Learning from E-Databases in an E-Data World

The last two decades have been marked by a profound revolution in the creation, storage, and use of information. The dream of ubiquitous information environments may be at hand, but how well do they support scholarly and scientific research?

Despite the opportunities offered by the digital medium, early approaches focused on replicating information, as opposed to using digital technology to transform information. In higher education, we concentrated on preserving information in the same modalities that had been used for centuries—static articles and maps, for example—and simply changed the storage and access medium.

Fortunately, lessons from today’s practices can provide insight into how to innovatively respond to the infrastructure challenges of enabling cyberscholarship. Comprising new forms of research and scholarship that are qualitatively different from the traditional ways of using publications and research data, cyberscholarship is based on the widespread availability of digital content.1

Systems Science in a Data-Driven World

Today, many of the exciting and innovative developments in science and cyberscholarship are evolving at the intersection of transdisciplinary domains. Combining disciplines leads to new visions of the infrastructure supporting systems science and the emergence of data science. The integration of heterogeneous experimental data, which today are stored in numerous domain-specific databases, is a key requirement. However, a wide range of obstacles related to information access, handling, and integration impedes the efficient use of these databases.

Massive amounts of data produced on a daily basis require more-sophisticated management solutions than are available in today’s database environments; the use of the Internet as an enabling infrastructure for scientific exchange has created new demands for data accessibility as well. Furthermore, new fields such as earth systems science, computational pathomics, climate change, biogeochemistry, paleoclimatology, and systems biology have further increased the requirements demanded of databases and data repositories. The limitations of the current database environment will be increasingly magnified in an era of e-Research and e-Science.

Finding Relevant Sources

Even in the Google era, it is difficult to identify suitable data sources and well-described repositories via the web. Transdisciplinary research requires researchers to locate relevant data repositories and databases outside of their known fields.

One critical component of the emerging cyberinfrastructure is the array of instruments and sensors deployed on the grid. We need to create a global registry of instruments and sensors so that scientists and scientists-in-training can obtain information about them, including how to use them.2 A description, at a minimum, of the relevant data set or database contents, and of the way in which the data are produced and/or derived from other data sources, should be mandatory.

Data Processing

Imagine trying to support collaborative e-Science projects without large-scale, automated data processing. In an era when we’d like the data to speak to other data, a large number of scientific databases aren’t equipped with programming interfaces enabling software developers to query those databases from within their own programs and systems.

Public access to these interfaces is rarely provided. The rationale for denial ranges from security concerns to financial considerations. Web-based access is unsuitable for bulk queries, and programming interfaces are only rarely available. When data downloading is not an option, content must be extracted from the web interface. This suboptimized approach requires customized data-extraction software for each data source and has many technical limitations.

When downloading is supported, flat files are often still the de facto standard for data exchange. Because domain experts lack an agreed-upon standardized format for flat files, many formats for the thousands of data collections exist. Self-described XML files that could be readily harvested would solve many of these problems, since generic XML parsers are widely available—but only a very small number of databases are currently provided in XML. The importance of XML has been increasingly recognized, and standardized XML-based data-exchange formats should be strongly encouraged.

Content, Missing Content

Many useful types of information are missing in widely used databases; little incentive currently exists to (re)supply the missing data. As a standard practice, funding agencies should require the submission of fully described results to
public databases. To minimize the risk of human error during data submission, databases must implement appropriate curation protocols and supporting software. Since errors in data repositories and databases are a known problem, data providers should establish reliable means of reporting, tracking, and correcting errors in a timely manner.

Can't We Talk?
We need close bidirectional communication among database providers and users to address problems. While the Web 2.0 world has begun to adopt social software and connectedness as a means of collaborating, in the database world, many providers still desire to control their silos and consequently are not open about their data-curation processes, nor about schema and content changes. Error reporting and tracking is not the rule.

Missing in Education
Many use problems with scientific databases can be traced to a lack of interest in and basic understanding of data management among scientists, whereas informaticians may not be aware of the needs of scientists. The learning curricula for both informaticians and research communities should be better defined to equip future practitioners.

Access
Financial and political issues drive the most controversial dimension, that of ubiquitous access to data and databases. It seems obvious that free access for all to scientific data and databases would be beneficial, but the reality is more complex. Data curation with highly qualified staff is costly, and as a result, sustainability and financial issues arise. Most funding agencies do not provide long-term support for data curation, so alternative funding models are required. Depending on the funding model selected, different trade-offs result.

Some important databases are cost prohibitive and not widely available (e.g., Chemical Abstracts). Others are freely accessible through a web interface, although downloading is not permitted. Some providers block requests from entire domains when they suspect someone is attempting to “steal” data using automated data parsing from a web interface.

Licensing conditions of “free” licenses may impose considerable obstacles—for example, when database providers demand that the origin of the data be transparent to the user. Another licensing problem is data redistribution, which may not be permitted. The newest wrinkle is the demand that any publication making use of the database in any way must grant coauthorship to the database. Clearly, a universal legal framework for database interoperability is overdue.

Curation Requires Funding
The importance of databases is fundamental to entire disciplines such as chemistry and biology. However, long-term curation efforts are rarely supported, and most publicly available database providers have funding problems. Funding for the long-term curation of data repositories and scientific databases is essential. One can only wonder at the eventual state of massively scaled data repositories a decade hence if this is ignored.

An Evolutionary Direction: The Adaptive Web
Since we are in the early stages of developing the new paradigm(s) required to support data science and massively scaled data repositories, we have the opportunity (and obligation) to creatively reconceptualize our approach, lest we magnify current limitations in the scholarly communication chain. Increasingly, value resides in the relationships between researchers, papers, experimental data and ancillary supporting materials, associated dialogue from comments and reviews, and updates to the original work.

Typically, when hypertext browsing is used to follow links manually for subject headings, thesauri, and textual concepts and categories, the user can traverse only a small portion of a large knowledge space. To manage and take advantage of the potentially rich and complex nodes and connections in a large knowledge system such as the distributed web, users need system-aided reasoning methods that can intelligently suggest relevant knowledge.

As systems grow more sophisticated, we will see applications that support not just links between authors and papers but also relationships between users, data and information repositories, and communities.1 A mechanism that enables communication between these relationships, leading to information exchange, adaptation, and recombination, is required. That mechanism itself will constitute a new type of data repository. Designers are working on the next generation of information-retrieval tools and applications, which will support self-organizing knowledge on distributed networks driven by human interaction. This capability will allow a physicist, biochemist, or sociologist to collaborate with colleagues in the life sciences without having to learn an entirely new vocabulary.

Recent notable examples of distributed efforts that have succeeded with innovative approaches include diverse experiences such as the decoding of the human genome, the open source movement, and peer-to-peer networks. For those of us in higher education, it would be in our best long-term interests to optimize our communication systems to support a variety of approaches as we evolve our understanding of the coming adaptive web—as well as its impact on the building of data repositories that support both current and new forms of scientific communication. If we believe it is prudent to hedge our bets, many alternatives should be propagated and stimulated.

Notes

Rick Luce is Vice Provost and Director of Libraries at Emory University.