Latin American Electronic Data Archive

by Kent Norsworthy

In October 2009, LLILAS and LANIC launched the Latin American Electronic Data Archive (LAEDA, http://lanic.utexas.edu/laeda/) project under a four-year U.S. Department of Education Technological Innovation and Cooperation for Foreign Information Access (TICFIA) grant. TICFIA provides funds for the development of innovative techniques or programs that address national teaching and research needs in international education and foreign languages.

Our LAEDA project seeks to collect, preserve, and provide access to data sets relevant to Latin American research, policy analysis, and teaching. The focus of the collection is on electoral data, household surveys, and data relevant to social policy evaluation. The original idea for LAEDA grew out of requests by Latin American Studies faculty for systematic access to data sets for research and teaching. In approaching the challenge of data set acquisition, delivery, and preservation, the project draws on multiple collaborators to contribute data sets, expertise, and technical capacity for project development.

The amount of statistical data produced in Latin America has grown exponentially in recent years. Many government agencies—census bureaus, national statistics agencies, development ministries, electoral institutes, etc.—now make some of this data available via the Web or CD-ROM. A variety of international institutions, private firms, scholars, and nongovernmental organizations also collect and disseminate statistical data, from surveys of public opinion to social development indicators. The increased production of electronic data in Latin America has been a welcome development for social scientists, both because it has expanded the amount of data potentially available to them and because it has made the data available in a format that can be easily manipulated and analyzed with statistical programs like Stata or SPSS.

Unfortunately, as a practical matter, there is limited access to this data for researchers. Most institutions place only a fraction of the electronic data they produce on their Web sites, requiring users to purchase the data on CD-ROM or to attempt to obtain it directly through special requests. Moreover, both public and private institutions often remove data from their Web sites after a relatively short period of time. Important "data gaps" now exist, including cases where electronic data has evidently been lost forever. For example, electoral data from Ecuador and Guatemala in the 1980s is no longer available at all in electronic format from the producing institutes in these countries. In cases like these, researchers are forced to rely on summary or aggregate data, typically in paper format only, which represents a fraction of what was produced originally in electronic format.

LAEDA is a joint project of LANIC, LLILAS, and the University of Texas Libraries. Our institutional partner is Mexico's Secretaría de Educación Pública (SEP). Several individual faculty members are working with the project from the UT Government and Sociology departments, as well as UT's Population Research Center. A LAEDA Advisory Board has been established to provide input for focused acquisition of data sets and to ensure the comprehensiveness and relevance of the LAEDA holdings (see sidebar).

Building and maintaining a data archive for microdata revolves around three core activities: data acquisition, preservation, and distribution. Project design encompasses acquisition, preservation, documentation and metadata, removal of personal identifiers, and access and distribution.

The LAEDA data acquisition process is fundamentally driven by a data collection policy that specifies the geographic and topical scope and the objectives of the archive. The policy also includes a list of criteria that will be used to determine the value and appropriateness of data sets for potential inclusion in LAEDA, including items such as geographic coverage, sample size, prior removal of personal identifiers, etc.

LAEDA encourages researchers to contribute copies of appropriate data sets they may have to ensure their long-term preservation and to provide enhanced access. LAEDA will also seek agreements with national electoral authorities and statistics agencies in Latin America for ongoing acquisition of data sets. Issues we are addressing as part of this process include confidentiality, intellectual property rights, and preservation.

Assuring the long-term preservation of valuable data sets is an essential function of LAEDA. Regardless of how well a survey may have been constructed and carried out, the resulting data sets are inherently fragile and subject to numerous threats. Factors such as natural disaster, human error, and deterioration of the bit streams and the media that hold them can lead to the corruption of data files or outright loss. LAEDA's preservation strategy works at two levels. The archive itself is being

built on the basis of a robust, standards-based data management plan that seeks to guarantee the physical integrity of the data across time, including a backup plan for disaster mitigation. Additionally, the entire contents of LAEDA will be ingested into the University of Texas Digital Repository, a preservation framework with strategies built upon the Open Archival Information System reference model.

The utility to researchers of a given data set is driven not only by the quality of the data itself, but also by the quality of the accompanying documentation. For those data sets deemed to have insufficient or incomplete documentation upon ingest, LAEDA staff will work to acquire or produce the necessary documentation. This includes items such as a description of the data collection techniques, sample design, copies of questionnaires used, coding instructions and classifications, as well as information on confidentiality and removal of personal identifiers.

Upon ingest, each data set in LAEDA will have a metadata record associated with it. Metadata will be used to enhance preservation and to facilitate discovery by users. Using the Dublin Core element set, metadata in LAEDA will include items for each data set such as title, author, date, geographic coverage, language, description of the survey, etc. Both the data documentation and the metadata schemes will rely upon established standards and best practices such as the Data Documentation Initiative (DDI) and Dublin Core (DC).

LAEDA is developing an official data distribution policy that establishes the terms and conditions of use for data hosted in the archive. The policy will be closely articulated to a tiered Web-based access system that will allow for different access levels depending on the nature of the data and the affiliation of the end user. Also under development is our central repository for data sets on Latin America, which will provide added granularity and depth of research materials. The ability to download and manipulate original data for analysis will allow scholars and students to reframe research questions and carry out quantitative as well as qualitative analysis.

In terms of the electoral data, the subnational sources we are collecting are of particular importance for two reasons. First, unlike the national level data that are generally widely available, electoral results at the level of electoral district or municipality are especially subject to the systemic shortcomings of collection, accessibility, and dissemination discussed above. Second, subnational data are increasingly important as teaching, scholarship, and policy making have evolved analytically to focus on attributes of national institutional arrangements, party systems, and parties that can only be studied with detailed subnational data. For example, fundamental concepts such as the nationalization of parties and party systems are built upon the measurement of electoral support at the level of electoral district. With LAEDA providing comprehensive sources of subnational electoral data, research on Latin American party systems will be better able to keep pace with research on advanced democracies and, more important, tackle features of national political systems that are distinctive to the democracies in the region.

Since the project launched in October 2009, LAEDA staff has focused significant efforts on determining and defining best practices across a variety of disciplines for data archive dissemination. We have conducted research to document exemplary user interfaces, services offered, site features, database tools, etc. Two directories of resources have been compiled to be used as a point of reference for current standards and best practices and for input into development of the LAEDA backend system.

Initial acquisition activities are under way, and to date we have gathered municipal and provincial electoral data sets from Bolivia, Chile, Ecuador, Guatemala, Peru, and Venezuela covering nearly 100 electoral exercises between 1958 and 2006. Project staff has carried out extensive work on each of the data sets gathered to date. This process has included activities such as standardization of variable naming practices, a thorough check of each data set to detect anomalies in the data or in the labeling scheme, and updating of certain variable labels.

We also are gathering supporting documentation and supplementary information for each data set, including things like glossaries, a description of the original data sources, etc. We also have begun Web archiving activities in order to gather copies of electoral summary data, as well as election rules and regulations, currently hosted on the official Latin American electoral tribunal public Web sites.

To contribute data sets to LAEDA, please contact c.palaima@austin.utexas.edu

---

Kent Norsworthy is LANIC Content Director.