

**TOWARDS A UNIFIED DATA AND METADATA SYSTEM AT THE CENSUS
BUREAU**

**by
Bo Sundgren
Statistics Sweden
and
Martin V. Appel, Daniel W. Gillman, William P. LaPlant, Jr.
U. S. Bureau of the Census**

ABSTRACT

This report is adapted from "Towards a Unified Data and Metadata System at the Australian Bureau of Statistics - final report", written by Bo Sundgren, dated 1991-12-02. Dr. Sundgren gave his permission to have his original document modified.

The purpose of this paper is to serve as a "Straw Man" document, elicit ideas, and lead to a draft plan for developing a unified data/metadata system at the Census Bureau.

This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.

I. INTRODUCTION

The purpose of this paper is to serve as a "Straw Man" to generate a dialogue, elicit ideas, and lead to a draft plan for developing a unified data/metadata system at the Bureau of the Census (BOC).

For the past few years the BOC has been striving to modernize its data collection, data processing, and data dissemination practices. These efforts have been undertaken as part of the National Performance Review, the CASIC effort, or as part of the BOC's ongoing census or survey redesign efforts. Work is also ongoing in the development of an open systems architecture and the defining of standard practices such as data archiving, electronic mail, and database documentation among others. A Reinvention Lab was established to look at the post data collection activities, and it produced a vision of how it thought the Bureau should operate. There is also work underway to develop a CATI/CAPI tracking system, to disseminate data via the National Information Infrastructure (Internet), to redesign the AGR/ECON census, and the development of a statistical metadata standard and a metadata repository. The above items are only a

partial list of the changes that BOC is undergoing; these are given to make the reader aware of the many efforts that are underway to improve BOC practices. It appears to the authors that many of the components of the Integrated Processing System (IPS) described in the Reinvention Lab's report are currently being worked on in some fashion, but, except for the area of data collection, there is still too little coordination or cross pollination of ideas among the different groups involved.

The authors are not proposing the initiation of a large unified project that would result in building the IPS as defined by the Reinvention Lab, but rather the adoption of that document as the philosophical under-pinnings of BOC processing. The reality would be that as activities are redesigned, they would be required to provide standard access "doors" and standard output formats. If developers want to develop project specific input and output methods, that would be their prerogative, but at least other Bureau groups would have the option of using the latest designs or software routines, or in the case of a rapid development effort, a survey manager could select from many pre-test components. With the constant rapid changes in methodology and technology this incremental approach is superior to trying to define and build a complete and unified system at once.

The Census Bureau has a number of dispersed survey, census, and infrastructure redesign activities underway. This document proposes a strategy for providing links between the data associated with these activities. Much of this report is adapted from Sundgren, 1991C. Dr. Sundgren gave his permission to have his original document modified. The document also incorporates much of the work of the Reinvention Lab and the proposal for a BOC metadata standard and repository.

II. BACKGROUND CONSIDERATIONS: BOC METADATA AND METADATA HOLDINGS

A. Statistical Metadata Definition¹

Statistical metadata is descriptive information or documentation about statistical data, i.e. microdata, macrodata, or metadata. Statistical metadata facilitates sharing, querying, and understanding of statistical data over the lifetime of the data.

The three types of statistical data (electronic or otherwise) are described as follows:

Microdata - data on the characteristics of units of a population, such as individuals, households or establishments, collected by a census, survey, or experiment.

Macrodata - data derived from microdata by statistics on groups or aggregates, such as counts, means, or frequencies.

Metadata - data which describe the microdata, macrodata or other metadata.

The extensive nature of statistical metadata lends itself to categorization into three components or levels:

Systems - the information about the physical characteristics of the application's data set(s), such as location, record layout, database schemas, media, size, etc;

Applications - the descriptive information about the application's products and procedures, such as sample designs, questionnaires, software, variable definitions, edit specifications, etc;

Administrative - the management information, such as budgets, costs, schedules, etc.

The systems, applications, and administrative components help to differentiate the sources and uses of statistical metadata.

B. Basic Purposes of Metadata and Metadata Systems in a Statistical Office

¹ Defined with assistance from participants at Statistical Metadata Workshop, November 14-15, 1995, Bureau of Labor Statistics, Washington, DC.

Statistical metadata and metadata systems have two basic purposes:

- the **end-user oriented purpose**: to support potential users of statistical information; and
- the **production oriented purpose**: to support the planning, operation, and evaluation of statistical production systems.

A potential end-user of statistical information needs to:

- identify,
- locate,
- retrieve,
- process,
- interpret, and
- analyze

statistical data that may be relevant for a task that the user has at hand.

The end-user's task typically belongs to one of the following types of activities:

- problem-solving,
- decision-making,
- planning/monitoring/evaluation,
- research.

This type of user of statistical information and metainformation is typically an **external user**, working outside the statistical office. However, it could also be a person working in the statistical office.

The production-oriented user's tasks belong to the following types of activities:

- planning/design/maintenance,
- implementation/operation, and
- evaluation.

The typical user of production oriented statistical metadata and metadata systems is an internal user working inside the statistical office itself or within a sponsoring agency, possibly in a managerial position.

Production oriented statistical metadata and metadata systems can also be used for supporting individual statistical surveys or systems of such surveys.

A **system of statistical surveys** is a set of related statistical surveys. **Statistical coordination** of definitions, classifications, etc. can be a vital function of a metadata

system.

C. **Input-Orientation and Output-Orientation in Statistical Production**

Most statistical offices are **input-oriented** in the sense that the statistical surveys they conduct or manage are also the natural building-blocks in its organization. The Census Bureau is an example of such a statistical office.

The input-orientation of the Census Bureau is further strengthened by the fact that there are relatively few natural or hard links between different survey programs. Surveys are largely independent operations, and their data are compared and integrated only on the macro level, after the underlying survey cycles are complete or a survey program terminates.

Other features that may stimulate logical and physical integration of survey production systems are the use of common sampling frames and positive (or even negative) coordination of the samples for different surveys.

However, the Census Bureau, like other advanced statistical offices, is working to become more **output-oriented**. Output-oriented means the statistical office focusses on meeting the needs of its customers. The planning and development of output-oriented databases and metadata repositories are steps in this direction.

Output-oriented database systems relate data from different surveys. They need special software and metadata tools for reconciling data from different sources and for helping the users to interpret and analyze the data.

D. **A Data Management Structure for the Census Bureau with Increased Emphasis on Output-Orientation**

An output-oriented database system consists of two major components:

- the **database** component, containing data and metadata; and
- the **user interface** component, containing a user interface, supported by software and (additional) metadata.

The database component of an output-oriented database system could be logically located in at least three different places:

- (a) **output-locally**, the database is associated with the user interface component;

- (b) **input-locally**, the database is associated with each one of the surveys that have to provide the output-oriented database with data and metadata;
- (c) **centrally**, in a common (logically central) Data Library² (DL) which provides all output-oriented systems with the data (and metadata) that they need, and which in turn is more or less continuously updated with data from the underlying surveys.

As the number of output-oriented database systems grows at the Census Bureau, alternative (a) will lead to a very complex systems architecture, with a lot of complex communication (see the upper part of figure 1) and physical duplication of data.

In addition, alternative (a) will lead to costly repetition of (more or less) the same systems development and maintenance activities, and (probably) a lack of uniformity in data and metadata management, which will in turn inevitably lead to a lack of uniformity in the interfaces between the Census Bureau and its information users.

Alternative (b) seems attractive from a theoretical point of view, but it requires a very sophisticated system for distributed database management. It would be surprising if there is a software product that would satisfy Census Bureau requirements, even if kept on a relatively modest level. The technical and conceptual problems involved are overwhelming.

Like alternative (a), alternative (b) implies a complex communication pattern between the output-oriented database systems and the surveys providing the data and metadata; once again the upper part of figure 1 is applicable.

Alternative (c), with a **logically** central DL, minimizes the volume and complexity of the necessary communication between the surveys and the output-oriented systems; cf the lower part of figure 1. If m is the number of surveys, and n is the number of output-oriented system, the maximum number of communication interfaces will be $(m + n)$ with this architecture, instead of $(m * n)$ as with the other alternatives.

The Census Bureau should adopt alternative (c) as the basis for its future data management strategy, and it should initiate and systematically carry out a number of well synchronized activities in order to implement this data management strategy. These

Data Library - for the purpose of simplicity the DL includes databases, metadata repositories, and tools.

activities will be elaborated later.

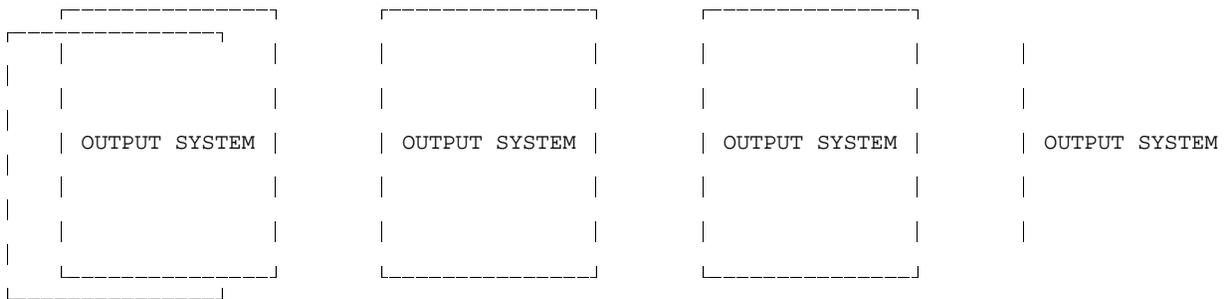
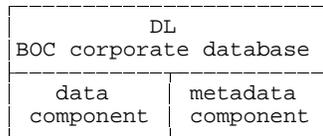
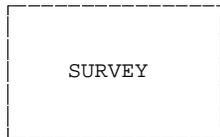
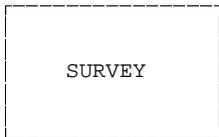
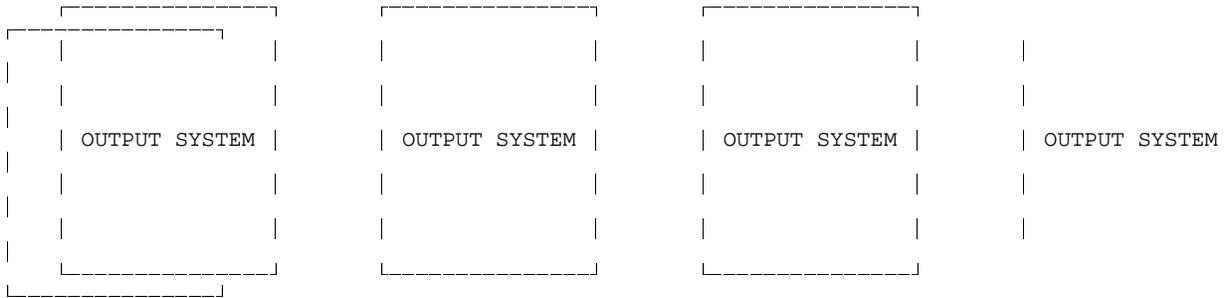
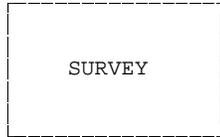


Figure 1. Two alternative architectures for data and metadata management in a statistical office. The architecture in the lower part of the diagram corresponds to alternative (c) in the text. The upper part of the diagram covers both alternative (a) and alternative (b).

E. Some Census Bureau Metadata Tools and Metadata Repositories

The BOC does have some metadata tools and repositories. Most of the systems are primarily end-user oriented, some (StEPS, IPS, etc.) are primarily production oriented, and some (Metadata Repository, etc.) are a mixture of the two. All the systems which are in production are end-user oriented.

Some of the important metadata tools and repositories that exist at the Census Bureau or are under construction are briefly described below. See Appendix A for more information.

Automated Reference Rack (ARRk) is a hypertext based system designed using Lotus SmartText. Telephone clerks at the Census Bureau use ARRk to help the public find appropriate published data through short descriptions of each available file. Descriptions contain information about subjects, coverage, storage medium, cost, and ordering information.

BOX Files is a mechanism for incorporating metadata into ASCII data files (Bean, 1991). The major advantage of this system is that the systems level metadata that describes each file are automatically carried along with the data in a file. The BOX file format is the basis for a recently issued Census Bureau information technology standard for archiving data, although this decision is under review.

Data Extraction System is a general system for extracting data from master data sets into one of several popular data formats, including SAS data sets. It also uses files stored in the BOX format (see above).

Surveys-On-Call is an on-line system for accessing publicly available Survey of Income and Program Participation (SIPP) and Current Population Survey (CPS) data. This system is currently in production and is accessible over the Internet (via the BOC home page) and by modem. Surveys-On-Call is a UNIX based menu system which is based on the Data Extraction System.

Extract is a system in use with CD-ROM products sold by the Census Bureau. Libraries and other public facilities often have these products on the shelf. Extract is a menu driven dBase

application which allows the user to construct extract data files. Metadata is also available and can be appended where necessary in the extracted files.

CENSAS is a project for an automated data and information delivery system based on Decennial Census data for both internal and external customers. Extracts are delivered to the user as SAS datasets to which the user can apply the desired tallies or other statistical analyses. A Beta Test version is available.

CENDATA is an on-line Census Bureau data system containing current and historical data, both demographic and economic. Examples include Foreign Trade Statistics, Quarterly Financial Reports, County Business Patterns, Wholesale and Retail Trade, Center for International Research, Agriculture County, and Manufacturers, Shipments, Inventories, and Orders Survey data.

FERRET (Federal Electronic Research and Review Extraction Tool) is a data extraction tool available on the internet that allows users to find information about monthly demographic survey data using a World Wide Web browser. Users can select microdata or macrodata and download files as SAS data sets or ASCII files.

StEPS (Standard Economic Processing System) is an integrated survey processing system the objective of which is to eliminate redundant processing by combining existing survey systems into one system. The scope of the StEPS system includes access to basic survey processing functions and some additional functions.

IPS (Integrated Processing System) is envisioned to be the umbrella for a compatible set of automated tools to design, conduct, and manage Census Bureau surveys and censuses in an effort to improve cost effectiveness, timely reporting, data quality, and data access. The overall goal is to provide a framework for the integration of generalized processing system components with data collection and other tools.

DADS (Data Access and Dissemination System) is the name for the Census Bureau initiative to develop and implement data access and dissemination focussed on the 2000 Decennial Census and Continuous Measurement data sets, but with the ability to accommodate other data sets having geographic detail, such as those produced from the Economic and Agricultural Censuses. The main objective of DADS is to provide one general (electronic) system for all access to Census Bureau data.

Standard for Survey Design and Statistical Methodology Metadata (SDSM) is a standard under development at the Census Bureau to specify the metadata necessary to describe survey designs, processing, analyses, and data sets completely. Approval for the

standard will be sought through the formal standard development procedures of the Census Bureau, and the process is expected to be completed by September 1996.

Metadata Repository is the project to build a logically central repository of metadata based on the SDSM (see above). The metadata repository is intended to be a source of metadata for all the Census Bureau programs so comparisons of designs, processing, analysis, or data can be made across time and survey programs. A proof-of-concept system has been built.

III. OUTPUT-ORIENTED DATA AND METADATA MANAGEMENT

General Description of a Unified Data and Metadata System at the BOC

A unified data and metadata system at the BOC will be based upon an architecture (c), as specified above, data and metadata standards [], and a corporate DL as a source of data and metadata that are of common interest for the BOC and its external users. Access to the system and the "glue" that ties it together are provided by the **internet** and **intranet** - that part of the internet available internally to the BOC and protected from the outside by a **firewall**.

The DL will be **logically central** in the sense that it will appear to the user to be a single database and it will function as a switch or a clearing-house between the input-oriented survey systems for data collection, etc, and the output-oriented systems for retrieval, analysis, presentation, and distribution of statistical information, emanating from (often) several different survey sources.

Will the DL also be **physically central**? The DL can be physically distributed, as long as all data and metadata can be conceptually interfaced in accordance with certain well-defined standards. Such an architecture would have to be very sophisticated from a technical point of view, but robust, efficient solutions to these problems do exist. Because a physically central DL presents problems which the BOC as an organization will find difficult to overcome, the distributed architecture will have a higher probability of success.

Because of the technical problems in implementing a DL for all BOC survey data and metadata, work should be planned in a step-by-step fashion. A functioning DL could be built around a few surveys, then others could be added as the technology and infrastructure allow.

The DL itself will interact, via data, metadata, and control flows, with other local systems, both on the input and on the output side.

IV. PERSPECTIVES ON A UNIFIED DATA AND METADATA SYSTEM

We shall now look at the unified system from four different perspectives:

- (a) a survey perspective;

- (b) an external user's perspective;
- (c) a management perspective;
- (d) a technical perspective.

A. The Survey Perspective³

On the input side, surveys will have automated production systems (e.g. IPS, StEPS) including a **knowledge base** containing documentation about survey designs, processing, analysis, and data sets. This includes historical information about previous survey designs and a data element registry containing information about all variables.

These survey-based knowledge bases are the ultimate sources of metadata concerning the outputs from the BOC. However, most of the metadata in the survey knowledge bases will probably stay there, but it will be stored in such a way that it can be referred to, and consulted by, the central and output-oriented components of the DL automatically, as well as by means of human intervention.

The knowledge bases of survey production systems will be built up incrementally as a result of operations, maintenance, and evaluation activities during survey re-design.

For example, when a new time version of a periodically repeated survey is going to be designed, the metadata from earlier repetitions of the survey will be available in the knowledge base, and could be used as a starting-point for the design of the new time version of the survey.

Resource sharing and coordination will be highly facilitated by a unified system for BOC data and metadata management. For example, if a new survey is to be designed, it will be natural to look for similar surveys, and the knowledge bases of such surveys, to be taken as a starting point for design activities, and this natural behavior will in itself encourage resource sharing and coordination.

The final outputs of a survey production system will be:

- (a) a set of **final observation data** concerning individual objects (statistical units) within one or more populations

³ Economic Surveys, Demographic Surveys, and Decennial Census are three possible perspectives within BOC.

of objects of interest; the final observation data are **microdata** which have been collected and prepared (coded, edited, etc.), until they have reached a status suitable for archiving and (possibly) later use;

- (b) a set of **estimates of statistical characteristics**, or **parameters**, concerning some population(s) and (sub)domains of objects of interest, that is, **macrodata** that have been produced (by means of aggregations, etc.) from the set of final observation data;
- (c) a set of **metadata describing the microdata and macrodata** with regard to contents, quality, storage, etc.

From the individual survey perspective, the job may seem to be finished when outputs (a), (b), and (c) have been delivered to the corporate DL. However, this is not quite true. The DL will be able to send important information concerning the survey's market and customers back to survey program managers. Every time a user requests some information, the feedback information could be accompanied by "dollars and cents" to make it more tangible and possibly more interesting. Whenever a user request **cannot** be served by DL data and metadata, the DL will report this back, too. Thus a unified system could be used as an instrument in a control process for improving the output and performance of the BOC.

The unified system could give some additional short-term benefits to the survey production systems. For example, the knowledge base component of the DL would contain standard instruments (methods and software) for performing standard operations on observation data and aggregated data that are structured and stored in accordance with corporate standards. Aggregations, tabulations, graphical presentations, and various types of statistical analysis could be examples of such standard operations that would be supported by corporate standard instruments.

B. The External User's Perspective

For an external user the DL will represent the **information potential** of the BOC; the user will view this information potential through the "glasses" of an output-oriented system.

Each output-oriented system may offer a variety of ways for users to locate relevant BOC data. Some of the important methods are:

- WAIS searches via the internet, especially coupled with **GILS** (Government Information Locator Service, FIPS 192) records;

- Hypertext and other key-word searches;
- Thesauruses, taxonomies, or other classification methods;
- Tools based on the SDSM standard, especially using the table of contents outline as a check list (see Appendix A);
- SQL and other database or repository query languages.

Every output-oriented system will provide its particular users with a view of the BOC information potential which is well tailored to the specific needs of the particular users. The tailoring could be accomplished by such things as

- subsetting and structuring the total information potential in a suitable way, focusing on information of high relevance for the particular user group, and leaving out information of little or no relevance;
- providing search ways and search mechanisms that are known to be "natural" for the particular user group;
- providing tools for processing, modelling, and analysis that are known to be of special interest for the particular user group.

The total information potential of BOC will always consist of data of various degrees of availability. Some data are aggregated and public. Such data may be available to any user within seconds, provided they are stored on easily accessible media. Other data may be aggregated and public, but still not so accessible, because they are not requested sufficiently often to motivate the costs of certain types of storage.

Most microdata cannot be released to external users for confidentiality reasons. Such data will have to be greatly modified or aggregated and re-examined after the aggregation before they can be made available to external users.

Regardless of the varying degree of availability of different parts of the BOC information potential, the unified system should - through its metadata component - provide any user with a view of the total information potential, which is as complete as possible with regard to the needs of the particular user. Then, when the user has identified what he or she would be interested to retrieve and process, the processing of the request may proceed a little differently, and take a shorter or longer time, depending upon

- whether the requested data are explicitly stored, or whether

they have to be derived (for example, by means of aggregation from underlying microdata);

- whether the requested data are (or are based upon) public or confidential data;
- where the physical data are actually stored;
- whether all data and metadata needed to serve the request are automatically available from the DL itself, or whether there are references to some other data and metadata sources (for example "survey-local" sources), which have to be resolved by some kind of human intervention.

C. The Management Perspective

We have already noted that managers on all levels may get interesting feedback information from a unified system concerning the demand for different parts (or "non-parts") of the BOC information potential. Thus a unified system will create, at least in theory, new possibilities to get certain non-trivial measurements of the value of different BOC outputs. This is especially interesting since the benefit side has always been the most difficult part of cost/benefit calculations in statistical services and in the public sector in general for that matter.

The unified system could be linked to and, to some extent, integrated with the budgeting and internal cost accounting system of the BOC. How far one should move in that direction is of course a matter of management policy. Anyhow, the "survey-local" production systems can, of course, generate some cost data, and, like other types of metadata emanating from the surveys, they can be communicated to corporate systems, where they can be combined and analyzed.

An DL could also be used for reallocating certain corporate costs - and revenues.

D. The Technical Perspective

The DL should be structured into a number of "conceptual layers" and a number of "functional components", and both the conceptual layers and the functional components should be as independent of each other as possible. One should be able to change the design of a layer, and to add, delete, or change components, with as few implications as possible for other layers and components, respectively.

The conceptual layers and functional components, which need to be most stable over time, are the layers and functions close to the

interface between the central DL and its surrounding, input- or output-oriented "satellite systems". These layers and functions should have a general, simple, and robust design; it has to be agreed upon by all interested parties at an early stage of the overall DL design process.

The **relational data model** is appropriate as a basis for this level of the DL design, but it has to be extended with a few concepts and functions that are needed for the modelling and management of statistical data and metadata. Modelling and management of **hierarchical classifications** is one such area for extensions, where the BOC has already gained important experiences and know how.

Layers and functions which are more "system-internal" - either within DL itself, or within the "satellite systems" - could be developed and optimized according to a "step-by-step" strategy, starting from simple but maybe somewhat crude and technically less than optimal solutions, and then refining and optimizing as needs occur. But even for the internal layers and functions, it is extremely important that they are as independent of each other as possible.

A special case of this **principle of functional independence** is that data and metadata management functions should be as independent of each other as is logically possible.

According to the principle of functional independence, the corporate database itself should be kept as independent as possible from the user views, directories, search mechanisms, and access paths leading into it, so that, on the one hand, physical reorganization of the database, and on the other hand, additions of new user views, directories, search mechanisms etc, can be carried out as independently of each other as possible.

In the user-oriented system layers it should make as little difference as possible whether the underlying data are stored as macrodata, microdata, or metadata. Optimization considerations may call for transformations between these three storage levels, but such reorganizations should be invisible to the external users.

V. PRINCIPLES AND PLANS FOR IMPLEMENTATION

A. General Design Considerations

The proposed architecture for BOC data and metadata management, architecture (c) in section II-D, will not evolve spontaneously; a "let-go" policy will likely lead to architecture (a).

If architecture (c) is regarded as the desirable architecture for the BOC (or a substantial portion), some explicit decision making and agreement on standards is required for

- data structures (content and physical) for logically central storage of data and metadata;
- interfaces between output-oriented systems and the users;
- interfaces between production-oriented systems and the internal users;
- interfaces between the output-oriented systems and the logically central database component;
- interfaces between the survey systems providing data and the logically central database component.

B. Design Principles

Based on the general design considerations listed above, some specific design principles can be listed. They are

- Principle of Functional Independence - each aspect of the unified system should remain as independent as possible from each of the other parts, e.g. the database and metadata repository will be independent from each other;
- Use existing BOC, FIPS, national (e.g. ANSI), and international (e.g. ISO) standards for
 - (a) metadata content (e.g. SDSM, Data Element Standards);
 - (b) data and metadata storage (e.g. IRDS⁴, Relational Model)
 - (c) data and metadata retrieval (e.g. SQL, GILS)
 - (d) data and metadata transfer (e.g. SDTS⁵)

⁴ Information Resource Dictionary System: ISO/IEC 10027 or ANSI X3.138 and FIPS 156. The international and U.S. standards are slightly different.

⁵ Spatial Data Transfer Standard, FIPS 176, developed by the Federal Geographic Data Committee.

- (e) user and system interfaces (e.g. http⁶, CORBA⁷)
- Use the internet for access and publication, including
 - (a) use World Wide Web browsers (e.g. Mosaic, Netscape) for the graphical user interface;
 - (b) use WAIS, hyper-text, and other search engines for accessing information;
 - (c) use html to publish documents;
- Use distributed architecture;
 - (a) use distributed database for microdata and macrodata;
 - (b) use distributed metadata repository;
 - (c) inter-connect individual output-oriented systems with middleware.

C. Implementation Plan

The proposed system and architecture will have to be built in small steps, acquiring additional agency support as time goes on. The traditional business approach of gaining full support from top management first does not work at the BOC. The corporate culture is such that success is achieved by building small systems, proving their worth, gaining new converts, and adding to the system. The cycle is repeated until a full system is complete. This approach is time consuming, but experience has shown it succeeds because the converts feel they have a stake in the success of the project.

The general approach to implementing the unified system and DL will be to build the systems in three main phases. There will be no time schedule for this; the success of any one of the phases depends on achieving success at all the earlier phases.

Phase one is the time for experiment and groundwork. It mainly consists of the following:

- Securing partnerships with other groups willing and interested in developing the systems;

⁶ Hyper-Text Transfer Protocol

⁷ Common Object Request Broker Architecture, a type of middleware.

- Determining relevant standards and developing ones when they are needed, especially the SDSM;
- Educating others throughout the agency in order to sell the new ideas and gain additional support;
- Build proof-of-concept systems to
 - gain experience with new software systems;
 - determine which approaches are feasible and which are not;
 - demonstrate the effectiveness of feasible approaches;
 - further educate and gain more converts;
- Evaluate existing tools;
- Develop new tools, especially for new technologies;
- Gain some high level support.

Many of the parts of the the first phase will be repeated multiple times until effective solutions are found.

The second phase is geared toward building a fully functional prototype. The main activities are

- Build production prototypes of each major sub-system;
- Integrate the sub-systems into fully functional prototype;
- Demonstrate the system to gain wider support;
- Increase scope of systems for wider support;
- Secure support of top management.

The last main phase is the implementation phase. The major goal is to implement a functioning system. Since the unified system and DL will be standards based, there will have to be acceptance of many new standards at the agency level. Acceptance of the SDSM by the BOC, for instance, is critical to the success of this plan.

Even more importantly, the unified system and DL will represent an entirely new way the BOC will conduct its daily activities and business. The phased implementation plan presented here will help minimize the risk of rejection and ultimate failure of the system.

VI. References

SUNDGREN 1989: Bo Sundgren, "Statistics Production in the 90's - Decentralization without Chaos", ISI/IAOS conference in Bilbao, Spain, 1989. Also available from Statistics Sweden as R&D Report 1989:11.

SUNDGREN 1991A: Bo Sundgren, "What metainformation should accompany statistical macrodata?" Discussion paper for the June 1991 Meeting of Working Party 9 of the OECD Industrial Committee. Also available from Statistics Sweden as R&D Report 1991:9.

SUNDGREN 1991B: Bo Sundgren, "Statistical Metainformation and Metainformation Systems". Report to the UN/ECE Joint Group on Metainformation Systems (METIS). Also available from Statistics Sweden as R&D Report 1991:11.

SUNDGREN 1991C: Bo Sundgren, "Towards a Unified Data and Metadata System at the Australian Bureau of Statistics - Final Report". Report to management of Australian Bureau of Statistics, 1991-12-02.

APPENDIX A

Automated Reference Rack (ARRk) is a hypertext based system designed using Lotus SmartText. Telephone clerks at the Census Bureau use ARRk to help the public find appropriate published data for their work. ARRk has short descriptions of each publicly available file which contains information about subjects, coverage, storage medium, cost, and ordering information. Hypertext links on subjects and coverage help the clerks locate files. But detailed information such as data dictionaries, record layouts, sample designs, questionnaires, and data quality measurements are not available through this source.

BOX Files is a mechanism for incorporating metadata into ASCII data files (Bean, 1991). The major advantage of this system is that the systems level metadata that describes each file are automatically carried along with the data in a file. The main purpose is to include data dictionary and record layout metadata for reading and processing the data, but the system is flexible and allows for many types of information to be stored. The BOX file format is the basis for a recently issued Census Bureau information technology standard for archiving data, although this decision is under review. ISO 8211 (Specification for a Data Descriptive File for Information Interchange, also FIPS 123) is functionally equivalent to BOX and may be more appropriate. Some software for converting datasets between SAS format and BOX format has been written, but additional software must be developed for other formats. As a result, the BOX file system has only seen limited use in Census Bureau surveys. Since the BOX system was designed primarily as a data transport system and the metadata for each data set are stored with the data, there is no mechanism to compare metadata across files, surveys, or programs.

Data Extraction System is a general system for extracting data from master data sets into one of several popular data formats, including SAS data sets. It also uses files stored in the BOX format (see above). Plans exist to expand the system to include the creation of FORTRAN or C "read" subroutines for any BOX file. This system has seen limited use, but could be part of a generalized software system for accessing data.

Surveys-On-Call is an on-line system for accessing publicly available Survey of Income and Program Participation (SIPP) and Current Population Survey (CPS) data. This system, previously known as SIPP-on Call, is currently in production and is accessible over the Internet (via the BOC home page) and by modem. Surveys-On-Call is a UNIX based menu system which provides users with easy access to Demographic surveys data. Users can define extract files from the screen as SAS data sets and receive documentation about these data sets. This system

makes wide use of the Data Extraction System (see above). A limitation of the system is that the online documentation available is not complete. A user must obtain a hardcopy documentation guide in order to use the system effectively.

Extract is a system in use with CD-ROM products sold by the Census Bureau. Libraries and other public facilities often have these products on the shelf. Extract is a menu driven Dbase application which allows the user to construct extract data files. Metadata is also available and can be appended where necessary in the extracted files. Through the use of index files, any type of metadata (or documentation) can be included on the CD-ROM. However, data definitions, code structures, record layouts, and the questionnaire are usually provided. Detailed information about survey design and data quality issues are not included. As with the other systems described so far, the available metadata is limited to describing the application data.

CENSAS is a project for an automated data and information delivery system based on Decennial Census data for both internal and external customers. Because of past difficulties in accessing customer-defined subsets of this data, the CENSAS staff has made its first priority the production of an automated system allowing internal customers to specify and receive such subsets (known as "extracts"). These extracts are delivered to the user as SAS datasets to which the user can apply the desired tallies or other statistical analyses. A Beta Test version is available.

CENDATA is an on-line Census Bureau data system containing current and historical data, both demographic and economic. Examples include Foreign Trade Statistics, Quarterly Financial Reports, County Business Patterns, Wholesale and Retail Trade, Center for International Research, Agriculture County, and Manufacturers, Shipments, Inventories, and Orders Survey data.

FERRET (Federal Electronic Research and Review Extraction Tool) is a data extraction tool available on the internet that allows users to find information about monthly demographic survey data using a World Wide Web browser. Users can select microdata items (individual survey question items) which can be used to create custom data queries. In addition, users can select macrodata (aggregated or summarized) tables to get preformatted survey data. Results of data queries can be output in SAS datasets or ASCII files. These results can be viewed on the screen or can be downloaded to a local computer. The SAS output allows you to get the results in pie charts, bar charts, or summarized on a U.S. map. The ASCII output can be brought into an Excel spreadsheet.

The FERRET system can be broken into four major parts. The first part is the user interface which is via the World Wide Web. The

Ferret repository contains metadata such as basic variable definitions, keywords, concepts, and other items. The Document Management System handles the documents which describe the survey design, processing, and analysis. It provides users the ability to modify or access documents via an on-line table of contents or WAIS searches. Finally there are two databases handling all the microdata and macrodata.

FERRET currently handles Current Population Survey data. Plans are to add other demographic survey data in the future. There is also an effort to adapt the system to economic survey data, too.

StEPS (Standard Economic Processing System) is an integrated survey processing system the objective of which is to eliminate redundant processing by combining existing survey systems into one system. The scope of the StEPS system includes providing the following basic survey processing functions:

- Data review and correction;
- Edits;
- Imputation;
- Outliers;
- Estimation;
- Estimation variance;
- Disclosure analysis;
- Time series;
- Queries (canned/ad hoc);
- Tables (canned/ad hoc);
- Management information; and
- Survey control operations (for scheduling of batch mode processes).

It will also provide the following additional functions:

- Generate standard and non-standard mail files for mail-out operations;
- Generate standard telephone files for telephone follow-up operations;
- Maintain standard variable names and flags;
- Maintain standard data structures;
- Allow entry of survey design specifications including edit and imputation parameters as determined by analysts or through automated historical data analysis
- Provide audit trails and backup capabilities;
- Provide access to SSEL; and
- Provide access to other economic area surveys and censuses.

The above provides a view of the functionality which StEPS will be designed to provide. Implementation details are not yet available. A prototype is expected to be developed by the end of 1996.

IPS (Integrated Processing System) is envisioned to be the umbrella for a compatible set of automated tools to design, conduct, and manage Census Bureau surveys and censuses in an effort to improve cost effectiveness, timely reporting, data quality, and data access. The overall goal is to provide a framework for the integration of generalized processing system components with data collection tools, as well as with other generalized systems under development such as generic reinterview, sampling/universe subsystems, and data dissemination technologies. General objectives of the system include providing a common structure for sharing and accessing Census Bureau data; allowing a survey manager to model a survey or census; allowing a survey manager to design and manage a survey or census; providing automated, standard documentation; and providing easy access and integration of generalized processing tools for modularized survey processing and analysis.

Work has begun to design an initial proof-of-concept system, but details of the plan and a schedule for completion have not been finalized yet.

DADS (Data Access and Dissemination System) is the name for the Census Bureau initiative to develop and implement data access and dissemination focussed on the 2000 Decennial Census and Continuous Measurement data sets, but with the ability to accommodate other data sets having geographic detail, such as those produced from the Economic and Agricultural Censuses.

The main objective of DADS is to provide one general (electronic) system for all access to Census Bureau data. The system will be designed to be fast, flexible, and cost-efficient. To achieve this, four cross-directorate teams were formed to study and recommend policies or designs for user input, promotion and outreach, pricing for products, copyright or trademark, corporate look and feel, data archiving, metadata and documentation, and coordination of various efforts and activities. DADS will attempt to incorporate other work, such as FERRET, where it is appropriate.

Standard for Survey Design and Statistical Methodology Metadata (SDSM) is a standard under development at the Census Bureau to specify the metadata necessary to describe survey designs, processing, analyses, and data sets completely. Subject matter experts throughout the agency are being consulted to ensure the SDSM describes all pertinent information. The SDSM is being written as an extension to the **Cultural and Demographic Data Metadata** draft standard which is being developed under the auspices of the **Federal Geographic Data Committee**. Approval for the standard will be sought through the formal standard development procedures of the Census Bureau, and the process is

expected to be completed by September 1996.

Development of the SDSM was motivated by four factors. The first is that development of data dissemination or integrated survey processing systems with combined metadata repositories requires a formal description of the metadata which is to be made available. The SDSM will specify a core required set which represents the minimum metadata necessary. The second motivation is the need for formal data models (metamodels) which can be used to design the metadata repositories. Third, the standard can be used as a means to develop check lists and tables of contents for documentation creation, modification, and access tools associated with data dissemination and integrated survey processing systems. A formal table of contents outline of the SDSM is also under development. It has been implemented on the World Wide Web and can be accessed via web browsers, such as Mosaic or Netscape. The formal table of contents is a readily understandable version of the SDSM, and the web implementation has the means for users to submit comments. The last motivation is the use of the standard (or table of contents) as a formal interface between disparate data dissemination and integrated survey processing systems. Implementing this interface will allow access of data and metadata across systems, leading to a seamless environment.

Metadata Repository is the project to build a logically central repository of metadata based on the SDSM (see above). The Open Workgroup Repository (OWR) software from Manager Software Products which is based on the **Information Resource Dictionary System** standard, **ANSI X3.138**, has been purchased for building the system. A proof-of-concept system has been built, and work to design and build a conceptual metamodel based on the SDSM has begun. The metadata repository is intended to be a source of metadata for all the Census Bureau programs so comparisons of designs, processing, analysis, or data can be made across time and survey programs. Database queries (SQL), internet searching (WAIS), and other data mining techniques will be used to make the information available in a variety of ways. Some specialized tools will have to be built for populating the repository and creating, editing, and accessing documentation.

Fundamental to all the metadata that is described in the SDSM is data element definitions. Data elements and their definitions are incorporated into all aspects of survey work. A first step to implementing the metadata repository is to build a data element registry, and one is being built to coordinate with the StEPS project (see above). This work will be based upon some developing national and international standards: **ISO\IEC 11179 - Specification and Standardization of data Elements**; and **ANSI 1125-D - Metamodel for the Management of Shareable Data**. The initial implementation will use the OWR software.