Appendix D

Digital Data Collections by Categories

INTRODUCTION

Digital data collections vary greatly in size, scope, usage, planned duration, and other dimensions. We distinguish between three functional categories of data collections: (1) research database collections, which are specific to a single investigator or research project; (2) resource or community database collections, which are intermediate in duration, standardization, and community of users; and (3) reference collections, which are managed for long-term use by many users. The following sections provide descriptions and examples of each of these types of digital data collections.

It should be noted that there are not always clear distinctions between these categories: data collections for large research projects overlap with community database collections, and many community data collections transition to become reference data collections. These categories are based on functional attributes of the collection rather than location or size of the data set, and some data centers support all three kinds of collections.

RESEARCH DATABASE COLLECTIONS

Description

Research database collections are the products of one or a few focused research projects. The collections may vary greatly in size, but are intended to serve a specific group, often limited to immediate participants. These collections have relatively small budgets and may be supported directly or indirectly, often through the research grants supporting the project that they serve. Funding is assured for only a short period of time. They typically contain data that is subject to limited processing or curation, and may or may not conform to community standards (e.g. standards for file formats, metadata structure and content, access policies, etc.). Often, applicable standards may be limited or rudimentary as the data types may be novel and the size of the user community may be small. The collection may not be intended to persist beyond the end of the project. Some research collections are accessible to the public through the Web, but many are not, and many of the Web links to research collections are ephemeral.

Examples

There are many thousands of research databases, and they are highly variable in size, number of users, consistency of data and metadata format, duration, and other attributes.

In the Earth Sciences, many research data sets result from field-based research projects.   Examples of data sets available on the web from recent field programs can be found at http://www.atd.ucar.edu/atd_data.html. A specific example is the data collection from the Fluxes Over Snow Surfaces (FLOSS) project, which is studying the surface meteorology of snow-covered rangeland in Colorado. This collection includes data from a wide variety of project measurement instruments http://www.atd.ucar.edu/rtf/projects/FLOSS/. Many research databases in the earth sciences use well-established file format and structures that conform to the requirements of major data systems funded by NSF or other agencies, such NOAA or NASA.

An example of a biology research data collection is the Ares Lab yeast intron database.   This site contains information and analyses about many specific segments of the genome of the yeast Saccharomyces cerevisiae. It was created and managed by a group that includes biologists and bioinformatics specialists. It is available at http://www.cse.ucsc.edu/research/compbio/yeast_introns.html.

In economics, some research data collections result from laboratory experiments. An example is NSF-funded research at the University of Virginia and collaborating colleges that collects data via online game-like programs. The project website contains computer programs and a data base of experimental results that can be further analyzed. Examples of these can be found in links from http://www.people.virginia.edu/~cah2k/research.html. Many other empirical economics projects create new datasets based on the compilation and analysis of economic, industrial, and behavior data. In many cases the project data collections are not available on the web, but may be available to other researchers from the author.

RESOURCE OR COMMUNITY DATA COLLECTIONS

Description

Resource or community data collections serve a specific science and engineering community.   They are typically between research and reference data collections in size, scale, funding, community of users, and duration. They typically conform to community standards, where such standards exist. Often these digital collections can play key roles in bringing communities together to develop appropriate standards where a need exists. In many cases community database collections migrate to reference collections.   In some fields, such as biology, resource data collections are often separate, directly funded projects. In other areas, such as the earth and environmental sciences, resource database collections are often managed under the umbrella of a data center that also supports research and reference databases.

Examples

Examples of resource data collections in the biological sciences include:

- The Arabidopsis Information Resource (TAIR) http://www.arabidopsis.org/ is managed by an organization that involves 20 developers (programmers and curators) and serves about 13,000 registered users and 5,000 laboratories. In early 2004, the collection contained around 3 gigabits of actual data and 16 gigabits for indexes for searching and analyzing data. The data are available to the public. Their continued availability depends on the duration of the project.
- PlasmoDB is a community data collection for the study of genomics of the malaria parasite Plasmodium. Researchers can view genomic data, obtain detailed information about individual genes, and access tools to facilitate analysis. http://www.plasmodb.org/bdbs.shtml.
- The Maize Genetics and Genomics Database (MaizeGDB) provides a similar set of databases and tools for maize research. MaizeGDB is funded by a cooperative agreement through the USDA Agricultural Research Service. http://www.maizegdb.org/.
- The Canopy Database Project supports data acquisition, management, analysis and exchange relating to forest canopy studies at all stages of the research process. It develops informatics tools, documents and publishes datasets that demonstrate use of these tools, characterizes fundamental structures of the forest canopy, and relates those structures to functional characterizations for retrospective, comparative, and integrative studies. http://canopy.evergreen.edu/home.asp

An example of a community database in the physical sciences is the LIGO Scientific Collaboration (LSC), which is a community resource for organizing technical and scientific research in the Laser Interferometer Gravitational Wave Observatory (LIGO). Around 500 scientists are involved in the collaboration. Access to the data is available only to members of the LSC, but the LSC is open to all scientists who apply and who propose an acceptable research plan – no groups have been rejected. LIGO data are characterized by very small signals buried in large amounts of instrument noise, and data are analyzed by internal teams consisting of instrument experts teamed with analysis experts. http://ligo.org.

In the earth and space sciences, many resource databases are housed within larger data centers that contain a combination of research, resource, and reference databases. For example the University Corporation for Atmospheric

Research (UCAR), which is jointly funded by NSF and NOAA, operates the Joint Office of Science Support, which provides scientific, technical, and administrative support services to help the research community plan, organize, and implement research programs and associated field projects. Its CODIAC data management system offers scientists access to research and operational geophysical data. It maintains data archives and provides data support for current projects and field programs, including aircraft data, ground radars, and satellite photos. http://www.ofps.ucar.edu/codiac/.

NASA's Earth Science Enterprise (ESE) has ten discipline-specific data centers, known as Distributed Active Archive Centers (DAACs) that process, archive, document, and distribute data from NASA's Earth observing satellites and field measurement programs. Each data center has its own data-delivery methods and data-analysis tools. Most contain a combination of resource and reference data collections. Data can be accessed through http://nasadaacs.eos.nasa.gov/search.html. Examples of these distributed active archives include:

- The Alaska Satellite Facility (ASF) DAAC at the University of Alaska, Fairbanks, operates under contract to NASA to acquire, process, archive, and distribute satellite Synthetic Aperture Radar (SAR) data for the U.S. government and research communities. The ASF DAAC archives both restricted and unrestricted data. Restricted data are available only to registered and approved users while unrestricted data are available to the general public.
  http://www.asf.alaska.edu/.
- The DAAC at Goddard Space Flight Center manages data related to the upper atmosphere, atmospheric dynamics, global precipitation, global biosphere, ocean biology, ocean dynamics, solar irradiance.
  http://daac.gsfc.nasa.gov/www/.

Another resource data collection is the Ocean Drilling Program database managed at Texas A&M University. The Ocean Drilling Program is supported by NSF and 22 international partners. It contains data relating to decades of ocean drilling. http://www-odp.tamu.edu/database/

REFERENCE COLLECTIONS

Description

Reference collections are intended to serve large segments of the general scientific and education community. Conformance to robust and comprehensive standards is essential to provide the diverse user access and impact that are the mission of these collections. Adoption of standards by reference collections often 'sets the bar' for a large segment of the community, effectively creating a 'universal' standard. Budgets are often large, reflecting the scope of the collection and breadth of impact, and are typically provided by long term, direct support from one or more funding sources.

Examples

Examples of biological reference data collections include:

- The Protein Data Bank, which serves as the authoritative, international repository for macromolecular structure information. This collection was first created more than 30 years ago and its activities are currently supported by a coalition of eight U.S. agencies. http://www.pdb.org
- Uniprot - the Universal Protein Resource, is the world's most comprehensive catalog of information on proteins. The UniProt Archive (UniParc) is a comprehensive repository, reflecting the history of all protein sequences. The UniProt Consortium is comprised of the U.K.-based European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB), and the U.S.-based Protein Information Resource (PIR). UniProt is supported, in part, by the National Institutes of Health and by the European Union. http://www.pir.uniprot.org/

Examples of space science reference data collections include:

- The SIMBAD astronomical database housed at the Centre de Données Astronomiques de Strasbourg in France. It provides basic data, cross-identifications and bibliography for astronomical objects outside the solar system. On October 1, 2004, Simbad contained over 3 million objects, 8.7 million identifiers, and nearly 15,000 bibliographical references. http://simbad.u-strasbg.fr/Simbad
- The National Space Science Data Center (NSSDC) serves as the permanent archive for NASA space science mission data, and includes data on astronomy and astrophysics, solar and space plasma physics, and planetary and lunar science. NSSDC archives about 20 TB of digital data from about 420 mostly-NASA space science spacecraft, of which the most current 3 TB are electronically accessible. In addition to serving as the permanent archive, NSSDC also serves as NASA's primary active archive for space physics mission data and for long-wavelength data (IR, etc.) from selected NASA astrophysics missions. It provides access to several geophysical models and to data from some non-NASA mission data. NSSDC also supports several public-interest web-based services that provide, for examples photo images of interest to the public. http://nssdc.gsfc.nasa.gov/

An example of a physical sciences reference data collection is the Physical Reference Data at the National Institutes of Standards and Technology. This collection contains high quality reference data on physical constants, atomic and molecular data, spectroscopy, and other areas. http://physics.nist.gov/PhysRefData/contents.html.

Examples of geoscience reference data collections include the reference datasets managed by the National Center for Atmospheric Research (NCAR). These includes hundreds of atmospheric, oceanographic, and geophysical datasets. As noted previously, some of these are research or community datasets, but evolve to become reference datasets over time. These can be accessed through http://dss.ucar.edu/. A specific example of a reference dataset at NCAR is the Re-analysis project which was carried out jointly with the European Center for Medium Range Forecasting. This project used the latest atmospheric global models and previously collected data (decades back in time) to derive past atmospheric circulation patterns. These are essential data sets for understanding how the atmosphere is changing and how well the simulation models can re-create the "observed" atmosphere. These data are accessible at http://dss.ucar.edu/pub/reanalyses.html

Examples of social science reference data collections include:

- SEDAC, the Socioeconomic Data and Applications Center, which is one of the Distributed Active Archive Centers (DAACs) in the Earth Observing System Data and Information System (EOSDIS) of the U.S. National Aeronautics and Space Administration. SEDAC focuses on human interactions in the environment. Its mission is to develop and operate applications that support the integration of socioeconomic and Earth science data and to serve as an "Information Gateway" between the Earth and social sciences. http://sedac.ciesin.columbia.edu/data.html
- The reference datasets from the Panel Study of Income Dynamics (PSID) conducted at the Survey Research Center, Institute for Social Research, University of Michigan. PSID, begun in 1968, is a longitudinal study of a representative sample of U.S. individuals (men, women, and children) and the family units in which they reside. The sample size has grown from 4,800 families in 1968 to more than 7,000 families in 2001. At the end of 2003, PSID had collected information about more than 65,000 individuals spanning as much as 36 years of their lives. In the last five years, more than 290 journal articles and 70 Ph.D. dissertations were based on the PSID. PSID datasets include public release data files that have been processed and edited, and are available to all users. Other PSID datasets are still undergoing active processing and revision by the project team and others, and would be considered to be research or community datasets. http://psidonline.isr.umich.edu/