and customers;</td>
</tr>
<tr>
<td><strong>repositories</strong></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>

Data Curation Network  datacurationnetwork.org
DCN FAQ

- Do researchers actually value these services?
- Won’t researchers curate their own data?
- Is it another community of practice?
- Aren’t you all large research libraries? How about other kinds of libraries?
- Can’t some of this (data curation) be automated?
- Why aren’t you also sharing a repository?
- Why charge $ for this?

What other questions do you have?
Thank you!

Tim McGeary
tim.mcgeary@duke.edu

Claire Stewart
cstewart@umn.edu

Sophia Lafferty-Hess
sophia.lafferty.hess@duke.edu

Jennifer Darragh
jennifer.darragh@duke.edu