

eDatabase Lessons for an eData World

After witnessing at close hand the last two decades of transformation from paper to digital, wherein evolution to digital publishing and digital libraries largely has followed the linear thinking of replicating processes rather than transforming processes, and where the early phases of the paradigm shift could be characterized by essentially a duplication of the current print medium, what observations and lessons might be applied to the many problems confronting data science and massive repositories?

Steering Clear of Alexandria

The free-enterprise model requires competition to allow the best solutions to emerge by stimulating the experimentation and urgency required to bring new ideas rapidly in to play. Experience has shown that centralized solutions may not be the best option for ensuring innovation over time. The notion of a centralized data archive has some attractive advantages, notably the convenience of a one-stop shop but it also carries significant disadvantages. Centralized approaches are traditionally more vulnerable to failure and a tendency to pick winners prematurely.

The shadow of politics looms over any initiative dominated by a single organization (e.g., ACS) or country, and centralized repository approaches are typically discipline centric, which engenders problems categorizing new, trans-disciplinary science. It is an opportune time to experiment and rethink the assumptions that underlie our systems, but let us start by examining some issues in our current system that remain unsolved prior to being scaled up to support eScience.

Systems Science in a Data Driven World

Today much of the exciting and innovative developments in science can be found at the intersection of trans-disciplinary domains, leading to the need for new conceptualizations of the infrastructure supporting systems science and the emergence of data science. A foundational prerequisite for systems science is the integration of heterogeneous experimental data, which today are stored in numerous domain specific databases. However, a wide range of obstacles that relate to access, handling, and integration impede the efficient use of the contents of these databases. An examination of a few of the limitations surrounding the use today's scientific databases might be insightful in thinking about one dimension of the data repository challenges ahead.

NSF/JISC Repositories Workshop
«GreetingLine», Emory University April 9, 2007

Massive amounts of data produced on a daily basis require more sophisticated management solutions compared to today's database environments, and the availability of the Internet as an enabling infrastructure for scientific exchange has created new demands for data accessibility. Furthermore, new fields such as earth systems science, computational pathomics, climate change, biogeochemistry, paleo-climate, and systems biology has further increased the requirements that are demanded of databases and data repositories. Although we require systems that support ubiquitous knowledge and information environments, many issues arise requiring better solutions. The limitations that characterize our current database environment will be increasingly magnified in an era of eScience; some of those limitations include:

Finding Relevant Sources – Even in the Google era it is difficult to identify suitable data sources and well described repositories via the web. A first step in building models that support transdisciplinary science requires the researcher to locate relevant data repositories and databases outside of one's known field. One critical component of the emerging cyber-infrastructure 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. A description, at a minimum, of the relevant dataset or database contents and the way in which the data are produced and/or derived from other data sources is mandatory. Unfortunately, well-described global registries are not the norm and not every database provides such meta-information.

Data Processing – Imagine trying to support collaborative eScience projects without large-scale, automated data processing. In an era where we'd like the data to speak to the data, today a large number of scientific database aren't equipped with programming interfaces enabling software developers to query these databases from within their own programs and systems. Although current production systems can support standardized interfaces, 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, contents must be extracted from the web interface. This sub-optimized approach requires customized data-extraction software for each data source, and has many technical limitations.

NSF/JISC Repositories Workshop
«GreetingLine», Emory University April 9, 2007

Where downloading is supported, flat files are often still the de facto standard for data exchange. Lacking a 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, however, 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, which is not the rule in all domains. To minimize the risk of human error during data submission, appropriate curation protocols and supporting software must be implemented. Since errors in data repositories and databases are a known problem, data providers should implement appropriate means report, track and correct errors in a timely manner.

Can't We Talk - It seems almost too obvious to state that we need close bi-directional communication between 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 silo and consequently are not open about their data-curation processes, nor schema and contents changes. Simple as it may seem, 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 and basic understanding of data management on the part of the scientific expert, while informaticians may not be aware of the domain needs. Because communication difficulties that arise from these problems clearly have educational roots, the learning curricula for both informaticians and research scientists 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. The most important problem here is the question of free vs. fee access to scientific 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, and alternative funding models are required. Depending on the funding model selected, different trade-offs result. Some important databases are not publicly available (Chemical Abstracts), while 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' their data using automated data parsing from a web interface. Licensing conditions of 'free' licenses may impose considerable obstacles, e.g., when database providers demand that the origin of the data is transparent to the user. Another licensing problem is the redistribution of data, which may not be permitted. The newest wrinkle is the demand of co-authorship in any publication that makes use of the database in any way. 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 long-term curation of data repositories and scientific databases is required, and 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 constructing solutions for massively scaled data repositories, we have the opportunity (and obligation) to creatively reconceptualize our approach, less we magnify the current limitations in the scholarly communication chain.

Increasingly, value resides in the relationships between researchers, papers, experimental data and the ancillary supporting materials, associated dialogue from comments and reviews, updates to the original work, etc. Typically, when hypertext browsing is used to follow links manually for subject headings, thesauri, textual concepts and categories, the user can only traverse a small portion of a large knowledge space. To manage and utilize the potentially rich and complex nodes and connections in a large knowledge system such as the distributed web, system-aided reasoning methods would be useful to suggest relevant knowledge intelligently to the user.

NSF/JISC Repositories Workshop
«GreetingLine», Emory University April 9, 2007

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

Recent notable examples where decentralized efforts have succeeded with innovative approaches include diverse experiences such as decoding the human genome, the open source movement and peer-to-peer networks. It would be in our best long-term interests to optimize our communication systems to support a variety of approaches while we evolve our understanding of the coming adaptive web and its impact on building our 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.