Digital Data Curation – Examining Needs for Digital Data Curators

Margaret Hedstrom, School of Information, University of Michigan

Paper presented at *Cultural Heritage online: Trusted Digital Repositories*, International Conference, Fondazione Rinascimento Digitale, Florence, Italy, 11-12 Dec. 2012.

Abstract

As data increases in volume, complexity and value, there is a growing recognition of the need for digital data curation. The demand for data curation is not limited to libraries, archives, museums and other cultural institutions. Government agencies, universities, scientific enterprises, and data-intensive private sector companies increasingly need enhanced capabilities to improve data quality, protect data from various threats, and to exploit data assets to their fullest. Digital curation is both a promising career track for information professionals and a skill needed by everyone who has responsibility for data-intensive work.

This paper analyzes the critical skill sets that constitute data curation, recognizing that data curation entails substantive knowledge of data, information processing, application environments, and anticipated use. In cultural institutions, mastery of procedures and requirements is becoming secondary to skills in problem solving and innovative development of solutions. Data curators in scientific and business environments need a hybrid of core knowledge of data curation and deep knowledge of the specifics of the research domain or business processes that generate data. Effective education and training of data curators, therefore, demands attention to both the general principles and practices of data curation and the specific requirements of each domain.

Introduction

Data are becoming critical assets in many sectors of the economy, government, health care, and culture. As a consequence, there is a growing recognition of the need for people with the knowledge, skills, and aptitude to manage data in such a way to enhance its value as an asset. In the United States, calls for increased attention to data management as a critical skill and growing career option have come from the highest levels of government, scientific leaders, and industry. Exactly what that bundle of skills, aptitudes, and knowledge looks like, however, remains to be defined. The knowledge requirements range from basic skills in data management that everyone who handles data needs in order to work effectively to highly specialized expertise in digital curation technologies, systems, and operations. The precise combination of these skills is also context dependent and may vary by sector and organizational setting. Effective and efficient education and training for digital data curators will require a flexible framework that accounts for core digital curation knowledge and skills, knowledge and expertise in data, and a deep understanding of domain-specific needs and requirements. Identifying

the basic knowledge and skill requirements is further complicated by the dynamic nature of digital curation tools, technologies and practices. This paper presents a framework for identifying skill and knowledge requirements across a spectrum from curation-centric needs to discipline or application specific requirements.

Current Practice in Data Curation

Most models for data curation that are in effect today assume that data curation is a shared responsibility organized around a life cycle model where data producers play a role in data curation while data are collected and actively used and archives or repositories assume responsibility for curation of data once it has become inactive but needs to be maintained for future use by a designated community. The OAIS reference model, for example, makes a distinction between the producers, the archive, and consumers [Reference Model]. Interaction between the producers and the archive result from negotiations around the content, structure, and rights associated with submission information packages (SIPs). Where the regimen of the OAIS model is not strictly followed, repositories typically manage the heterogeneity of potential inputs by placing constraints on the types of data, file formats, and software dependencies they will accept and by establishing requirements for metadata and documentation.

There are several advantages to "normalizing" data at the point of ingest into a repository. First, it limits the variety of data and file types that the repository must handle to those that the archive organizing has expertise in and some confidence in its ability to maintain for the long-term. Second, this approach gathers all of the necessary metadata and documentation from the producers while they are still known and accessible. Third, this approaches assures that the data can continue to be used independent of its original production environment or any of the people involved in its creation. As data increases in complexity, volume, and value to both its designated community and secondary users, the sharp line between responsibilities of the producers of data and the needs and requirements of curators of data starts to blur, and as a consequence the ability of curators to manage data for the long-term becomes dependent on both the ability and the willingness of data producers to comply with the archive's requirements. Working effectively with the data deluge requires rethinking of the expectations, requirements, and division of responsibility between producers and the archive with deep implications for knowledge and skill requirements.

Curation-Centric Knowledge and Skills

One end of the spectrum of data curation knowledge and skills is "curation-centric" expertise that forms the essential knowledgebase of individuals who work in organizations that are dedicated to managing, preserving and disseminating data. Such organizations vary widely in size from small repositories to very large data centers. Independent of size, they also vary in scope from serving one specific designated community to a broad amorphous public. These organizations may be self-standing independent archives or units embedded within an organization. Regardless of the size, scope and organizational arrangements, their mission or line of business centers on

managing, preserving, and providing access to data. People who build careers in data curation collectively share a set of knowledge and skill requirements. Perhaps the most fundamental requirement is deep understanding of the data needs and requirements of the communities that they serve, including expertise in the data types, analytical methods and tools, standards of evidence, ontologies, representation schemes, and data practices of the data producers and users. Second, they need knowledge of best practices and standards for data management, representation, storage, security, and long-term archiving. Third, they need facility with the technologies used to produce data, used by the archive to maintain data, and used by the consumers of data to make productive use of data. This bundle of knowledge and skills is necessary to move from current practice where curation-centric organizations impose requirements that the producers and consumers must meet to one where the curation-centric organizations responds to the requirements of the producer and user communities.

There are two additional dimensions to the curation-focused end of the spectrum that have implications for the nature and type of education required. Ideally, the people who staff curation-centric organizations will identify themselves and be recognized as professionals with career aspirations that they can fulfill in an environment of growth and advancement. Furthermore, they will experience, participate in, and ultimately shape the technological environments in which they work. Educating a generation of data curators in the facile execution of current best practices will not produce the human capacity needed, nor will it provide the foundation for lasting and satisfying careers. A more effective approach will require emphasis on solving curation problems rather than executing curation procedures, developing skills in anticipating and responding to technological change, and an orientation toward innovation and continuous improvement.

Domain-Specific Data Curation Knowledge and Skills

At the other end of the spectrum are the knowledge and skills everyone who handles data needs to get the most out of data and to meet their obligations for responsible data stewardship or domain-specific requirements. The precise nature of the knowledge required is highly context-dependent with many factors bearing on what constitutes the relevant knowledgebase and skill set. Domain-specific knowledge and skills typically consist of very deep knowledge of a narrow set of factors. In science, for example, each discipline and sub-discipline has its own data types, ontologies, analytical methods and tools, standards of proof, and parameters of acceptable uncertainty and variance. In some fields, a common instrument or research protocol provides a frame of reference for judging the quality, relevance and utility of data. Some fields are organized around a common taxonomy or nomenclature. Some fields are heavily dependent on observations of actual phenomena or events, while others rely on experiments, models and simulations. Many fields in the social science and life sciences engage with human subjects where researchers are obliged to protect the identity of individuals and maintain confidentiality.

The distinguishing feature of domain-specific data curation knowledge and skills is the value proposition. While curation-centric organizations share a mission of managing, preserving, and disseminating data, curation in the domain-specific context is a means to

an end, not an end in itself. Therefore, the value proposition for developing data curation capacity within a domain, be it science, government or business, is how does data curation contribute to the organization's core mission. For scientists, investments in data curation need to be motivated by demonstration of value in the form of increased accuracy and validity of results, new discoveries, or deeper insights into scientific problems or by efficiency gains such as reducing time and effort spent searching for data, cleaning noisy data, or manipulating data to fit a scientific need. In business, data curation has to contribute to the bottom line by enabling a firm to remain competitive or produce new product lines or new services.

Data Curation in the Cultural Heritage Domain

Data curation in cultural heritage, like any other sector, includes a mix of curation-centric knowledge and domain-specific knowledge. If deep understanding of the data needs and requirements of the communities that cultural heritage institutions serve is a fundamental requirement, who are these communities and what are their data needs and requirements? Typically, organizations in the cultural heritage domain define the public as their primary community, and, if they hold research collections, researchers whose study relies on the type of material, region, historical period, etc. represented in their collections. In addition, cultural heritage organizations create, collect, and manage large stores of data that underlay their basic "business" operations, including acquisition and collection management data, registries and catalogs of holdings, user and visitor information, and data on internal operations.

Individuals whose careers in cultural heritage organizations are oriented around data curation will need expertise in the technologies, standards and practices for storage, management, long-term preservation of digital assets. In a data-driven economy with rapid technological change, expertise in today's technologies and good practices is not sufficient, not only because the knowledgebase will become outdated and the skills become obsolete, but also because curators in the cultural heritage sector need to help design, build, manage, and evaluate the next generation of digital curation systems and applications and the generations of technology that replace them. Cultural heritage organizations that hold research collections and serve researchers need curators with expertise in the data types, analytical methods and tools, standards of evidence, ontologies, representation schemes, and research trends in the fields of research that the collections support. Serving the general public effectively requires creative use of digital assets in marketing, education, and other types public-facing services such as web site interface and interaction design, packaging digital content for education, and creating digital products.

Cultural heritage organizations vary along a continuum from engagement primarily with retrospective conversion of physical and analog collections to digital form to acquisition and stewardship of born-digital content. Serving the user communities for these new forms of collections requires recognition that digital collections enable new types of inquiry and the creation of new cultural heritage services and products. For example, cultural heritage organizations can take advantage of digital assets to create new ways of searching and discovering "hidden" collections, provide on-line access to materials,

create virtual tours, or virtually reunify dispersed collections [Punzalan]. Exploiting these opportunities requires creative people with an aptitude for innovation, deep knowledge of the technological and commercial opportunities for new types of products and services, and knowledge of the ethical and legal constraints on reuse of digital content.

Filling the Skills and Knowledge Gap

The wide variety of skills and the many areas where deep expertise is required creates a daunting situation for cultural heritage organizations. There are several strategies, however, to meet the demands for digital data curators in the cultural heritage sector. Solutions begin by placing cultural heritage needs and requirements in the larger economy and ecology of data curation. Cultural heritage organizations do have domain-specific needs and requirements, not the least of which is a commitment to long-term sustainability of their digital assets. However, starting with the recognition of the fundamental common areas of knowledge and expertise that cultural heritage organizations share with other organizations opens more options for educating, recruiting, and training a generation of data curators who are well-versed in the fundamentals of data management, embrace the creative application of technology to solve problems, and are able to adapt to the specific context in which they will build their careers.

References

Reference Model for an Open Archival Information System (OAIS). Consultative Committee for Space Data Systems, Recommended Practice, Issue 2. CCSDS 650.0-M-2, June 2012. http://public.ccsds.org/publications/archive/650x0m2.pdf

Punzalan, Ricardo. Virtual Reunification: Bits and Pieces Gathered Together to Represent the Whole. Dissertation. University of Michigan, 2013.