



Developments in International Information Systems and Services



Recent Developments in Electronic Access to the Data of Science

Brian Vickery

Abstract

Data are the bricks in the structure of science, and there is a need to preserve such data indefinitely for subsequent use. The paper discusses the nature of scientific data and demonstrates the great increase during recent decades of the volume of data. This growth has been particularly enhanced by the techniques of remote sensing of both stellar and terrestrial data, whose development is briefly described. The transition of data storage and handling to an electronic format is outlined, along with the development of data access via the Internet. Particular examples of databases in biology, astronomy, and earth sciences are described as well as the growth of access to distributed databases through “grid” technology. Some problems that hinder access are discussed.

Data have been likened to bricks in the structure of science, bound together by a mortar—the theoretical concepts that create the structure. However theory may change, the data remain, the cumulative material out of which science is built.

Scientific data therefore have a permanent importance. They are needed not only at the time of their generation by observation or experiment but also for the future. There is the perpetual possibility—and actuality—that old data will be revisited, to make up the building blocks of new theory, in the same or a related scientific field. Hence data must be permanently preserved in a form that permits access by future scientists (and technologists who use scientific data) in whatever field they are working.

The U.S. National Academy of Sciences gives some examples of the need for long-term data preservation (National Academy of Sciences, 1997). In the life sciences the report cites such observational data as experimental agricultural field tests going back to the nineteenth century, breeding histories of domestic ani-

mals and plants, and hospital and other medical records. The study of earth processes may involve the scrutiny of retrospective observations going back centuries or even millennia.

Traditionally, of course, scientific data have been preserved in print, papers, books, encyclopedias, and compendia; in near-print reports; in manuscript laboratory or field notebooks or logbooks; in graphs, diagrams, drawings, maps, photographs, cinefilm, or even models; or in such instrumental recordings as seismograms or stellar photographic plates. Today electronic handling of scientific information is becoming the norm, affecting all forms of data preservation. This paper aims to review some aspects of this transition.

The Nature of Data

Data exist in every field of science, as shown in Table 1. Various levels of data are distinguished: “raw” data, or the primary observations recorded; “calibrated” or edited data, in which anomalies or known instrumental biases are corrected; “reduced” data, or data expressed in some standard way; and “critically evaluated” data or “best values,” where there is more than one observation of the same entity available. For data to be useful to later workers, more than the actual data needs to be recorded. In all fields of science the technique by which the data were obtained should be given (perhaps in some detail). In many fields of science the time and place of the observation also needs to be added. Data are never collected without a purpose—which may well affect the manner of their collection and recording—and so some reference to this purpose and to any associated theory can also be valuable.

A general problem among disciplines is the relatively

Table 1. Some Categories of Data for Various Areas of Science

Physics
Physical properties of material
Empirical data demonstrating relations
Elementary particles
Atomic structures
Radioactive data
Nuclear reactions
Chemistry
Measures of physical properties of chemicals
Observations of chemical behavior and reactions
Crystal structures
Chemical structural formulas
Biosciences
Descriptions of plants, animals, and microorganisms, and of their internal organs
Observations of animal behavior
Physiological data
Diseases
Biochemical reactions
Genetic data of various kinds
Astronomy
Stars and other stellar objects, their positions, motions, and parallaxes; their radiation emissions and spectra, etc.
Celestial events such as eclipses, novas, solar flares, comet appearances
Earth Sciences
Geological structures
Geographic descriptions
Properties of rocks
Properties of minerals
Properties of soils
Weather data—temperatures, pressures, wind, precipitation, etc.
Oceanographic data—depths, currents, salinity, tides, etc.
Events such as volcanic eruptions, earthquakes, hurricanes
Environmental Science
Food chains and other ecological interactions
Pollution data
Events such as epidemics and famines

low priority attached to data management and archiving. Archiving is often viewed as something that happens after the research is completed and is done by someone other than the researcher. The most important current deficiencies are in the adequacy of documentation and in access to data in a usable form. Another problem relates to inadequate directories describing what data are available, where the data are located, and how and under what conditions users may access the data. Technological developments also pose an important challenge: data may become inaccessible if they are not regularly migrated to new storage media as hardware and software used to access the data become obsolete. Thus, regular

updates and migrations of data are needed to ensure continued access.

The Volume of Data

As is well known, scientific activity and publication have been expanding exponentially for several centuries, their size doubling every twenty years or so. It is only logical that the amount of scientific data in existence should have shared this growth. A simple illustration may be offered by considering the size of the Landolt-Börnstein tables: the 1883 edition was a book of 261 pages; the 1960–86 edition covered 52,000 pages—a doubling in size every twelve years. A doubling period of thirteen years over the last two centuries is reported by Joachim Schummer (1997) for known chemical compounds; Chemical Abstracts Registry in May 2002 recorded nearly twenty million compounds (Chemical Abstracts Registry System, 2002).

It may be roughly estimated that the 52,000 pages of the Landolt-Börnstein edition cited might contain 150 megabytes of data (0.00015 terabytes, one terabyte being a million million bytes). To put this in perspective, we may cite Howard Resnikoff and James Dolby (Salton, 1975, p. 207) in the United States, who in 1971 estimated some order-of-magnitude figures: a typical university library, 0.5 terabytes; a national library, 15 terabytes. (These figures assume that all pages consist of ASCII text: if the images in the collections were digitized, the total byte sizes would increase considerably.) Relative to these figures, the volume of physical data did not at that time seem too large to handle, even though the Landolt-Börnstein figure represents only “critically evaluated” data, which are a good deal less voluminous than the raw data on which they were based.

But within recent years the volume of scientific data has been growing enormously in many areas of science. This growth has been especially rapid because of the use of remote sensing systems, which give rise to data in the form of images. To give one example, it was estimated that in 2000 raw data holdings at the U.S. Earth Resources Observation Systems Data Center would total 642 terabytes and would be doubling every two or three years (National Academy of Sciences, 1997). It is also estimated that each year the detectors at the Centre Européen de Recherches Nucléaires (CERN; European Organization for Nuclear Research) particle collider in Switzerland record 1 petabyte of data (1 petabyte equals 1,000 terabytes). The Large Hadron Collider, under construction at CERN, is expected to produce

100-petabyte data sets. The Terraserver Web site—perhaps the world's largest on-line atlas—provides rapid access to five terabytes of aerial, satellite, and topographic images of Earth (at present restricted to the United States) (Terraserver, 2002).

Let us make some further comparisons. A recent estimate (Bergman, 2000) puts the total information directly accessible on the World Wide Web in 2000 at 17 terabytes or more—although it is also suggested that the size of databases attached to the Web but not directly browsable, the “deep web”—for example, DIALOG—may total 7,500 terabytes (DIALOG itself is 11 terabytes overall). The size of a large full-text database at the U.S. Patent Office is put at 2.4 terabytes. With all these figures we are now talking about information and data collections very much larger than those existing a few decades ago.

Scientific Data Activity

International scientific interest in access to data was evident as early as the 1920s, when the *International Critical Tables* started publication. But a new stage began in 1957–58 with the International Geophysical Year, in which sixty countries collaborated in making observations. To cope with the data collected, three World Data Centers were established. In 1966 the International Council of Scientific Unions set up a Committee on Data for Science and Technology (CODATA). Its aim was to improve the quality and accessibility of data and to facilitate international cooperation. One of its tasks was to identify the categories of data with which it would be concerned: not only quantitative data (e.g., physical, chemical, meteorological, and physiological) but also qualitative (e.g., chemical structures, classes of rocks, stellar spectra, biological taxonomy, and gene sequences). The data could be in the form of numerical values, diagrams, graphs, models, maps, or symbols. It could be primary observational or experimental data or derived by theoretical calculation.

In some fields—for example, chemistry, meteorology, geological survey, and astronomy—the volume of data became immense, as we have seen. The amount of scientific research data became far greater than could be reported in the publication system, and even that which was published was inadequately tagged and indexed, making access difficult. Some specialized journals devoted to data publication began to be produced, such as the *Journal of Chemical and Engineering Data* (as early as 1959), *Atomic Data and Nuclear Data Tables*

(1969), and the *Journal of Physical and Chemical Reference Data* (1972). Data centers—institutions concentrating on the production, evaluation, and dissemination of a particular field of scientific data—began to grow in number (Vickery, 1999).

A more recent development is the *Journal of Astronomical Data*, which accepts raw and reduced data from observations as well as from theoretical research, Ph.D. theses with extensive data tables, catalogs, graphs, and articles accompanied by the data on which they are based. Contributions briefly describe the astrophysical background and motive for the scientific work done, give a complete description of the data-collecting process, and assess the accuracy of the data as well as the data themselves. All data are published on CD-ROM (*Journal of Astronomical Data*, 2002).

International cooperation in the generation and management of scientific data is not new, particularly in astronomy, but it has continued to grow. One of the oldest international observing systems is that of observations by ships worldwide of weather and sea surface temperatures, transmitted by radio.

Some cooperative projects, such as the thermodynamic or spectroscopic properties of a group of chemicals, involve small collaborative teams who communicate via the Internet. Larger and more costly projects require long-term, formal arrangements between worldwide institutions and government agencies, also much aided by electronic communication. An example of a major project involving transnational data flow was the International Decade of Natural Disaster Reduction (1990–1999) which aimed to study the processes that cause such disasters as earthquakes, volcanic eruptions, floods, and hurricanes (World Meteorological Organization, 2000).

Electronic Data Records

The use of computers to acquire, massage, archive, access, retrieve, display, and evaluate data in a variety of forms has continuously expanded during the last three or four decades. Electronic records of data (specifically data banks, which will be referred to here as databases) may be created either by transferring paper records to machine-readable form or by recording data electronically. The first route has been much used in the physical and biological laboratory sciences, but in the observational astronomical and earth sciences, direct electronic-data recording has become predominant.

With the coming of on-line hosts, some databases

began to be publicly available, particularly in the field of toxic chemicals (Environmental Chemicals Data and Information Network [ECDIN], Hazardline, Hazardous Substances Data Bank [HSDB], Registry of Toxic Effects of Chemical Substances [RTECS], and more recently the Dictionary of Substances and Their Effects [DOSE]). STN International hosts about twenty databases covering the physical and mechanical properties of thousands of materials, and there are now many Internet sites hosting material safety data sheets (Rossmasler, 1980). Bibliographic databases such as those of *Chemical Abstracts* or Derwent include much numerical data.

But many more databases have been built for access by specialists only. To give some examples, in 1977 the U.S. Geological Survey had over fifty systems handling geoscience data in digital form, totaling perhaps 0.5 terabytes (Rossmasler, 1980). The Evaluated Nuclear Structure Data File has existed in electronic form since the 1970s, at first in the form of magnetic tapes, and is now available through the Internet from the U.S. National Nuclear Data Center Web site (National Nuclear Data Center, 2000). The Collider Detector at Fermilab records high-energy physics data, again initially on magnetic tape. By 1995 the volume of data had reached 20 terabytes (National Academy of Sciences, 1997). With the growth of networked communication among scientists a growing number of databases have become widely available on-line (e.g., the Beilstein and Landolt-Börnstein compendia), nowadays through the World Wide Web.

The total number of electronic data banks is now very large. Actions taken to improve access to them have been of three kinds: attempted integration of the data in a world center; if the data remain dispersed, provision of comprehensive directories and indexes to them; or provision of electronic access in an integrated way to distributed databases.

Starting in the 1960s, conceptually we may describe the stages of transition to electronic data access as follows:

1. Conversion of existing printed data to machine-readable form.
2. Distribution of electronic data via magnetic tape or CD-ROM.
3. Direct recording of data in electronic form.
4. Distribution of electronic data via on-line hosts, via ftp and gopher over the Internet, and eventually via the World Wide Web.

5. Sharing and exchange of electronic data to build up comprehensive databases.
6. Creation of directories to dispersed sets of databases.
7. Construction of integrated, distributed database access.

These stages cannot be directly related to specific dates, since they have occurred at different times in various fields of science. A useful general discussion of the “data deluge” is provided by Tony Hey and Anne Trethethen (2003).

Biology Data

The handling of genetic data is a good example of stages 1 to 5 in the above sequence. The structure of DNA was established by James Watson and Francis Crick in 1953, as a long, repetitive string of nucleotides, each with one of four nitrogenous bases. Each triplet of successive bases determines the production of one of twenty amino acids, the building blocks of proteins. A string of successive triplets forms a gene, determining the production of a specific protein. This is the basis of the genetic code. During the 1970s the number of known nucleotide and gene sequences increased rapidly, and in the 1980s the data came to be stored electronically, making it accessible on-line. Several major data banks were established, such as GenBank; the data library of the European Molecular Biology Laboratory (EMBL); and GENINFO. Data were distributed from such centers on magnetic tape, on-line, and later on CD-ROM. In 1990 Derwent made available GENESEQ, data derived from published patent applications. In 1993 it was noted that EMBL data could be accessed via the JANET network in the United Kingdom and ARPANET in the United States (Sillince & Sillince, 1993).

The project to map the whole human genome received its first funding in 1987, and in 2000 completed its basic task, feeding the results into GenBank. Today GenBank, EMBL, and other centers exchange data, and in February 2002 more than 17 billion amino-acid bases from over a hundred thousand living species were recorded, the collection doubling every fourteen months (National Center for Biotechnology Information, 2002). The size of GenBank is estimated to be 0.03 terabytes, which may be compared with the size of the bibliographic database PubMed, estimated at 0.04 terabytes (Bergman, 2000). Other genomic data is held privately by commercial firms.

CODATA task groups and commissions are now

engaged with other biological data projects. There is currently no master inventory for the world's plants against which species diversity, genetic resources, and pharmaceutical materials can be assessed. Because of major inconsistencies in the classifications used in local inventories, putting together a master inventory involves bringing the classification of each group of plants into worldwide consistency. The inventory will initially include the higher (vascular) plants—about three hundred thousand species (CODATA, 2002).

CODATA is also involved in a cooperative program with the International Unions of Biological Sciences and Medical Sciences to accelerate the development of a Global Master Species Database, with the goal of creating a virtual index of all the world's known species of organisms.

Astronomy Data

The increasing efficiency of cameras and photon-collecting devices used by astronomers, either from ground-based telescopes or from space experiments, is generating an unprecedented accumulation of data. In the 1980s dedicated computer archives and databases, accessible remotely, were the appropriate answer to the data retrieval problem. Later on, in order to face the diversity and complexity of access to data for the astronomers as part of their research work, integrated information systems began to build up gradually. Examples are the U.S. National Aeronautics and Space Administration (NASA) Data Archive and Distribution Service (ADS) and the European Space Information System.

Data were needed where the expertise to handle them was located, that is, as close as possible to the data providers, who were usually also the first users, able to understand and process the raw data, and who possessed the routines for producing final data in physical units—a process that sometimes implied several years of iterative improvements. But such teams were generally reluctant to devote time, manpower, and money to the different problems of data distribution to a wider community (e.g., documentation, homogenization, and building of friendly user interfaces).

At the same time the growing complexity of data systems called for a change of concept. The scientist had to manipulate not only data but also information, not only observational data but also, for example, documentation about the data, knowledge about the instruments, calibrations, bibliographical references of published re-

sults, and cross-references to other data sets. Pointers and links between these different pieces of the same puzzle were needed.

The World Wide Web provided easy ways of making all kinds of data and information available through the network; of bringing together data collected at different physical locations, provided they all used the Web http protocol; of creating links between data tables, publications, images, or data sets; and of gaining access to this variety of data, documentation, and multimedia information through browsers. The astronomical community showed an early interest in these developments. Several dozen Web servers were already available in mid-1993, several hundred in mid-1994, and one thousand at the end of 1994. We will now look at just a few of the facilities available today (Egret & Albrecht, 1995).

One example illustrates well the transition from printed compendium to database to Internet presence (Vickery, 1999). The conversion of star catalogs—the oldest form of recorded scientific data—into database form began in the 1960s. In 1970 the International Astronomical Union set up a bureau to distribute information on machine-readable catalogs. Two years later the Centre de Données Stellaires (CDS; Center for Stellar Data) was formed in Strasbourg and started to collect all such existing databases. (The center is now known by the same acronym, CDS, but its name has changed to Centre de Données Astronomique de Strasbourg.) During the next decade similar data centers were set up in the United States, Germany, the U.S.S.R., and Japan. The same stellar object could be named in various ways in different catalogs, and a “cross-identification” file of synonyms had been started in 1971 (Jaschek, 1989). The next step was to use this file to connect the different catalogs, and at Strasbourg an integrated database, SIMBAD, was established that combined all the information on over three million stellar objects. In addition to this database others were compiled as a result of special astronomical projects. For example, the data collected by the International Ultraviolet Explorer satellite (from 1978 on) was housed in Spain. The archives of data from the EXOSAT (x-ray information) and IRAS (infrared) satellites, launched in 1983, were housed in the Netherlands. The Space Telescope archive was set up in Germany. Other data archives were at the U.S. National Space Science Data Center. Specialized communication networks for use by astronomers were developed during the 1980s, for example, the British STARLINK, the Italian ASTRONET, the U.S. NASA

Science Network, and the international Space Physics Analysis Network (SPAN). In 1993 the CDS made publicly available an astronomical information service on the World Wide Web (Centre de Données Stellaires, 2000), providing access to over 1,400 star catalogs, including SIMBAD.

Another facility provides an example of transfer of data from printed to electronic form. The Astrographic Catalogue was the result of an international effort from 1891 to 1950. It was designed to photograph and measure the positions of all stars brighter than magnitude 11.0. In total, over 4.6 million stars were observed, many as faint as thirteenth magnitude. The sky was divided among twenty observatories, by declination zones. Each observatory exposed and measured the plates of its zone, subsequently publishing the plate measures. The U.S. Naval Observatory has completed data entry of the printed volumes of the Astrographic Catalogue and has converted the star positions to a common system (U.S. Naval Observatory, 2002).

The Astronomical Data Center, at NASA Goddard Space Flight Center, in Greenbelt, Maryland, is a cooperative effort between the National Space Science Data Center and the World Data Center A for Rockets and Satellites and its parent organization, the Space Science Data Operations Office. The Astronomical Data Center archives hold more than 670 catalogs of astrometry, photometry, spectroscopy, radio, and other miscellaneous data for stellar and nonstellar objects. The data were acquired through exchanges with the CDS and other astronomical data centers throughout the world and by direct contributions from the international astronomical community (Astronomical Data Center, 2002).

Remote Sensing

As previously noted, remote sensing has become an important source of scientific data (Gibson, 2000; Short, 2002). Remote sensing uses instruments or sensors to view the spectral and spatial relations of observable objects and materials at a distance, typically from above them, or in astronomy by looking outward. The use of photography to record the star images seen through telescopes began in the second half of the nineteenth century (Jaschek, 1989). Today the Hubble Space Telescope archive is providing a wealth of new stellar photographic data (Hubble Space Telescope, 2002).

Remote sensing of the earth's surface began in the 1840s as balloonists took pictures of the ground, using the newly invented photcamera. Aerial photography

became a valuable reconnaissance tool during World War I and came fully into its own during World War II. The entry of remote sensors into space began with automated photcamera systems mounted on captured German V-2 rockets, launched from White Sands, New Mexico. With the coming of the Soviet *Sputnik* in 1957, putting film cameras on orbiting spacecraft became possible. The first cosmonauts and astronauts used handheld cameras to photograph selected regions of land as they orbited the globe. Sensors tuned to obtain black-and-white television-like images of Earth were placed on meteorological satellites. The first such satellite, TIROS (Television and Infrared Observation Satellite), was launched in 1960, providing global quantitative views of temperature, cloud, and moisture distributions, as well as of surface properties (e.g., ice cover and soil moisture).

As an operational system for collecting information about Earth on a repetitive schedule, remote sensing matured in the 1970s, when instruments flew on Skylab and later on the Space Shuttle and on Landsat, which was the first satellite dedicated to mapping natural and cultural resources on land and ocean surfaces. A radar imaging system was the main sensor on Seasat, launched in June 1978. By the 1980s Landsat had been privatized, and widespread commercial use of remote sensing had taken root in the United States, France, the U.S.S.R., Japan, and other nations. At least forty-five land observation satellites will have been launched between 1995 and 2005. Much of this activity was, and is still being, driven by the increasing awareness that Earth's environments are in peril from man's activities and misuses.

Some of the uses of remote earth sensing are for geological mapping; for studying surface water, ice, snow, floods, or ocean waves and circulation; for analyzing land use and vegetation types; and for monitoring the environmental effects of man's activities and natural disasters.

Earth Data

Let us first look in more detail at Landsat. In the mid-1960s the idea of a civilian earth resources satellite was conceived by the U.S. Department of the Interior. Later NASA embarked on an initiative to develop and launch the first Earth-monitoring satellite to meet the needs of resource managers and earth scientists. The U.S. Geological Survey (USGS) entered into a partnership with NASA in the early 1970s to assume responsibility for archiving data and distributing data products. In 1972

NASA launched the first in a series of Landsat satellites. Further satellites were launched in 1975, 1978, 1982, 1984, 1993 (radio communication was lost with Landsat 6), and 1999. In September 1985 the Landsat system was commercialized, and the Earth Observation Satellite Company assumed responsibility for its operation under contract to the National Oceanic and Atmospheric Administration, or NOAA (2002). Overall Landsat program management remained the responsibility of NASA, NOAA, and the USGS. Throughout these changes the USGS Earth Resources Observation Systems Data Center retained primary responsibility as the government archive of Landsat data (Landsat, 2002; Landsat 7, 2002).

The World Data Centers created during the International Geophysical Year mentioned earlier have continued operating to provide international access to various types of observational geophysical data (World Data Centers, 2002). Five main World Data Center areas now exist, serving the United States, Europe, Russia, China, Japan, and India. A number of specialized World Data Centers have been set up, for example, for cosmic radiation, glaciology, geomagnetism, oceanography, and seismology. Some national centers now have subsets of their holdings designated as a World Data Center, for example, the National Geophysical Data Center (2002) and the Earth Resources Observation System (2002).

The World Weather Watch was established in 1961–62 by the World Meteorological Organization to improve the global system for meteorological observation and prediction. The Weather Watch is based on the establishment of world and regional centers at strategic locations around the globe. Three World Meteorological Centers were created in Washington, D.C., Melbourne, and Moscow, from where data are distributed to all nations through a global telecommunications system. Data are archived for retrospective retrieval by the World Data Centers and by Regional Data Centers (World Weather Watch, 2002). The National Climatic Data Center at NOAA was stated by BrightPlanet to be the largest of the “deep web” sites identified, estimated at 366 terabytes in 2000 (National Climatic Data Center, 2002; Bergman, 2000).

Among the varieties of earth data are those concerned with the environment. Environmental data and information embrace qualitative or quantitative descriptions of physical, chemical, biological, or ecological characteristics and processes of Earth. This information

includes knowledge about people, plants and animals, fresh and marine waters, surface and subsurface geologic materials and structures, the atmosphere, and ice. It also includes the factors that cause Earth’s physical, chemical, biological, and ecological characteristics to change. This subject area provides examples of major data directories.

Global change studies are concerned with such problems as the changing composition and chemistry of the atmosphere, carbon dioxide in the environment, the global water cycle, and the relations between human activity and natural processes. NASA has set up a Global Change Master Directory, the range of subject fields covered being listed in Table 2 (Global Change Master Directory, 2002).

Integrated Systems

Beyond the single comprehensive Web site and the directory of sites giving access to data, there are developing integrated, distributed systems of databases. Some of these projects are noted below. Science today is increasingly collaborative and multidisciplinary, and it is not unusual for teams to span institutions, states, countries, and continents. E-mail and the Web provide basic mechanisms that allow such groups to work together. But what if they could link their data, computers, sensors, and other resources into a single virtual laboratory? So-called grid technologies seek to make this possible, by providing the protocols, services, and software development kits needed to enable flexible, controlled resource sharing on a large scale (Foster, 2000).

Grid computing concepts were first explored in the 1995 I-WAY experiment, in which high-speed networks were used to connect for a short time high-end resources

Table 2. Global Change Master Directory Subject Areas

Subject	Subtopics Covered
Agriculture	Animal science, forestry, soils, etc.
Atmosphere	Temperature, winds, etc.
Biosphere	Vegetation, wetlands, zoology, etc.
Cryosphere	Sea ice, snow cover, glaciers, etc.
Human dimensions	Health, environmental impacts, etc.
Hydrosphere	Water quality, groundwater, etc.
Land surface	Land use, land cover, soils, etc.
Oceans	Temperature, circulation, coastal processes, etc.
Paleoclimate	Ice cores, tree rings, etc.
Radiance, imagery	Radar, infrared, etc.
Solar physics	Solar activity, sunspots, etc.
Solid earth	Volcanoes, rocks, minerals, etc.

at seventeen sites across North America. Out of this activity grew a number of grid research projects that developed the core technologies for “production” grids in various communities and scientific disciplines. For example, the U.S. National Science Foundation’s National Technology Grid and NASA’s Information Power Grid³ are both creating grid infrastructures to serve university and NASA researchers, respectively.

The Earth Observing System (EOS) of NASA consists of a science component and a data system supporting a coordinated series of polar-orbiting and low-inclination satellites for long-term global observations of the land surface, biosphere, solid Earth, atmosphere, and oceans. EOS aims at providing an improved understanding of Earth as an integrated system. One section of the system is EOSDIS, which provides a gateway to a number of distributed active archive centers. The gateway offers a single Web interface through which users can automatically search for, browse, and order data from various participating archive centers around the globe, including the Landsat data mentioned above (Earth Observing System, 2002). The National Science Foundation is funding the Grid Physics Network, to put in place the distributed computational environments for a number of forthcoming physics experiments, including two at the Large Hadron Collider. Another project, the Particle Physics Data Grid, is also under way to link America’s major particle physics research centers.

In Europe CERN is managing a project to provide access to distributed databases. The project is using two areas of science as a test bed. First is the exploitation of genomes and of the huge influx of data coming from postgenomic experimentation. The DataGrid project’s biology test bed will provide a platform for new algorithms for data mining, new databases, code management, graphical interface tools, and new strategies for computer-based experimentation. The grid technology will facilitate sharing of genomic databases. Second, the European Space Agency manages several missions of earth observation through satellites. These involve the downloading of about 100 gigabytes per day of images, and this number will grow in the near future by a factor of about five after the launch of the ENVISAT satellite in 2001. Some 800 terabytes of data are already stored in mass storage facilities, and this too is set to grow in the future. The grid will guarantee an improved way to access large volumes of data stored in a distributed European-wide archive (EuroGrid, 2002; “DataGrid computing,” 2002).

Access Difficulties

Despite all the work that has gone on to improve access to the data of science, problems remain (Rumble, 2000). The first is continued fragmentation into separate databases, a problem said to be particularly true of some areas of the physical sciences, such as materials science and chemistry. Many small data files exist, with a plethora of formats, a range of quality levels, and inadequate directories to locate them (National Academy of Sciences, 1997). CODATA has a task group on the standardization of data files concerned with physicochemical properties. Because of the multitude of existing properties and the variety of modes of presentation, the computer-assisted extraction of the numerical values of such properties from primary sources remains difficult.

The second problem is inconsistent terminology and nomenclature, particularly in biology but also in other subjects. For example, there are no consistent and internationally agreed classification schemes for land-cover vegetation or for soils. CODATA has a commission on standardized terminology for access to biological data banks.

Particularly in the earth and environmental sciences the multidisciplinary of approaches creates access difficulties. Some years ago the International Council for Scientific and Technical Information set up a group on interdisciplinary searching that reported on these problems (Weisgerber, 1993).

Now that there are hundreds—nay, thousands—of different data sources, there is the problem of uniformly describing them and providing access to them through appropriate directories (Nicholls, 2001). Another problem that has been identified is that the generation, holding, and distribution of data is often now the concern of commercial firms, which can have the effect of making access to some data costly or even restricted. Further, legal restraints on access to scientific data of all types are currently being implemented or discussed.

Conclusion

In 1969 the U.S. Committee on Scientific and Technical Communication presented a major report to the National Science Foundation. It made mention of data-collection activities, in some instances dating well back into the nineteenth century, such as those of the major environmental programs (e.g., geology, geodesy, weather, tides, and currents). But its main emphasis was on the production of critically evaluated data compilations. Storage of and access to the raw data of science was not

considered as a separate problem: it was assumed to be covered by the publication of journals and technical reports. Remote sensing had not yet made a major impact. Thirty years later the situation has completely changed. Interest in critical evaluation still exists, of course, but vast emphasis is now given to the preservation and transfer of primary data. The first major interdisciplinary conference on this topic was sponsored in 1997 by the U.S. National Committee of CODATA (CODATA, 1997). This paper has tried to give some indication of this transition.

References

- Astronomical Data Center. (2002). Available: adc.gsfc.nasa.gov (accessed 12 May 2002).
- Bergman, M. K. (2000). The deep web: Surfacing hidden value. Available: www.brightplanet.com/technology/deepweb.asp (accessed 15 November 2000).
- Centre de Données Stellaires. (2000). Simbad astronomical database. Available: simbad.u-strasbg.fr/Simbad (accessed 5 December 2000).
- Chemical Abstracts Registry System. (2002). Available: www.cas.org/EO/regsys.html (accessed 12 May 2002).
- CODATA. (1997). Conference on scientific and technical data exchange and integration, December 1997. Available: www.nas.edu/cpsma/97cont.htm (accessed 12 December 2000).
- CODATA. (2002). Task groups and joint groups. Available: www.codata.org (accessed 12 May 2002).
- DataGrid computing for data intensive science. (2002). Available: www.eu-datagrid.org (accessed 12 May 2002).
- Earth Observing System. (2002). Available: spsun.gsfc.nasa.gov/eosinfo (accessed 12 May 2002).
- Earth Resources Observation System. (2002). Available: edcwww.cr.usgs.gov (accessed 12 May 2002).
- Egret, D., & Albrecht, M. A. (Eds.). (1995). *Information and online data in astronomy*. London: Kluwer. With updated online references available: cdsweb.u-strasbg.fr/data-online.html (accessed 12 May 2002).
- EuroGrid. (2002). Application testbed for European Grid computing. Available: www.eurogrid.org (accessed 12 May 2002).
- Foster, I. (2000). Internet computing and the emerging grid. *Nature Web Matters*, 7 Dec 2000. Available: www.nature.com/nature/webmatters/grid/grid.html (accessed 12 March 2001).
- Gibson, P. J. (2000). *Introductory remote sensing*. London: Routledge.
- Global Change Master Directory. Available: gcmd.gsfc.nasa.gov (accessed 12 May 2002).
- Hey, T., & Trefethen, A. (2003). The data deluge: An e-science perspective. In Berman, F. (Ed.), *Grid computing*. New York: Wiley.
- Hubble Space Telescope. (2002). Images by subject. Available: hubblesite.org/gallery (accessed 12 May 2002).
- Jaschek, C. (1989). *Data in astronomy*. Cambridge: Cambridge University Press.
- Journal of Astronomical Data. (2002). Available: www.vub.ac.be/STER/JAD (accessed 12 May 2002).
- Landsat. (2002). Available: geo.arc.nasa.gov/sge/landsat (accessed 12 May 2002).
- Landsat 7. (2002). Available: landsat.gsfc.nasa.gov (accessed 12 May 2002).
- National Academy of Sciences. (1997). *Bits of power: Issues of global access to scientific data*. Washington, DC: National Academy Press.
- National Center for Biotechnology Information. (2002). Available: www.ncbi.nlm.nih.gov (accessed 12 May 2002).
- National Climatic Data Center. (2002). Available: www.ncdc.noaa.gov (accessed 12 May 2002).
- National Geophysical Data Center. (2002). Available: www.ngdc.noaa.gov (accessed 12 May 2002).
- National Nuclear Data Center. (2000). Available: www.nndc.bnl.gov (accessed 16 November 2000).
- National Oceanic and Atmospheric Administration. (2002). NOAA history: A science odyssey. Available: www.history.noaa.gov (accessed 12 May 2002).
- Nicholls, B. (2001). Digital libraries in the large. *Byte*, 30 July 2001.
- Rossmasler, S. A. (Ed.). (1980). *Data handling for science and technology*. Amsterdam: North Holland.
- Rumble, J. (2000). *Shaping the information revolution for 21st century science and technology*. Available: www.nist.gov/srd/Talks/codata-oct-99.ppt (accessed 12 May 2002).
- Salton, G. (1975). *Dynamic information and library processing*. Englewood Cliffs, NJ: Prentice-Hall.
- Schummer, J. (1997). Scientometric studies on chemistry, I: The exponential growth of chemical substances, 1800–1995. *Scientometrics*, 39, 107–123.
- Short, N. M. (2002). Remote sensing tutorial. Available: rst.gsfc.nasa.gov/Front/overview.html (accessed 12 May 2002).
- Sillince, M., & Sillince, J. A. A. (1993). Sequence and structure databanks in molecular biology. *Journal of Documentation*, 49, 1–28.
- Terraserver. (2002). Available: terraserver.homeadvisor.msn.com (accessed 12 May 2002).
- U.S. Committee on Scientific and Technical Information. (1969). *Scientific and technical communication*. Washington, DC: National Academy of Sciences.
- U.S. Naval Observatory. (2002). Astrographic Catalogue history. Available: ad.usno.navy.mil/proj/AC (accessed 12 May 2002).

- Vickery, B. C. (1999). A century of scientific and technical information. *Journal of Documentation*, 55, 476–527.
- Weisgerber, D. W. (1993). Interdisciplinary searching: Problems and suggested remedies. *Journal of Documentation*, 49, 231–254.
- World Data Centers. (2002). Available: www.ngdc.noaa.gov (accessed 12 May 2002).
- World Meteorological Organization. (2000). Natural disaster reduction in the twenty-first century: Science and technology can make the difference. Available: www.wmo.ch/web/Press/Press637.html (accessed 12 May 2002).
- World Weather Watch. (2002). Available: www.wmo.ch/web/www (accessed 12 May 2002).