

The Coming Metadata Deluge

James D. Myers

National Center for Supercomputing Applications

With the rise of computing, our ability to produce content – from raw data to summarizing documents – has exploded. People speak of “information overload” and “data deluge” to describe the problems caused by this explosion for those trying to find, analyze, comprehend, and store the growing body of material available. To date, the automation of processes to capture the history and context of data has not kept pace, making the development of systems for data preservation, curation, and discovery extremely labor intensive. However, it is not clear that this situation is permanent, and, in fact, there are many reasons to think that there will soon be a ‘metadata deluge’ as our ability to capture and share metadata catches up with our capability to produce data. If such a metadata deluge occurs, it would profoundly affect the role of libraries and the design of curation and preservation infrastructure. An analogous change occurred with the data deluge – as our ability to create, store, and share content increased, our ability to organize information became a bottleneck, and infrastructure such as the World Wide Web arose. The Web, which directly supported the ability for experts and non-experts alike to organize information, created a market for third-parties to re-organize existing material and for ‘competing’ entities to offer alternate organizations. The Web also enabled those without the means or expertise to maintain content to none-the-less develop collections. More recent innovations such as blogs, wikis, and community spaces (e.g. MySpace and virtual 3-D worlds such as Second Life) go even further in enabling content creation and organization without technical expertise or owned infrastructure.

While this stack of Web technologies does not support the requirements for creating, managing, curating collections and for long-term information preservation, there are emerging extensions in this area that have the potential to spark a transformation in these areas analogous to the Web transformation of information publishing and organization. For example, global identifier schemes such as Handles and Digital Object Identifiers and Life Science Identifiers now provide Web gateway mechanisms to create persistent URLs. More recent schemes such as the Archival Resource Key (ARK) bring a more Web-centric view and additionally provide a means of decoupling the roles of the initial information provider and subsequent curator(s) in the way identifiers are generated and in how metadata is attributed. Extensions to HTTP such as WebDAV and URIQA provide means of managing versions and metadata in XML and RDF formats. Specifications such as the Java Content Repository API (JSR 170 and JSR 283) are standardizing the same types of functionality at the programming interface level. These technologies are making their way into large data grid and repository software, but they are also being used directly in scientific applications and environments to enable up-front capture of metadata and data provenance information. For example, the CombeChem project has developed an experiment planning and execution environment that captures the experimental design, the provenance of data in specific experiments, and electronic notes related to the experiments as a single web of RDF information that can subsequently be

searched, viewed, and potentially harvested. The Collaboratory for Multiscale Chemical Science (CMCS) with which I have been involved provides a similar service for applications to record any and all information related to data and experimental procedures that has focused more on connecting information across scientific disciplines and related to dynamic community assembly and evaluation of reference data and associated computational tools. Many other examples could be cited – from ones like these that are primarily driven by the goal of directly increasing researcher and community productivity to those that are more specifically focused on the long-term curation of data.

Working from the Web analogy, emerging semantic and content management technologies, and the exploratory projects using them, one can anticipate a period of rapid change in the curation and preservation of digital data and in the role of libraries. Very rich information – with all the detail captured and/or used by all instruments and applications in scientific experiments – will be available via standard protocols in self-descriptive schema and directly available, given authorization, for inclusion in institutional repositories, community databases, reference collections maintained by scientific associations, etc. The information collected and any additional annotations generated by third parties will be transferable (thanks to unique identifiers) and the data/metadata could be migrated, cached, replicated as needed by the organizations interested in it. Questions about what to collect may become much more graded – should the content be indexed only, should it be cached for performance, should it be copied to reduce risk or to extend the retention period, what metadata and ancillary data should be indexed, cached, copied along with the primary artifacts of interest? It is possible to imagine that different institutions may make very different choices in these areas to customize their solutions and provide added value for specific user bases, with or without global coordination or concepts as master copies or tiered collections.

In planning for the next-generation of digital data curation and preservation capabilities, it is important to question our assumptions. While the expertise gained over centuries in curation and preservation will be central to robust solutions, it will be necessary to disentangle principles of information management from practices that actually represent compromise based on the current limits of technologies and organizational structures. Conversely, while technological progress will play a driving role, complex socio-technical issues will be faced in defining practical solutions that align with cultural and economic realities and are ‘just complex enough’ to serve society’s needs. If the web analogy is broadly valid, we are about to enter a period of rapid progress, new ideas, and new partnerships that will dramatically change and improve our ability to understand the world’s information.