THE SOCIAL SCIENCES DATABASE ARCHIVE:
The Purpose, Design, and Implementation

Academic Technology Support
California State University, Los Angeles

Introduction

The Social Sciences Database Archive (SSDBA) is a CSU Specialty Center in development at Cal State L.A. The SSDBA will provide support to the political and social sciences system-wide . The principal resources associated with the SSDBA are the databases available through the Inter-university Consortium for Political and Social Research (ICPSR). The ICPSR is a consortium based at the University of Michigan which collects and maintains an archive of thousands of databases useful for political and social research. It is the ICPSR data sets, along with the Census data and the Roper and Field services databases which are commonly referred to in the CSU as the "social sciences databases".

As a confederated member of the ICPSR, the CSU has agreed to provide a single-point-of-contact and to request only a single copy of each database in the ICPSR archive. If multiple CSU campuses request a database, it is the responsibility of the CSU confederation to reproduce and distribute the database and associated documentation. By forming a confederation and accepting responsibility for internal distribution, the system realizes a $62,000 annual savings in membership fees.

Prior to the re-organization of the Office of the Chancellor's Division of Information Resources and Technology (IRT), the academic computing staff at IRT had responsibility for preparation and archiving of files, and distribution of the ICPSR datasets was a service provided by the. At the time of the re-organization, the responsibility for providing these services was delegated to the campuses. The SSDBA was proposed as a campus-based alternative to the services offered by the IRT staff. The SSDBA will provide the technical support necessary to maintain the confederation agreement with ICPSR. In that capacity, the SSDBA staff will archive and distribute copies of all tapes and code books received from the ICPSR.

For campuses wishing to provide the services related to the processing and use of these data sets locally, participation in the SSDBA can be limited to only the tape distribution services. There is, however, a significant amount of work

associated with the processing and maintenance of these data sets before they are useful to researchers. The SSDBA and associated services is being developed as a cost effective alternative for campuses which do not have or wish to invest the time of local technical and support staff to process and prepare the social science databases for students and faculty.

Design Goals

To be useful to researchers, the data sets received from the ICPSR must be cataloged so they can be located by title or by variable names. They must be reformatted into the file type used by the central computer so they can be loaded into the computer and then prepared for use with a statistical package. Preparation for use with a statistical package includes deciphering code books (sometimes handwritten), creation of a data dictionary describing each variable, and reformatting the file into the file structure of the statistical application. Very often, researchers are interested in only a subset of a database and/or the database is too large and difficult to work with. In these cases there is further processing which must be done to create subsets. And finally, the files must be reformatted into an "export file" so they can be moved to and used on local campus microcomputers and timeshare machines.

The goal for the SSDBA is to provide on-line access to all of the databases in the archive. The intent is to provide an easily negotiated environment, designed to make it possible for students and faculty to locate and retrieve materials without the intervention of technical and support staff. A combination of large scale mass storage devices, "field-oriented" and "text-oriented" data management technologies, and systems integration tools will be used to create an environment in which researchers can manage their own inquiries and data manipulation tasks. The objective of this system is to reduce the technical barriers and "EDP" delays to the point where the system does not prohibit the active engagement of the researcher's imagination and curiosity with the research materials.

Using the SSDBA

The SSDBA host will be available through both the TCP/IP and asynchronous services of the CSU Net. The SSDBA will be available to students and faculty of the participating CSU campuses as a "public resource." In the target environment, clients will be able to login to the SSDBA using telnet software in a character-oriented environment; search and view documentation; and locate, subset, and transfer data files from the SSDBA host to a local machine without having an account.

A resource which is offered as a "public" resource within the CSU can be managed only via telnet and a TCP/IP ethernet link. For campuses unable to

support telnet access from the workstation level, public access can be supported by telnet from a local timeshare machine. If campuses cannot support telnet access from a host or workstation, asynchronous communications will be available. In this case, the "public" mode of access to the SSDBA will be precluded, and accounts must be issued to individuals. There are several ways this can be accomplished with differing benefits and restrictions. It will be left to the campuses to decide which way will best meet their needs.

In the TCP/IP environment it will be possible to direct print services directly from the SSDBA host to a campus print server. (This will require support of the person(s) responsible for data networks and shared resources at each campus.) Initially, users of the SSDBA Resources will be encouraged to download print files and print at a local printer.

Consulting Services

As stated above, one of the primary design goals for the SSDBA is to provide a system which requires minimal support services for the end-user. The reasoning behind that goal is this - consultation with the faculty at Cal State L.A. led the systems development staff to believe there is a significant unrealized potential for the use of these resources in instruction, and that if the means of accessing and using these databases was improved there would be increased demand. In planning for the implementation of the business and social sciences databases at Cal State L.A., we felt it necessary to provide for the potential demand. But, if realizing the potential of these databases as a network service meant that each additional user required an incremental increase in staff time, it would be impossible to support growth. This was confirmed by the reports of institutions which have developed heavily used network services. Network services must be scalable. If there is anything approaching a linear relationship between utilization and support requirements services cannot be sustained. Thus, the design goal for the system was to provide a scalable resource.

It is anticipated, however, that the staff responsible for maintaining the social science databases will develop a familiarity with the collection of materials and the resources available through the ICPSR, Roper, and Field services beyond that of most faculty. And, it will be desirable to have this expertise available for faculty consultation. Consequently, the SSDBA services are structured to give campuses the option of subscribing for on-line services and providing consulting services locally or subscribing for consulting services provided by the Network Information Services Group (NISG) at Cal State L.A. If a campus opts to subscribe for SSDBA Consulting Services, a Research Analyst with a thorough understanding of the databases and statistical packages will be available to their faculty. To facilitate the consulting services, the Research Analyst will be able to create permanent accounts with adequate storage space for clients. The

Research Analyst will also be available for training, the development of documentation and instructional materials, and research activities related to the use of the SSDBA.

Campuses wishing to subscribe to the on-line services, but not the consulting services, will be expected to designate a local user liaison. The campus user liaison will be the single point of contact for questions regarding the SSDBA on-line services. Problems reported by users at the campus should go through the campus user liaison. The campus liaison will have access to the technical staff of the SSDBA to support the consulting services provided locally. In addition to providing technical support for the campus liaison, SSDBA support staff will provide the campus liaison with support materials. They will prepare and distribute student lab hand-outs to facilitate access to the SSDBA, publish a newsletter to keep students and faculty up to date on the databases and functionality of the system, and publish a User Guide to the SSDBA .

System Specifications

The computer which will be used is a Sun Microsystems SparcServer 470. This is a UNIX based system. The SparcServer was designed for client/server and timeshare applications. The SparcServer 470 is a 22MIP single processor RISC based system. The system architecture is designed for parallel processing and can be upgraded to a four processor system expandable to 672MBytes of memory should the demand exceed expectations. The databases will be stored on a "laser jukebox" - a large scale storage device. The jukebox has two disk drives and fifty-one removable platters. It will hold 34GBytes of information and presents itself to the SparcServer as a standard UNIX file structure. (This system is currently used by the Jet Propulsion Laboratories (JPL) to store images from the space probes.) The "laser jukebox" is also expandable.

The SSDBA environment will include several integrated database management tools to aid researchers in locating databases which are relevant to their studies. The ICPSR Catalog, a one thousand page document with abstracts describing the databases available through that agency, will be loaded into the TOPIC document management system. TOPIC is a search tool used to query text databases. It provides standard search strategies (boolean logic) as well as support for "expert system" profiles (see attachment) and it supports a "transparent" bridge into relational database management systems (RDBMS). Ingres is the RDBMS used on the CSLA servers. The integration of TOPIC for "text-oriented" information and Ingres for "field-oriented" information will make it possible to "browse" the contents of the ICPSR Catalog and a relational database developed by SSDBA staff describing each database variables in a single environment.

The UNIX operating system was designed at Bell Laboratories by systems development staff for systems development. For the last decade, the Department of Defense and the National Science Foundation have worked through grants to leading computer science departments such as Berkeley and Carnegie Mellon to enhance UNIX and make it a powerful tool for systems integration projects. (Ingres was developed at Berkeley as part of a UNIX extension.) The UNIX development utilities, Ingres, TOPIC and the major statistical applications will be used to develop a simple, intuitive work environment for social science researchers.

Funding

The Cal State L.A. proposal to provide system-wide access to the social science databases is subscription based. Campuses wishing to use the SSDBA On-line Services and/or the SSDBA Consulting Services are being asked to pay a subscription fee for each of the services. In the first year of operation participating campuses to pay a one-time start-up fee (TABLE 1.) to establish the hardware and software systems to support the SSDBA. The resources required for start-up are

	FTEF	Annual Fee
Small Campus	Under 600	$5,650
Medium Campus	600 to 1,000	$8,500
Large Campus	1,000 and Above	$11,350

TABLE 1. SCHEDULE OF FEES FOR SSDBA **START-UP**

fixed, they cannot be reduced according to the number of participants. But, the CPU requirements for the operation can be sized, and Cal State L.A. is willing to assume a larger portion of the cost for this resource based on participation. At eighteen campuses participating, Cal State L.A. will assume one third of the cost and re-direct that resource toward another project. At lower levels of participation, Cal State L.A. will purchase a larger share of the CPU resource for other campus uses. The following are the EDP resources which will be made available for the SSDBA:

1. up to two thirds of the CPU capacity of a SUN SPARCserver 470,
2. 500MBytes of volatile disk space for temporary workspace,
3. 2GBytes of volatile disk space for consulting services,
4. 21 GBytes of WORM disk space,
5. intelligent UPS support for the above
6. bridges and gateways necessary for connection to the CSUNet,

7. relational database management system,
8. document management system, and
9. BMDP, Minitab, SAS, SPSS statistical software.

Since the cost of operations for the SSDBA and the technical staff requirements and support service for the campus liaisons are scalable, it will be possible to offer a fixed schedule of fees for the On-line Services. The fee for access to the SSDBA On-line Services and support services for the liaisons are based on the relative size of the participating campus. The range and provision of these services (TABLE 2.) will not be affected by the level of participation. At lower levels of participation the technical support staff will be re-directed to other activities.

	FTEF	Annual Fee
Small Campus	Under 600	$4,750
Medium Campus	600 to 1,000	$7,000
Large Campus	1,000 and Above	$9,500

TABLE 2. SCHEDULE OF FEES FOR SSDBA ON-LINE SERVICES

As explained above, the consulting services for user support are not scalable. The staffing requirements will increase with the number of participating campuses. To accommodate that variability, and to still offer a fixed schedule of fees for the SSDBA Consulting Services, it was proposed that each participating campus receive a share of the SSDBA Research Analyst's time. The cost of the service

	Hours Per Week	Hours Per Year	Annual Fee
Small Campus	1.48	39	$4,000
Medium Campus	2.22	58	$6,000
Large Campus	2.96	77	$8,000

TABLE 3. SCHEDULE OF FEES FOR SSDBA CONSULTING SERVICES

and the amount of time a campus is entitled to will be based on the size of the campus (TABLE 3). If participation is at the highest expected level, which is eighteen (18) campuses, it will require a full position to meet demand. If participation is lower, the surplus time of the Research Analyst will be re-directed to campus services.

As stated above, campuses wishing to limit their involvement in the SSDBA to ICPSR tape distribution only will be able to do so by participating in the cost of the confederated membership and distribution materials only. The confederation membership fee was not affected by the re-organization of IRT. The cost of distribution materials will be handled on an item-by-item basis. Table 4. is a summary of the fees which will be associated with the SSDBA Start-up, On-line Services, and Consulting Services.

	LARGE	MEDIUM	SMALL
Start-up costs	$11,350	$8,500	$5,650
Cost of SSDBA On-line Service	$9,500	$7,000	$4,750
Cost of SSDBA Consulting Services	$8,000	$6,000	$4,000
First Year - Start-up, On-line & Consulting Services	$28,850	$21,500	$14,400
Ongoing Annual - On-line & Consulting Services	$17,500	$13,000	$8,750

TABLE 4. **SCHEDULE OF FEES FOR SOCIAL SCIENCE DATABASE ARCHIVE**

Campus Participation

The SSDBA will be operated as a CSUNet service by the Office of Academic Technology Support (ATS) at California State University, Los Angeles. The Director of ATS will be responsible for the management and operation of that service. All campuses, including those which will limit their participation to ICPSR confederation membership only, will be asked to designate a Campus Liaison from their user services organization to coordinate distribution of materials, services, etc. with the SSDBA staff. For campuses subscribing to SSDBA services, the Campus Liaison for each member institution will act as liaison between the "user services" organization at the home campus and the SSDBA administrative and technical support staff and will coordinate user services between the member institutions and the SSDBA.

To ensure quality support to faculty and students, the Social Sciences Instruction and Research Council (SSIRC) has tentatively agreed to act as a User Advisory Committee to the SSDBA. This group will advise the ATS Director on priorities for database processing, development of utilities and on-line services, training, and documentation. It is expected that the Campus Liaison will also coordinate campus requests with the SSIRC representative on each campus.

In addition to the designation of a Campus Liaison, campuses subscribing to services will be asked to designate locally a Campus Representative to a SSDBA

Policy Advisory Committee. The Policy Advisory Committee will be responsible for advising the Director of ATS on general policies, fiscal issues, and in defining areas for possible special project funding.

Implementation

The implementation of the SSDBA project is being done in two phases. In the first phase, the target is to provide access to 50-100 of the most frequently accessed databases and their codebooks using the same (or comparable) search utilities currently used on the Central CYBER. The objective is to ensure that the services currently available are in place as quickly as possible. Development on this phase has already begun and participating campuses will be expected to make payment through the Office of the Chancellor as soon as FY1991-1992 budgets are received.

As soon as we have completed the first phase, the second phase of the project will begin. This will be the implementation of the target environment described in the sections above. It is estimated that this phase of the project will last eighteen months to twenty-four months. Deliverable modules will be made available as developed. Additional databases will be made available continuously after the start up date until all the databases have been installed.

Contacts

If there are technical or administrative questions regarding the SSDBA, please contact one of the following:

Project Leader
Don Carder
(213) 343-4530
dcarder@atss.calstatela.edu

Business Manager
Nancy Kudo-Hombo
(213) 343-4530
nkuhodo@atss.calstatela.edu

Supervisor, Network Information Services Group
Janet Valade
(213) 343-2575
jvalade@atss.calstatela.edu

FILE:BROCHURE.DOC REV:43
CREATED:3/24/91 BY:D. CARDER
PRINTED:4/15/91 9:58 AM