What Is a Data System, Anyway?

4.1 INTRODUCTION: Search engines such as Google have become a mainstay in the toolbox that we use to approach almost any problem that requires access to information. The reason these programs have become so important is that they bring a modicum of order to the vast array of textual and graphical information available on the Internet. Think about it: there are literally millions of people who, independently, are making information available on the Internet, and yet we can often find the information in which we are interested in a matter of seconds. And all of this is accomplished with no centralized control, no entity saying what should go on the Internet or how it should be presented.

Remarkable. Why does this all work so well? There are a number of reasons: for example, a very large number of Web sites are in the same language, English; we are good at defining simple text searches that will find what we want; and powerful computers are available to do the indexing. But the following are the two most important reasons:

- A small number of well-defined protocols are used to transmit the data (HTTP at the highest level), as well as to format it (HTML, XML, etc.) for presentation.
- The "data" on which the system operates are text—letters, numbers, punctuation—with a major fraction encoded in the same form, ASCII.

It is therefore relatively straightforward to write programs to display this information, as well as to write programs to access and index it for subsequent searches.

Yet despite the incredible power of current search engines coupled with Web browsers, they provide only pointers based on textual information and are highly constrained in the type of information they can retrieve and display—generally only textual and graphical information. Of equal importance to the research scientist is access to remote data sets often held in large, complicated binary structures. Progress is being made in allowing seamless access to these data, albeit much more slowly than the progress that has been made in dealing with textual information. This slowness is due to the lack of a universally accepted access protocol and to the highly idiosyncratic way in which individual data providers organize their data. In addition, the decentralization of data resources, an attribute that is at the core of information on the Web, requires a fundamental shift in the way we think about data systems.

Historically, data systems have been developed by a centrally managed group, which took responsibility for all aspects of the system. None of these systems provided for the complete range of data system functionality—from discovery to analysis—over a broad range of data providers and data types. The first generation of data systems involved a single computer that could be accessed only locally. Such systems consisted of data, a search capability, and an ability to manipulate the data. Adding data to these systems was tedious at best. Programs had to be written to ingest data from storage media. Then the ingested data had to be cataloged and stored in a format that the data system could readily access. All aspects of these systems were controlled by the system builder, and such systems did not scale well either from the perspective of added functionality or from the perspective of the data available.

The ability to network computers and, in particular, the advent of wide-area networks resulted in a substantial change in the approach taken in the development of data systems. The first significant system elements to be built were online directory services designed to aid in the discovery of data. Initially, directory entries provided the user with a point of contact, such as a phone number or postal address, at the originating archive. As more data became available electronically, directory entries associated with these data were augmented with the electronic address of the data, generally the URL of the associated FTP site. To obtain data of interest, the user located the data set in the directory, then contacted the site holding the data and ordered them for either electronic or mail delivery. Once delivered, the data could be entered by the user into his or her application program for analysis.

The next step in the evolution was to build systems that integrated the discovery and delivery functions. Such systems are often referred to as "one-stop shopping" systems. They provide for a seamless connection between the data-discovery component and the data-delivery component, but little beyond that. These systems do not provide an end-to-end, integrated solution to data access; they are missing the component that provides seamless access to a data-analysis capability.

Because "one-stop shopping" systems are limited to data that can be ordered through the system, work has continued on the development of system elements that provide for data discovery only—such as the Global Change Master Directory, or GCMD (http://gcmd.nasa.gov). These systems provide electronic pointers to a much broader range of data sets, but they still lack the ability to order these data in a consistent fashion.

Independently of the discovery and order system elements, several large projects are developing data-access protocols that allow delivery of a well-defined data stream to the user's application package. The oldest of these is OPeNDAP, the Open-source Project for a Network Data Access Protocol (http://opendap.org/). OPeNDAP software hides the format in which the data are stored; the user requests a subset of data of interest from a participating archive via a URL, and the data are extracted from the archive, sent to the user's computer, and instantiated in the workspace of the user's analysis or visualization package. More recently, the Open Geospatial Consortium, or OGC (http://www.opengeospatial.org/), has developed an architecture addressing the same basic issue of data delivery, although there are some fundamental differences between the architectures. The most important difference is that OPeNDAP is a discipline-neutral data-access protocol (i.e., it works equally well with medical data, space physics data, Earth science data, etc.), whereas the OGC protocols are designed specifically for use in disciplines where geospatial location is critical.

Taken together, the efforts listed above point to the rapid evolution from centrally designed, implemented, and maintained data systems toward the development of data system elements. Different system elements are developed and managed by different groups, and there may be multiple versions of a given element—for example, different directory efforts. End-to-end data systems will be constructed from collections of these elements and hence involve distributed responsibility within the system. Furthermore, different end-to-end systems will likely use some of the same components. So, what defines a data system in such an environment?

To have a meaningful data system, a program (or system element) that interfaces the various system elements in such a way that the user can move seamlessly from data discovery through delivery to analysis is required. This system element is referred to as the *data system integrator*. It effectively defines the data system. The primary objective of the data system integrator is to aid the user by reducing the burden of the preliminary operations in a task requiring data analysis—those operations that are peripheral to the actual analysis of data. The OPeNDAP Data Connector, or ODC (http://opendap.org/ODC/), is an example of a rudimentary data system integrator. The user selects data via either the GCMD or a data set list maintained by OPeNDAP, requested data are extracted from the archive and moved over the Internet using the OPeNDAP data-access protocol, and analysis is performed either in the ODC or in any one of a number of application programs to which the ODC can deliver the data. (For textual and graphical information, a Web browser is the system integrator; a search engine, such as Google, is the information locator; and the browser is the analysis tool.)

Higher education institutions will play a central role in all aspects of the development of end-to-end data systems. They will contribute to the evolution of the fundamental data system elements. They will provide the initial user base that will test these system elements. And they will be among the more important data providers contributing data to the system. There is, however, one related area in which colleges and universities have taken a step backward. In the past, researchers often published their data, in paper form, as technical memoranda or some equivalent, and these reports were archived in the institutional library. With the advent of the Web, such technical reports have all but disappeared, with researchers publishing their data on personal Web sites; thus, the institutional commitment to a long-term archive of the data is waning. This is a trend that must be reversed; colleges and universities must provide a mechanism for researchers to publish their data electronically for permanent archival in the institutional library. Otherwise, a significant fraction of the data will be lost forever.

4.2 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 *cyber scholarship*. Comprising new forms of research and scholarship that are qualitatively different from the traditional ways of using publications and research data, cyber scholarship 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 cyber scholarship are evolving at the intersection of trans-disciplinary 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. Trans-disciplinary research requires researchers to locate relevant data repositories and databases outside of their known fields.

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.[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 sub-optimized 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.[3] 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.