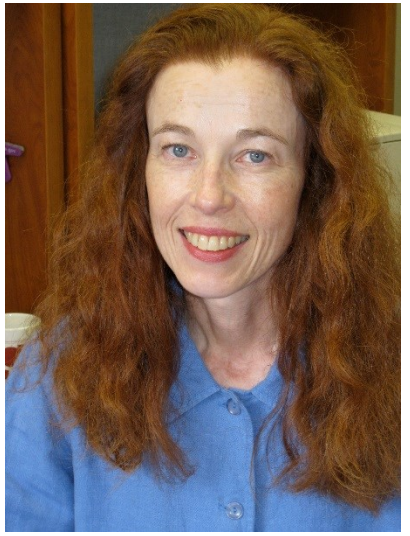
Open Data Meets Digital Curation: An Investigation of Practices and Needs

Christopher (Cal) Lee
School of Information and Library Science
University of North Carolina at Chapel Hill

11th International Digital Curation Conference
February 22-25, 2016
Amsterdam, The Netherlands

Acknowledgements

- Funding for project – Institute of Museum and Library Services (IMLS)
- Support for this presentation – Council on Library and Information Resources (CLIR)
- Collaborators and co-authors:
  - Alice Bishop, CLIR
  - Suzie Allard, University of Tennessee
  - Christopher (Cal) Lee, University of North Carolina at Chapel Hill
  - Nancy McGovern, Massachusetts Institute of Technology (MIT)



Background and Motivation

- Research data is a valuable resource for a variety of stakeholders across all sectors of society.
- In the US, research funded by the government produces a significant portion of these data.
- US law mandates that these data should be freely available to the public in “open access” (OA), i.e. unrestricted access and unrestricted reuse.

U.S. Executive Order on Open Access to Data (2013)

- “Increasing Access to the Results of Federally Funded Scientific Research” - required federal agencies with annual research and development expenditures > $100 million to create public access plans
- Applied to nineteen federal agencies, some with multiple divisions
- Additional direction provided by Executive Order “Making Open and Machine Readable the New Default for Government Information” (9 May 2013) and a memorandum (OMB m-13-13) with specific guidelines
- Common foundations established by these documents provided guidance for agencies as they drafted public access plans

Digital Curation and Open Access – Crossing the Streams?

- Two streams:
  - Substantial activity – much within academia and cultural heritage sector – to both define and support competencies for digital curation work
  - Significant move toward provision of open access to data created with public funds
- Neither stream has a single clearly defined professional home - both are undertaken by individuals with a vast array of disciplinary backgrounds, job titles and institutional contexts
- Institute of Museum and Library Services (IMLS): key U.S. player in the development of conceptual and professional approaches to digital curation, with a strong interest in implication of open access mandates

CLIR Project on Open Access Imperative

- In 2013, IMLS funded the Council on Library and Information Resources (CLIR) to conduct a project to help IMLS and its constituents understand implications of the US federal OA mandate and how to address needs and gaps in digital curation
- Three research components:
  1) **content analysis** of federal agency plans supporting open access to data and publications, identifying commonalities and differences (led by **Suzie Allard**)
  2) **case studies** (interviews and analysis of project deliverables) of seven projects previously funded by IMLS to identify lessons about skills, capabilities and institutional arrangements for data curation activities (led by **Cal Lee**)
  3) **analysis of capacity building efforts**, including educational programs, competency models, workforce indicators (led by **Nancy McGovern**)

US Agencies' Response to the Mandate

- Original timeline for plans' release was delayed by federal government sequester in 2013
- As of December 6, 2015, 16 agencies (or one operating unit) had made open data access plans public (blue below)

Department of Agriculture (USDA)	Department of Homeland Security (DHS)	Environmental Protection Agency (EPA)
Department of Commerce	Department of Housing and Urban Development (HUD)	Institute for Museum and Library Services (IMLS)
Department of Defense (DOD)	Department of the Interior	National Aeronautics and Space Agency (NASA)
Department of Education	Department of Labor (DOL)	National Institute of Standards and Technology (NIST)
Department of Energy (DOE)	Department of Transportation (DoT)	National Oceanic and Atmospheric Administration (NOAA)
Department of Health and Human Services (HHS)	Department of Veterans Affairs (VA)	National Science Foundation (NSF)
		Smithsonian Institution

Observations on Agency Public Access Plans

Open Data Infrastructure	Roles and Responsibilities	Making the Data Public
• Environment for coordination of open data activities not yet fully realized • Shared definition of “data” helps identification of potential collaborations • Clearly defined boundaries for the scientific data to be included (or not) • PubMed Central = highly adopted platform for research articles • Implementation won’t occur simultaneously across agencies • Recognize that research data management planning is part of a larger effort to enable public access	• Two issues that are not thoroughly addressed: role of education and costs/cost recovery • Access and storage provisions for data not as mature as those for pee-reviewed publications, requiring more effort from prospective users	• The plans for public access vary in how they approach empowering the public with the data • Recognize essential role of metadata for discovery and access

Case Studies of IMLS Digital Curation Projects

- Identified seven recent (2010-2013) IMLS-funded projects with significant digital curation objectives
- Aimed for diversity of project objectives, curation functions, data types
- Semi-structured interviews with key project personnel
- Coding and analysis of interview transcripts and project documentation

Project	Primary Focus
Creating a Better World by Sharing Research Online	Institutional repository (IR) to provide access to the university's research output
Databib	Annotated online bibliography of research data repositories
Datastar	Study researchers' data sharing and discovery needs and enhance a linked data platform to meet those needs
ETD [Electronic Theses and Dissertations] Lifecycle Management	Guidance documents and software tools for life-cycle data management and preservation of ETDs
Improving Data Stewardship with the DMPTool	Identify and propose strategies to address challenges to adopting the Data Management Planning Tool (DMPTool)
Virtual Archiving for Public Opinion Polls	Demonstrate and promote streamlined workflows for getting research data into data archives
What's on the Menu? – From Software to Funware	Support crowdsourcing of menu transcriptions

Insights and Observations from Case Studies

- Successful initiatives are part of ongoing capacity building
- Digital curation requires control over software
- Essential digital curation activities involve working with data and active engagement with stakeholders
- Convincing resource allocators is a key factor in many settings
- Value of releasing early prototypes to test with real data
- Meeting user needs involves many inferences about behaviors and expectations
- Potential to facilitate further interaction between users.
- Metadata satisficing is essential
- Open access involves not just enabling data discovery but also enabling new forms of interaction with and among users
- Value of pushing into Producer practices and behaviors

Capacity Building: Curriculum, Competencies, and Careers

- High-level goal: explore current capacity for preparing digital curation professionals
- Analysis of existing educational offerings, models and job postings
- Recognizes efforts in various countries, but major focus is on US context (given charge from IMLS)

Capacity Building – Observations and Insights

Curriculum Development and Implementation	Curation Competencies
• Increase in the number of certificate programs, with sustainability yet to be determined • Significant amount of curriculum material pertaining to digital curation that could be adapted, extended, or built upon • Online resources are plentiful but many also potentially at risk • Content that includes organizational and technological examples are desirable, fill a demonstrated need, but are challenging to update and sustain.	Profiles of four prominent frameworks: • Digital Curation Curriculum (DigCCurr) Matrix – DigCCurr Project • Digital Curator Vocational (DigCurV) Curriculum Framework • Staffing for Effective Digital Preservation - National Digital Stewardship Alliance (NDSA) • Preparing the Workforce for Digital Curation - National Research Council (NRC) Mapped to common categories: • Contexts • Management and administration • Functions

Conclusions

- Open data efforts and digital curation efforts can benefit each other, but the intersections are often not fully explored
- As digital curation roles and responsibilities emerge and change, new opportunities to engage and collaborate with data will open up fresh frontiers across the range of research domains
- Need for increasingly interdisciplinary approaches that will encourage and enable innovation of a kind we only now imagine.
- Stay tuned for final report: *The Open Data Imperative: How the Cultural Heritage Community Can Address the Federal Mandate*