EDITOR'S SUMMARY

Prompted by proliferating titles for those charged with managing digital data, archivists, embedded data managers, data librarians and data users explored terminology at the 2014 Annual Meeting of the Society of American Archivists. Digital data creation may originate with a submission to a repository, copies ingested at different locations or even reassembly of existing data. Contrary to conventional concepts for archives, data is not complete but may represent one version, a level in a process or point in a workflow. Data must be accessible throughout versions and stages and often across a network of locations. Additional discussions focused on data ownership and responsibility for stewardship, the need for a common vocabulary to support interoperability by managers in varied roles, and the reincarnation of data as it is reused. The session made evident the critical need to reframe communications among those involved with data management to overcome barriers rising from vocabulary differences.

KEYWORDS

terminology
archivists
librarians
digital libraries
archives
librarianship
document management
data set management

Archivist! Data Librarian! Asset Manager! Do the Differences Really Matter?

Reflecting on Breakout Discussions at the Society of American Archivists 2014 Annual Meeting

by Wendy Hagenmaier, Dana M. Lamparello, Karen S. Baker, Janina Mueller and Stewart Varner

Research data management librarian; digital asset manager; archivist/digital data specialist; born-digital processor; curation archivist; data curator – all of these positions have come online recently to address the explosive growth of digital data. What they all have in common – despite their varying titles – is digital data management. So why the variety of titles? Are we using different names for the same work? We developed this session at the Society of American Archivists 2014 Annual Meeting [1] as a series of breakout discussion groups around the themes of data creation, access and reuse to highlight areas in which data management roles overlap. The panelists represented four archetypal roles commonly encountered in the digital data world – archivist, embedded data manager, data librarian and data user – but were motivated by a desire to break down barriers extant among such positions and to explore the diversity of data and information needs in practice. Our session was intended as a first step in collectively developing a common conceptual understanding of semantics and roles to bridge disparate professional communities, including the archives community, research data management community, digital curation community and digital humanities community, among others. This column is a brief overview of the session, but we encourage you to join the continuing discussion by commenting here: goo.gl/yCRHqG

Data Creation

Digital data creation is a complex phenomenon encountered by the archetypes in differing circumstances. Researchers, for instance, create data in the field and laboratory. Yet dataset creation may also be recognized as occurring at the moment of repository submission. Or given the reproducibility of digital data, copies of the same dataset may reside in several different repositories where each views ingestion as a moment of creation. And finally, data creation may be identified during assembly of a new collection from pre-existing data.

Wendy Hagenmaier, Georgia Institute of Technology, can be reached at wendy.hagenmaier<at>library.gatech.edu; Dana M. Lamparello, Chicago History Museum, can be reached at lamparello<at>chicagohistory.org; Karen S. Baker, University of Illinois at Urbana-Champaign, can be reached at ksbaker2<at>illinois.edu; Janina Mueller, Harvard Design School, can be reached at jmueller<at>gsd.harvard.edu; and Stewart Varner, University of North Carolina at Chapel Hill, can be reached at svarner<at>email.unc.edu.

Across the archetypal roles, two concepts represent a shift in traditional archival thinking: awareness of data versions and data levels. With data versioning, the "doneness" of data is no longer taken for granted. Rather than producing final data records, research practices may result in the creation of many versions of a dataset due to corrections or upgrades. There also may be a multitude of data products resulting from transformations of the data. Data creators have developed data levels, distinguishing raw, refined, derived and interpreted data. Each level represents a well-defined step in the processing, transformation or presentation of data. The concepts of data versions and levels are foreign to many archivists and digital preservationists.

Data Access

Currently, each archetype is grappling with several challenges that complicate access to digital data, including the unprecedented quantity of data and the legal tangle of intellectual property issues endemic to such volume. Adding to the complexity is the evolutionary nature of digital records: as aforementioned, data are no longer "done" or definitively inactive; our roles must shift from providing access to discrete records to providing access to interactions and dynamic entities. Other current challenges include the seemingly impossible task of architecting systems that anticipate future users' access needs and provide meaningful access to levels of data through sustainable repository workflows. In our breakout groups, we discussed how providing creators with early organized access to data during the creation process is our entrée to encouraging their participation in curation efforts. And to ensure future access, we identified the need to become involved in data creation, policymaking and system building.

In order to address these challenges and transitions, ideal data access should involve network models – networks of discovery, data stewardship institutions and data professionals. The definition of access needs reshaping, as it will no longer mean access to one-stop-shop institutional or disciplinary repositories that must contain all records, but rather access to a network of linked information sources. Portable, replicable, linkable data (and metadata) mean that data will be accessible in different places and discovery will unfold in myriad locations. Digital data will become a renewable resource living in an ecosystem of repositories.

Additionally, a need exists to reevaluate how we define ownership (that is, not always as an exclusive right) and stewardship (not purely institution-based but repository-based, inclusive of national, global, inter-institutional, intra-institutional and community efforts). This emerging network model takes alliances to a new level, one that may be uncomfortable for traditional archives and libraries. It is a model that requires additional infrastructure and a change in mindset to create interoperability among institutions. Above all, the network model demands a common vocabulary to empower each archetype to collaborate much more closely than ever before. And yet, despite significant shared priorities and responsibilities, data managers are not always doing the same work; that is, specialization of roles can be very important. But our specialized roles must be interoperable and speak a common language.

Data Reuse

Understanding how and at what points data are created and ensuring their accessibility by redefining traditional concepts of access are the first steps to ensuring their future use. Emerging technologies and proliferating modes

of scholarly discourse are greatly expanding potential uses and creating new kinds of data users driven to the archives in search of raw materials for new projects. Scientific data may be used to replicate the results of an experiment or they could form the basis of an entirely new line of inquiry. Digital humanists looking for linguistic or social patterns across centuries of newspaper articles can mine digitized text collections. Historical social networks may be reconstructed and visualized based on information extracted from archived correspondences.

With the energy and excitement surrounding open data and the digital tools used to wrangle it, new methods and techniques will continue to emerge, and researchers will continue to find new ways to put archived data to work. Because we cannot anticipate every possible use, managers of digital data must maintain active and open lines of communication with end users and work together to push the limits of our collections. Of course, such collaborations will produce new data that will need to be accessible and ready for reuse, thereby establishing a new data creator, a new point of data creation and a new data level.

Conclusions and Invitation to Join the Conversation

The major takeaways from our discussions highlighted the following:

1) the iterative and cyclical nature of digital data work and the flexibility required for ever-evolving records;
2) the need for a common vocabulary to bridge professional divides and enable interoperability among specialized data roles, including reconsideration of key, traditionally archival concepts; and
3) a glaring need to reframe communication among digital data managers, creators and users as a natural part of data workflow, rather than treating it as a difficult crossing-boundaries effort.

Additionally, we recognize each archetype came from a fairly well-funded academic institution, which strongly impacts available resources and expectations in digital data arenas. We plan to include more environments, archetypal roles and types of data as we continue these discussions. Again, we encourage data managers from all types of environments and perspectives – including you – to join the discussion: goo.gl/yCRHqG. ■

Resources Mentioned in the Article

[1] Archivist! Data Librarian! Asset Manager! Do the Differences Really Matter? (August 16, 2014). Session 708, Society of American Archivists 2014 Annual Meeting. Description available at http://sched.co/1qX7PB6