Editorial

MICHAEL SEADLE
Dean, Faculty of Arts and Director, Berlin School of Library and Information Science, Germany

World Digital Libraries 7(1): v-vi (2014)

Five of the six articles in this issue address the use of digital content. This editorial looks at how the structure of the content in digital libraries could evolve to make it possible for them to become the basis of large scale data mining operations as part of a distributed scholarly knowledge management system. To explain this, it is helpful to consider the characteristics of knowledge management systems from a very high level perspective.

A traditional knowledge management system relies on highly curated data at relatively small levels of granularity, often in a relational database. The curation of this data is an important aspect of its quality and a key driver of its cost. The data is often selected and entered by hand at the institutional level, and a variety of competitive, legal, and sometimes practical considerations inhibit sharing.

The management of scholarly data operates at a much higher level of granularity with the journal article or the chapter as its typical base level, and with the responsibility for curation shared among a relatively large number of scholarly publishers and an even larger number of scholars who do peer review. The per byte costs of the curation are significantly less, but the aggregated utility of these forms of scholarly data is reduced because of the format inconsistency and the quality of the contents.

Data mining, text mining, and distant reading using computing algorithms are all mechanisms that attempt to extract useful information from the heterogeneous mass of scholarly data. Data mining scholarly content across a range of sources is a relatively new concept, and when applied to text-based information, it is like text mining. Distant reading is a new and relatively open concept that is defined in contrast to "close" reading and uses computing mechanisms to discover not just words, but context-sensitive information. Regardless of the name, the process requires a degree of consistency in the format to "read" efficiently across a heterogeneous mass of scholarly data. It also requires a plausibility analysis in order to filter out doubts from valuable content. When both conditions are met, accessible scholarly data effectively becomes a knowledge management system.

The easiest problem to solve technically is to standardize the consistency of the data. One option is to structure text-based content with XML, preferably using a common Document Type Definition (DTD) that every document in the set uses. In fact, there is already a great deal of text in ASCII and XML formats. The conversion

process is well tested and with contemporary optical character recognition, the vast majority of western language text content in the world could be made available in this format. Legal and commercial issues restrict the degree of freely accessible information, but broad scale university licensing makes the access possible for many in academic institutions. Standardization is nonetheless a problem. Much of the scholarly data is in PDF format, which includes ASCII content that can generally be extracted easily, but the ASCII content in PDFs is relatively unstructured and may need cleanup. Line breaks are an example. This kind of cleanup can be standardized at a modest cost. Imposing a formal structure on content extracted from a PDF file requires more context-based analysis. It is doable, but non-trivial.

Making reasonable machine-based judgements about the quality of content is more complex and costly, but is essential if a distributed digital library-based knowledge management system is to be feasible. The criteria are probably not generalizable across disciplines, which inevitably raises costs. Quality judgments over time depend on balancing internal and external information. Librarians and scholars tend to make initial judgments based on factors, such as publisher and author reputation, which are broadly measurable using bibliometric information, such as citation analysis, depending on the field and the degree to which that community accepts the validity of impact factors. Download statistics could play a role too, as well as other alternative metrics. Internal quality measures are harder to define. The simplicity of language is to a certain extent measurable by machines, but clarity is not yet reliably measurable. The quality of logic has no metric and is difficult to define even for humans. As natural language processing improves, it should be easier for machines to provide quality metrics for language and logic, but that time may be some years distant. It is easier for machines to provide a content analysis, both in simple single-word or phrase terms, such as a list of articles that include the word "metadata" or the phrase "knowledge management", and in terms of more complex context relationships, such as "metadata" in the same sentence or paragraph as "knowledge management". The effect of a content analysis like this is to push the granularity of usable information down to a level well below an article or chapter. This is still higher than the granularity of information in typical knowledge management systems, but may be close enough to be useful.

Viewing scholarly content as a searchable knowledge-base is not new. Google's Ngram Viewer already does this by running queries across the whole corpus of the scanned content from the Google Books project. The results give some crude but useful metrics, and Google has the resources to standardize and process large amounts of data. The goal of this editorial is to suggest that a similar approach to knowledge management could be possible with other forms of scholarly content.