

## What Is a Data System, Anyway?

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/>).

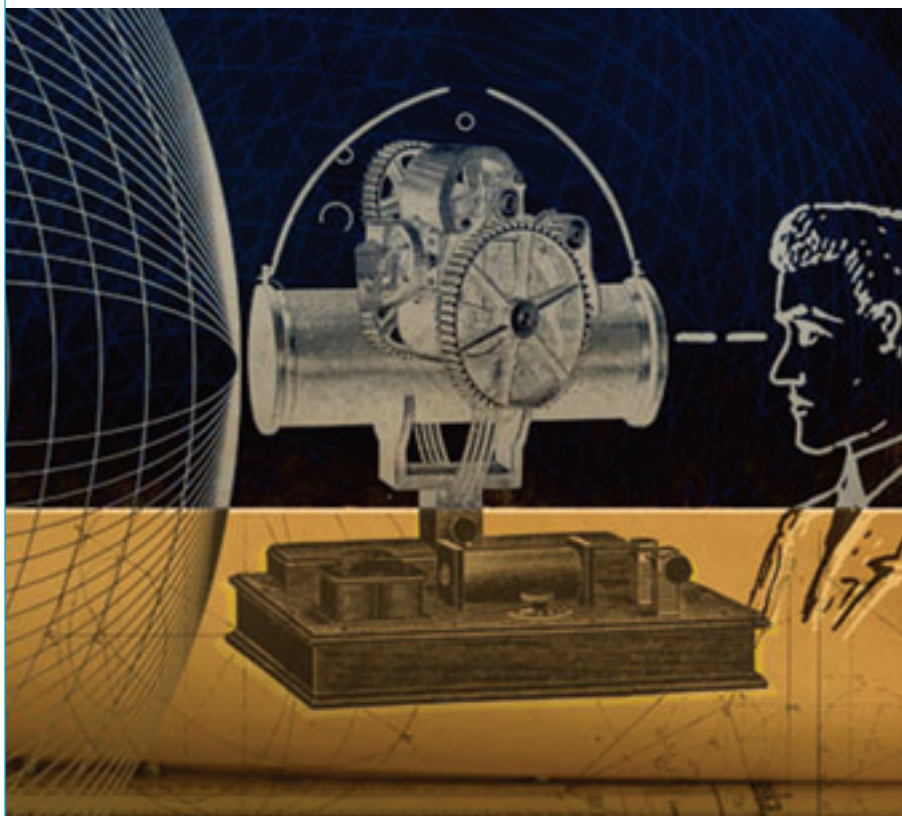


Illustration by Dung Hoang, © 2005

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.

**Peter Cornillon is Professor of Oceanography at the Graduate School of Oceanography, University of Rhode Island, and is President of OPeNDAP.**

