STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

Working Paper No.8
English only

Work Session on Statistical Metadata
(Geneva, Switzerland, 18-20 February 1998)

Item 3 of the provisional agenda

TRANSITION PLAN FOR A UNIFIED APPROACH TO METADATA MANAGEMENT AT THE BUREAU OF THE CENSUS

Submitted by Bureau of the Census, United States [1]

[1] Prepared by Cheryl R. Landman, Martin V. Appel, Daniel W. Gillman and Samuel N. Highsmith, Jr.

I. VISION STATEMENT

1. The vision for metadata management at the Bureau of the Census (BOC) is to build and maintain a logically central corporate metadata repository (CMR). The CMR will be used as a library of information about surveys, products, datasets, and variables. It will contain pointers to survey data and the documents that describe survey designs, processing, analyses, and data sets. Tools for updating, querying, managing, and registering metadata in the CMR will be available. The CMR will function as the library of information for electronic data dissemination and automated survey processing/design systems. General access to the CMR will be through the Internet/Intranet.

2. This vision supports the Bureau's Data Access and Dissemination System (DADS) vision to improve accessibility to information we collect and process. Metadata management is the enabling technology to build closer coordination across program and support areas for sharing survey processes.

II. INTRODUCTION

3. Over the last several years, the BOC has developed prototype metadata components--tools, models and standards in "support of statistical data production and managment". This paper provides a "common solution that could be shared by others", specifically addressing the transition from research to production for metadata and data administration. It also provides the BOC metadata current status and planned next steps, highlighting "practical, new experiences". This paper presents a transition plan for moving to a unified approach to metadata management, essentially migrating in stages from a research/prototype to a full production environment. The paper is organized in the following manner - a brief general introduction and statement of purpose, followed by the transition plan, a proposal for the establishment of a metadata management staff, current status of metadata at the BOC and planned next steps. The Appendix provides a vision of the BOC Metadata Repository production environment.

4. The management of statistical information is part of the core business of the BOC. Our customers increasingly expect more data released much sooner than was previously acceptable. With ever increasing Internet utilization by our users, the BOC must continue to be progressive in releasing data electronically. DADS is the vehicle for improving data access and meeting our increasing customer demands for BOC information. While we continue to expand the amount of data available via Internet (DADS), these data are of little value without additional descriptive information. Metadata is that additional descriptive information which describes data content and organization. It provides the information that makes data useful, understandable, and shareable. Metadata in computer-readable format is the means to help automatically manage, share, re-use, combine, analyze, and evaluate data. With the establishment of the CMR model, we are beginning to build a corporate electronic library for accessing BOC data regardless of origin. The CMR will contain the metadata for survey designs, processing, analyses, and data sets for all surveys the BOC performs. Links from the CMR to the data files, documentation, and images (such as questionnaire forms) will be available.

III. PURPOSE

5. The ***main objective*** of this plan is to describe implementation of metadata management to support the organization's efforts to achieve the goals of the strategic plan—greater customer satisfaction, greater productivity, and better public perception and cooperation. This plan will transition metadata management from its current prototype stage to full production.

This will be accomplished through four distinct stages, which we call the transition stages.

IV. TRANSITION STAGES

6. **Phase 1--(Current through January 1998)**, the Prototype Stage, includes the initial research and technology evaluation followed by building, testing, and evaluation of a prototype system. This is the current phase and has resulted in agreement between DADS, Demographic and Economic program areas, Publication/Marketing Support and Statistical Research areas to share one metadata model. During this prototype phase, a number of tools to populate and manage the CMR are being built and tested. The next step is to build a dynamic CMR from the DADS Prototype 2 metadata which was based upon the shared metadata model.

7. **Phase 2--(February 1998 through February 1999)** is the Transition Stage in which the current prototype system will be replaced by a production CMR. This stage includes but is not limited to the identification of key projects for incrementally developing the production CMR. The CMR will be a physically distributed but logically central Intra/Internet based database application. During the Transition Stage the necessary support and organizational changes will be identified and put into place. Necessary new standards and policies will be identified and implemented. Tools necessary to access and populate the CMR will be built and distributed for general BOC usage. This transition phase will result in both an agreement across the BOC to build, populate, and utilize one logical CMR and the definitions and tools necessary to make it a practical course of action. We will have in place a standard methodology for describing and accessing BOC metadata.

- After release of the second DADS prototype, modify and extend the DADS metadata repository for adoption as the CMR. Develop mechanisms to keep the BOC metadata model current with the metadata standard and active repositories which are built with the metadata model.
- Form two teams to accomplish the short term goal of registering and classifying metadata: a Subject Matter team and a Prototype Development team. Their purpose is to propose ideas and oversee implementation of solutions for the efficient registration and classification of metadata. The Subject Matter team defines the components of the Web based templates and business strategies for collecting, registering, and retrieving metadata. The Prototype Development team in conjunction with the DADS staff would develop prototype tools that would be fully integrated and implemented within DADS.
- Develop a long term (2-3 years) strategy or how the BOC can implement systems which reduce costs and accommodate changes in technology. Determine staff and funding operations for production operations. Put in place a metadata management staff --the Electronic Information Support Staff (EISS)-- see next section for details.
- Market and train staff on the capabilities of the CMR throughout the BOC. This will be an ongoing effort led by the EISS.

8. **Phase 3--(February 1999-February 2000):** At this point in the transition, the EISS should be in place, classification teams and prototype develop teams should be well along in their efforts, and a migration schedule exists. The DADS metadata repository should be populated and be established as the production CMR for the BOC.

- With EISS in place, implement data and metadata administration at the BOC using DADS Repository.
- Modify metadata management staff--EISS roles and responsibilities as a result of Phase 2 input.
- Organize and combine staff that have metadata responsibilities to focus our efforts and eliminate redundancy.

9. **Phase 4--(February 2000-February 2001)** is the Full Production Stage in which the CMR is fully integrated into the mainstream of the BOC. Now the focus is on continued marketing to BOC program managers. Formal assistance and guidance is available and aggressively provided to program managers electing to move to this technology.

- Formal training in the use, installation, and administration of the CMR is provided to BOC users and administrators. The CMR will greatly increase BOC metadata utilization by both internal and external customers. The exact configuration will depend on what we learn about the operating requirements and capabilities of the CMR from internal BOC customers.
- The Subject Matter team evolves into a steering committee (part-time role) for registration and adjudication of metadata while the Prototype Development team would provide the system support for DADS including metadata. If DADS is already supported, this team would be eliminated.

10. The phases described above will move the (CMR) technology from the Prototype stage (phase 1) to the Transition stage (phase 2), Pre-production stage (phase 3), and finally to the Full Production stage (phase 4). The Research areas have initiated and coordinated the research effort as part of phase 1 but now project leadership should be assumed by a new staff responsible for overall production and implementation, the EISS. This staff will be responsible for the overall production implementation of BOC metadata management. The EISS will guide the evolution from research to full production and must be in full partnership with DADS.

V. ESTABLISHMENT OF THE ELECTRONIC INFORMATION SUPPORT STAFF (EISS)

11. **Proposal:** An Electronic Information Support Staff (EISS) should be established to manage the BOCs metadata resources. The EISS would in essence form the nucleus of the future DADS staff. The EISS will provide services to all divisions and offices in the BOC. The staff's customers are internal BOC divisions, offices, and individuals. Their goal will be to provide service—not only responding to current needs and requests, but also actively seeking to understand future directions and requirements.

12. During the transition period, the Metadata Research areas will transfer responsibility for planning and implementation of the CMR to the EISS. Their role will initially be to guide and coordinate further development of the BOC Metadata model and Metadata standard.

13. **Summary of Responsibilities:**

- CMR—Negotiate the links for the database repositories. Consensus building for the CMR and tools.

- Internet Metadata Dissemination—Ensure standards are met for metadata released on the internet. Participate as BOC representative in Federal Web Consortium. Develop link for "One Stop Shopping"

- Metadata Collection—Develop tools and standards for collection. Program divisions will be responsible for the collection/entry to CMR.

- Metadata Standards Implementation—Participate in the standards process. Steer the implementation of corporate national/international standards. Specifically steer the development of standards for BOC-wide definition adjudication, registration authority, security procedures/systems etc.

- Data and Metadata Administration—Coordinate the building of corporate tools for input, maintenance and access. Represent the BOC's corporate perspectives at internal and external meetings related to metadata. Steer the phased transition to a production environment. Establish interim mechanism that can eventually become the maintenance, training, and continued development site for metadata.
- Education and Communication—Market/promote the corporate use of metadata concept and its uses. Develop communication channels for internal and external users exchange of information. Develop education/training modules, demonstrations to survey planners, etc.

- Research—Coordinate corporate metadata research, usability testing, internal and external meetings to exchange ideas/plans/ status of metadata in other organizations; coordinating the exchange of information collected by internal and external units about research and production successes. DADS/data access research, looking to the future environment also would be a critical research component.

- Additional responsibilities could encompass coordinating Bureau-wide research and implementation of generalized systems for such areas as photocomposition, disclosure avoidance, sampling error calculations, standardized editing and processing tools that could be incorporated as modules within the DADS.

14. **Staff Functions and Roles:**

The EISS will be responsible for developing a business plan for the metadata repository which links with the strategic goals of the BOC. This business plan must quantify specific benefits and savings attained by migrating to this environment, such as cost and time savings for the re-use of data, quicker survey development, reduction on survey development and processing costs, better customer service and so forth. The business case will also identify functions that could be eliminated in other areas as a result of the EISS as well as recommend organizational modifications necessary for a cost-effective metadata environment. Specific EISS staff responsibilities include the following major functions that are outlined in detail below: systems management, repository management, data and metadata administration, tool research and development, customer service and education.

15. **Systems Management**

- Determine system requirements -- Hardware and software requirements that provide a fast and reliable Intra/Internet metadata repository for bureau-wide use.

- Determine staff resources and skills that CMR maintenance will require.
- Manage the resources (human and hardware/software) needed to maintain the system
  -- Manage the systems' funding and budget.
  -- Establish a funding strategy for this shared corporate resource.
- Address security issues.
  -- Serve as a liaison with the Security Office for security issues resolution.
- Determine system users' training needs and arrange to meet them.

16. **Repository Management**

- Coordinate the administration of the various databases that comprise the CMR.
- Assign and manage processes that govern the function of the CMR.
- Develop and provide standards, procedures, and education for using the CMR.
- Provide central support for all information related to the CMR operation.

- Develop guidelines to assist users in meeting the established/approved metadata policies.
- Work with established bureau-wide groups such as the Internet Users Group and the Standards Management Team to ensure that the repository serves users' needs and employs Bureau-wide standards.

17. **Data and Metadata Administration**

- Establish a Registration Authority function and adopt policies for metadata registration.
- Develop a process that consult with the divisions to establish the rules for metadata registration.
- Develop a process that moves the Bureau toward a centralized metadata environment.

18. **Tools Research and Development**

- Keep abreast of changes in technology and arrange for upgrades as needed.
- Conduct periodic needs assessments.
- Review internal and external user feedback and recommendations.
- Monitor Internet and Intranet applications of other federal government agencies and the private sector.
- EISS Subject Matter jointly participate with Information Technology area (IT) in Federal Web Consortium collaborative efforts, as appropriate.

19. **Customer Service and Education**

- Educate Users.
- Communicate standards.
- Build training modules.
- Train users on the CMR and associated tools.

20. **Coordination and Liaison**

- Consult with knowledgeable staff throughout the organization, e.g., the Marketing Services, Customer and Publications areas, the Internet Users Group, the Standards Management Team, IT Liaisons, and so forth.
- Resolve all metadata issues and as necessary report issues and resolutions to the oversight group-- Data Dissemination Policy Steering Committee.
- Monitor adherence to the established/approved metadata policies and act as a help desk to the program areas. (Note that the program divisions should have responsibility and control of their metadata and thus will be responsible for ensuring its within the standard.)
- Establish mechanism to ensure CMR meets the customers requirements and customers' are satisfied with its performance, maintenance, response, etc.

21. **Organization**

The EISS should be established as a service to all divisions and offices in the BOC. It should have a manager who is well respected for his/her knowledge of the organization's subject matter and for having a customer service orientation. The office should have a small core staff (possibly 3-4) plus contractor support. Basically following the DADS model.

22. The core staff should consist of persons with knowledge and skills in technical and subject-matter areas plus management, outreach/marketing, customer service, electronic library and cataloging skills, contracting, systems analysis, and security[2]. One staff member should come from each of the major subject-matter areas—demographic, decennial, and

[2] Additional functionalities need include: Systems Administrator, Registrar, Repository Manager (DBA), Tools Developer

economic—and have the ability to serve as a communication link between subject-matter areas and systems experts.

VI. CURRENT STATUS OF METADATA AT THE BOC

23. **DADS Prototype 2.** The DADS project has adopted and is implementing the metadata repository model developed by the research areas.

- **Tools.** Research area is currently developing prototype tools for registration of surveys, products, data sets, variables and documents.
- **Document Management System.** The Standards Management Team has recommended the use of a bureau-wide document management system--PCDOCS. This will aid the development of common tools to share data and documents.
- **Metadata Standards.** The first standard (Survey Design and Statistical Methodology Metadata), addressing the business model and how documents are catalogued is going through bureau-wide review and modifications based upon user input.
- **Memorandums Of Understanding** (MOU). Several MOUs have been developed and signed between the Statistical Research and Marketing/Publications Support areas and Demographic and Economic Program areas. These MOUs describe cooperative, shared development projects involving the use of the CMR with Internet Data Dissemination and Automated Survey Processing systems at the BOC. Ongoing projects such as DADS, Demographic extract tools and Economic generalized system and the Product Registry have all agreed to use the CMR to assist in their work.
- **Models** The models necessary for the design of the CMR have been built. Each model represents a dimension of the complete model for the CMR. The models are listed below:
- **Business Data Model** - It defines the metadata and relationships necessary to describe the BOC's business practices (surveys and survey data).
- **Data Element Registry Model** - A data element registry is a mechanism for managing the names, definitions, and other attributes of data elements.
- **Metamodel** - This is the model describing schemas of information about data sets, such as specifications for record layouts or database schemas. It also contains versioning, security, user type, and search category information.

VII. NEXT STEPS

24. The EISS should be established.
Near-term activities of the EISS:

- Selection of a program manager to head the office. The program manager should immediately begin meeting with divisions to understand their needs, plans, and expectations.
- Conduct a needs assessment with all divisions and offices.
- Conduct of a benchmarking study.
- Conduct of a technical assessment.

25. Based on the results of these studies and assessments, develop a plan for full production support of the BOC's repository.

- Continue development of tools for automatic registration.
- Expand MOUs to cover Geography Division, archiving, etc.
- Develop a real survey example for all components of the metadata standard.
- Begin educating users about the positive aspects of metadata.
- Additional issues that could be driven by the metadata environment and that need to be immediately addressed for the next phase of DADS include disclosure avoidance, photocomposition, and variance estimation.

VIII. CONCLUSION

26. The Bureau of the Census has been successful in integrating the metadata model in several key prototype projects in support of statistical data production and management, developed prototype tools and has initiated metadata standards. Now, it must migrate the metadata support from the research environment to an area that will develop, support and maintain the metadata models and tools in a full production environment. The above outlines our roadmap to migrate to a full production environment and envisions the Census metadata repository of the future as shown in the Appendix.

APPENDIX

THE CENSUS METADATA REPOSITORY OF THE FUTURE

Production Environment:

- **Architecture**
  The CMR will be a logically central database of metadata for the BOC. This means it will appear to the user as a single database, but it will be maintained as several distinct databases using the same underlying model. Each functional area within the BOC that needs to manage its own metadata can do so by acquiring the CMR model and tools (see below). The model and tools will be used to build local metadata repositories that function as parts of the CMR.

- **Hardware**
  The hardware necessary to implement the CMR is the servers, routers, and disk systems for maintaining the Internet inside and outside the firewall at the BOC. There must be sufficient disk space available to store the metadata for each survey the BOC conducts. This includes space for all the descriptive documents produced by survey designers and analysts.

  The tools for the CMR (see below) will reside on client (PC) systems, so access to the Internet must be available from the desktops of users of the CMR.

- **Software and Database**
  The software needed to implement the CMR is a Database Management System (DBMS) in common use across the BOC. The ease of connecting multiple databases will increase dramatically when implementers use the same DBMS. Since Oracle DBMS is in common use throughout the BOC, this should not be a problem.

  Open Database Connectivity (ODBC) software and proprietary software for client/server systems will be needed to connect the tools to the CMR. Common use of Oracle will increase the usefulness SQL*NET for this purpose.

  Use of Web tools and software will enable the tools to be used by a wide audience without the procurement and installation of specialized software on client hardware systems.
  Tools are needed for updating, querying, and registering metadata in the CMR. To be most effective, easily distributed, and useful, tools will be Web-based and use open or standard software as much as possible.

The metadata query tools (or browsers) are under development as part of the development of DADS and StEPS. Prototype browsers have also been developed as part of the CMR development in SRD.

Tools for updating and registering metadata are necessary to make the CMR work. Prototypes of these tools are under development at this time. Metadata update tools are necessary for the maintenance of the CMR, especially for the data and metadata administration functions.

Registration tools (see below) are the vehicle for submitting metadata to be entered into the CMR. Tools and templates for each type of metadata class (documents, data elements, datasets, surveys, and products) will be used to collect and submit metadata to the CMR. Prototypes of these tools are also under development.

- **Standards**
  The design of the CMR is based on two standards. These are: 1) **Survey Design and Statistical Methodology (SDSM)**, a proposed BOC standard (LaPlant, *et al*, 1996; Census Bureau, 1996a; or Census Bureau, 1996b); 2) **Metamodel for the Management of Shareable Data** (draft) **(MMSD)**, ANSI X3.285. The purpose of the **SDSM** and **MMSD** standards is to specify the basic metadata models necessary to build the CMR. The SDSM has been submitted to the Standards Management Team for formal review. The review process is expected to be finished in Fall, 1997.

REFERENCES

CASIC Inter-divisional Team, “Integrated Processing System”, December 1994

CASRO, “Integrated Computing Environment Initiative”, April l997.

DADS Metadata Team, “Metadata Strategy Report”, February l997

Gillman, D. W., Appel, M. V., and LaPlant, W. P. Jr. (1996), A Design Principles for a Unified Statistical Data/Metadata System@, Proceedings of SSDBM-8, Stockholm, Sweden, June 18-20, 1996.

Gillman, D. W. and Appel, M. V. (1997), “The Statistical Metadata Repository: an Electronic Catalog of Survey Descriptions at the U.S. Census Bureau”, Presented at the IASSIST /IFD’97 Conference - Data Frontiers in the Infoscape , Odense, Denmark, May 6-9, 1997.

Gillman, D. W. , Appel, M. V., and Highsmith (1997), “Building a Statistical Metadata Repository, Presented at “Metadata’97 Second IEEE Metadata Conference”,.Silver Spring, MD, October 16-17, 1997.

Sundgren, Bo, Appel, M. V., Gillman, D. W., and LaPlant, W. P. (1996), “Towards a Unified Data and Metadata System at the Census Bureau”, Census Annual Research Conference, Arlington, VA., March 18-21, 1996.