Int J Digit Libr (2015) 16:3–4
DOI 10.1007/s00799-015-0143-5

The new knowledge infrastructure

Michael Lesk

Published online: 27 February 2015

For many, information science starts with Vannevar Bush's July 1945 essay "As We May Think". Before the war, science had been an individual pursuit (think of Sinclair Lewis's *Arrowsmith*). Bush wrote that we had learned how to have teams of scientists; he knew of the atomic bomb, although he could only openly mention the development of radar. That was an amazingly fast and effective development by the team at the MIT Radiation Laboratory. Now, he asked, how could we use teams in peacetime?

Traditional humanities scholarship was also an individual pursuit. The same information technologies, however, that transformed science have also transformed scholarship in other domains. We can locate, read, and study materials. We can explore texts, images, sounds, videos, sculptures, costumes, interactive objects and scientific data. We can cooperate at a distance via electronics; I knew something had changed in the 1970s when I saw someone send an email message to somebody else in the same room. I also remember in the early 1990s when I needed to check a citation for a book and I looked it up online even though I had a copy of the book in the office I was using. Distance can be irrelevant at any scale. Cooperation can extend around the world—in the sciences, look at the IRIS project for seismological data, or the International Virtual Observatory, or the Protein Data Bank. In the humanities, look at Worldcat, or ARTSTOR, or the International Children's Digital Library, or the International Dunhuang Project.

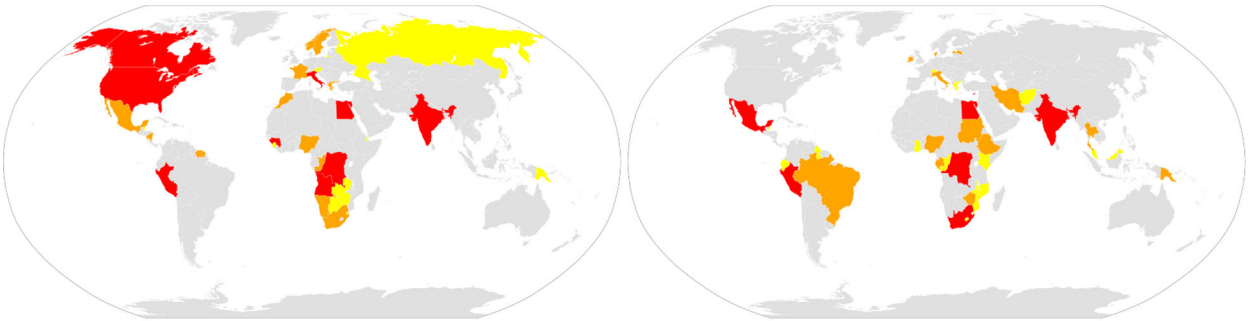
Increasingly, we also study using algorithms. Projects like the Sloan Digital Sky Survey collect so much data that it cannot be viewed by a single individual. Instead, the purpose of the data is for data mining. We have computational studies of authorship, stylometrics, analyses of paintings and musical performance, and other topics of scholarship to complement the scientific data mining of galaxies, chemical molecules, or weather events.

How do we enable these new kinds of scholarship? We need a new kind of knowledge infrastructure that will offer more than the rudimentary search and retrieval capabilities possible today. Conventional library subject indexing for books may be old-fashioned now that we have full text searching. A few years ago the Library of Congress floated a study even suggesting that the assignment of LCSH categories be phased out (the community objected). But we now need to search images, data, and other resources where text search is not immediately applicable. In addition, we have problems of quantity. The more material to be studied, the more accurate searching must be, so retrieval algorithms of greater resolving power are needed. And it is not just that individual projects are gathering more data, but that data availability is extending across disciplines and around the world.

The knowledge infrastructure we need must emphasize interoperability across areas and institutions. Libraries led the way with cooperative cataloging and standards for electronic records. Open Archives protocol use has now spread as well, but museums are trying an even more ambitious kind of description with "linked open data." In the spirit of the Semantic Web, museums are putting their catalog information into RDF (resource description format). In this methodology, all information about an object is recorded as a subject–predicate–object triple, with the predicates and objects taken from official ontologies with very precise definitions. The British Museum and the Rijksmuseum are leaders in this effort, with support from the Andrew W. Mellon Foundation. In principle, linked open data should permit deductive logic software to operate across museum data bases. In

M. Lesk (✉)
Rutgers University, New Brunswick, USA
e-mail: lesk@acm.org

practice, we are still learning both how to create and how to use this information. As an example of its value, however, seven months ago I wished to create chloropleth maps showing where British Museum objects had originated (see below). Merely using place names confuses such locations as Memphis in Egypt and Memphis in Tennessee, or Rochester in Kent and New York. The official geographic ontology in the RDF data clears this up. Since the descriptions of RDF for museum data are coming from international cooperative projects, we can look forward to increased interoperability across museums in different domains (decorative arts, natural history, and so on) as well in different countries.



Origins of British Museum objects: left, 18th century objects; right, 19th century.

The next step, and a more complicated one, will be scientific data. Today the instructions for describing scientific data are very complex, with hundreds of pages of documentation for formats such as FGDC (Federal geographic data committee) or SEED (standard for the exchange of earthquake data). These complex data formats do not necessarily interoperate effectively. The situation with medical data is similarly complex, with some accepted interchange standards such as DICOM (digital imaging and communications in medicine) for radiology, and with other standards that are proprietary including CPT (current procedural technology, sold by the American Medical Association).

Beyond data description is data use. Here scientific data exploration is ahead, rather than behind, the use in museums. But advanced algorithms using information resources are spreading to all areas of scholarship. Sentiment analysis of text based on Twitter messages, for example, shows that Hawaii is the happiest state and Louisiana the least happy. A remarkable example of image exploitation is Noah Snavely's "Building Rome in a Day" project, which took thousands of Flickr photographs of the Colosseum, discarded the out of focus ones, and then registered the images and computed a 3-D model using photogrammetry. Other examples of scientific collaboration with humanities scholarship has been the analysis of manuscripts, including Brent Seales work on "unrolling" a manuscript scroll using 3-D scans and image transformation software, or William Noel's reading unknown books of Archimedes from a palimpsest (an overwritten manuscript) . All of these projects are collaborative across institutions and knowledge domains. They show us how to build and use a more sophisticated knowledge infrastructure which can support teams of interdisciplinary researchers.

The *International Journal of Digital Libraries* offers current papers in the area of digital scholarship, especially the broader applications of modern techniques and methods. This issue ranges from text to video, network basics to linked data, and from the humanities to science. It introduces a wide vision of future research.