A Strategic Framework for Institutional Research Data Curation

PRESENTED BY:
Sayeed Choudhury, Associate Dean for Research Data Management, Johns Hopkins University
Anne Linton, Director, Himmelfarb Health Sciences Library, George Washington University
Working Group Goals

Pull from a variety of institutions to:

- Assess data curation current state and explore future paths
- Create a practical guide for institutions starting out or further along in the process of implementing a data curation program
- Create a document that would apply across a range of research interests

https://www.educause.edu/ecar/ecar-working-groups
Participants: Membership

- Co-chair: Mike Fary, Enterprise Data Architect, University of Chicago
- Co-chair: Dan Noonan, Assistant Professor, e-Records/Digital Resources Archivist, Ohio State University
- Sayeed Choudhury, Associate Dean for Research Data Management, Johns Hopkins University
- Esmé Cowles, Digital Infrastructure Developer, Princeton University Library
- Holly Croft, Assistant Professor, Digital Archivist, Georgia College
- Karen Estlund, Associate Dean for Technology and Digital Strategies, University Libraries, Pennsylvania State University
- Grace Lynn Faustino, Information Technology Project Manager III, Office of the Vice President for Research, University of New Mexico
- Thomas Hauser, Director Research Computing, University of Colorado, Boulder
- Anne Linton, Director, Himmelfarb Health Sciences Library, George Washington University
- Clifford Lynch, Executive Director, CNI
- Karen Menard, Assistant Vice President of Institutional Analysis and Research, University of Guelph
- David Minor, Director, Research Data Curation Program, University of California, San Diego
- Greg Monaco, Director for Research & Cyberinfrastructure Initiatives/Great Plains Network, Kansas State University
- David Ulate, Executive Director, Institutional Research and Planning, Foothill-De Anza Community College District
- Sarah Shreeves, Associate Dean for Digital Strategies, University of Miami Libraries
- Natalie Waters, Head librarian, Schulich Library of Physical Sciences, Life Sciences, and Engineering, McGill University

https://www.educause.edu/about/mission-and-organization/governance-and-leadership/member-committees/ecar-dc-working-group-members
Working Group Process

- Bi-weekly calls. Started with:
  - Scoping
  - Data curation definition and its role in the research lifecycle
  - Survey of membership to identify issues, maturity models
- Matrix of tasks and roles to guide document creation
- Document drafting (in process)
Objectives

▪ Explore relationship of data curation to data governance, stewardship, and protection/preservation
▪ Identify skills sets and professional roles for conducting data curation
▪ Review budget implications and organization structures to support data curation
▪ Promote data curation through research partnerships, coursework
▪ Provide minimum standards for best practices
Astronomy Image from Hubble Space Telescope
Image of Cancer Cells
Points of Deliberation

▪ Need for institution-wide conversation and strategy
▪ Build on past successes with DMPs, knowledge of funder requirement, library skills in organizing and cataloging data
▪ Place of data curation in research lifecycle
▪ Why curated data is important to the institution and the research enterprise
▪ Identify partners needed for success, identify leads
Focus on Research Data
Data curation is the process of ensuring that data can be understood and reused by interested parties across disciplines, organizations, and the passage of time. Curating data also implies making choices about where to invest limited resources and understanding likely needs for the data, ranging from experimental reproducibility to genuine repurposing. It subsumes preservation but goes beyond it. Stakeholders in the process include scholars, librarians, IT staff, funders, and policymakers.
Working Group: Initial Survey

- Organizational Structure
- Funding and Policies
- Services
- Staffing and Skills
- Infrastructure
- Future Plans
Survey: Organizational Structure

- **Who’s involved?**
  - 85%: faculty and library
  - 69%: academic departments
  - 39%: IT

- **Who manages the curation process?**
  - (42%): No formal management
  - In library (33%), IT (25%), and/or research (17%)

- Almost all schools (92%) don’t have a single point of contact; 75% of schools have multiple owners
Survey: Partners, Funding, and Policies

- Just over half have external partners (e.g., Hathi, DuraSpace, ICPSR, Portage Network)

- Funding:
  - 54%: Dedicated, ongoing funds
  - 46%: Research grants
  - 31%: Fees for services

- 54%: No policies in place for research data curation
Survey: Research Data Curation Services

▪ What services are offered?
  ▪ Storage (92%)
  ▪ Consulting (77%)
  ▪ Training (62%)
  ▪ Access & Analysis Tools (54%)
  ▪ Preservation Tools (46%)
  ▪ Management (39%)
  ▪ Graduate credit course, DMP tool (7.7%)

▪ When do you engage (data lifecycle stage)?
  ▪ Data Storage (54%)
  ▪ Archiving and Preservation (54%)
  ▪ Data Collection & Description (39%)
  ▪ Data Creation (31%)
  ▪ Discovery & Analysis (23%)

Only one school had metrics in place to measure the success and effectiveness of their efforts.
Survey: How are Services Offered?

- Mostly ad hoc (58%); 25% via formal service/institutional programmatic approach
- 92% of services are free; 31% also have chargeback models
- Primarily for (and used by) faculty and graduate students
- Marketed via:
  - 60%: Website
  - 40%: E-mail
  - 30%: Internal lists
  - 20%: Flyers
Survey: Staffing & Skills

- Staffing ranged widely -- from only part time or as add-on when time available, to 10 FTE/4 students

- Skills most lacking:
  - Data creation
  - Data re-use and transformation
  - Discovery and analysis

- Training offered primarily by library (69%), and often not at all (23%)
Survey: Infrastructure & Future Planning

- 31% thought they were lacking necessary infrastructure for diverse data formats; only 54% have data in their IRs.
- Most of the infrastructure is local (82%), with some in the cloud (36%) or off-premises (36%)
- All but one are expanding services:
  - “We have a set of recommendations before the office of research to formalize a research data services program.”
  - “I expect it to grow and become better organized/more centralized in the future.”
  - “Yes, our Coordinator works at it, but without institutional wide support it is very difficult.”
What’s in the Paper?

- Introduction: Data curation and the data lifecycle
- Research Data Curation Service Areas
- Things to Consider:
  - Policies
  - Funding
  - Staffing
  - Communication
Research Data Curation Service Areas

- Data Management Planning
- Data Discovery
- Data Creation
- Data Description
- Data Analysis
- Data Storage
- Data Access

Each area includes:
- Scope/description
- Roles/skills
- Sample activities
<table>
<thead>
<tr>
<th>Data Curation Tasks</th>
<th>Who does?</th>
<th>Who consults?</th>
<th>Who trains?</th>
<th>Who coordinates?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planning for Data Management</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DMP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Management</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Discovery</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data creation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data location</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data purchase</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data collection</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data manipulation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data extraction</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data cleaning</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Description</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metadata schema</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metadata creation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Lessons Learned

- Partnerships are essential for success: within departments (e.g., the library) and across the institution
- No one model will work at every institution
- Wide variety of maturity across institutions
- Changing nature of data curation
Contact Us!

- Sayeed Choudhury (sayeed@jhu.edu)
- Anne Linton (alinton@gwu.edu)
- Karen Wetzel, Senior Program Manager, EDUCAUSE
  - kwetzel@educause.edu
  - @eduwetzel
- EDUCAUSE Working Groups Website

Questions? Feedback please!