STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Joint UNECE/Eurostat Work Session
on Statistical Metadata
(6 - 8 March 2002, Luxembourg)

COMMISSION OF THE
EUROPEAN COMMUNITIES

EUROSTAT

Working Paper No. 4
English only

Topic (i): Infrastructure issues for statistical metadata

BUILDING A METADATA REPOSITORY TO SUPPORT THE 2002 ECONOMIC CENSUS

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

Invited paper

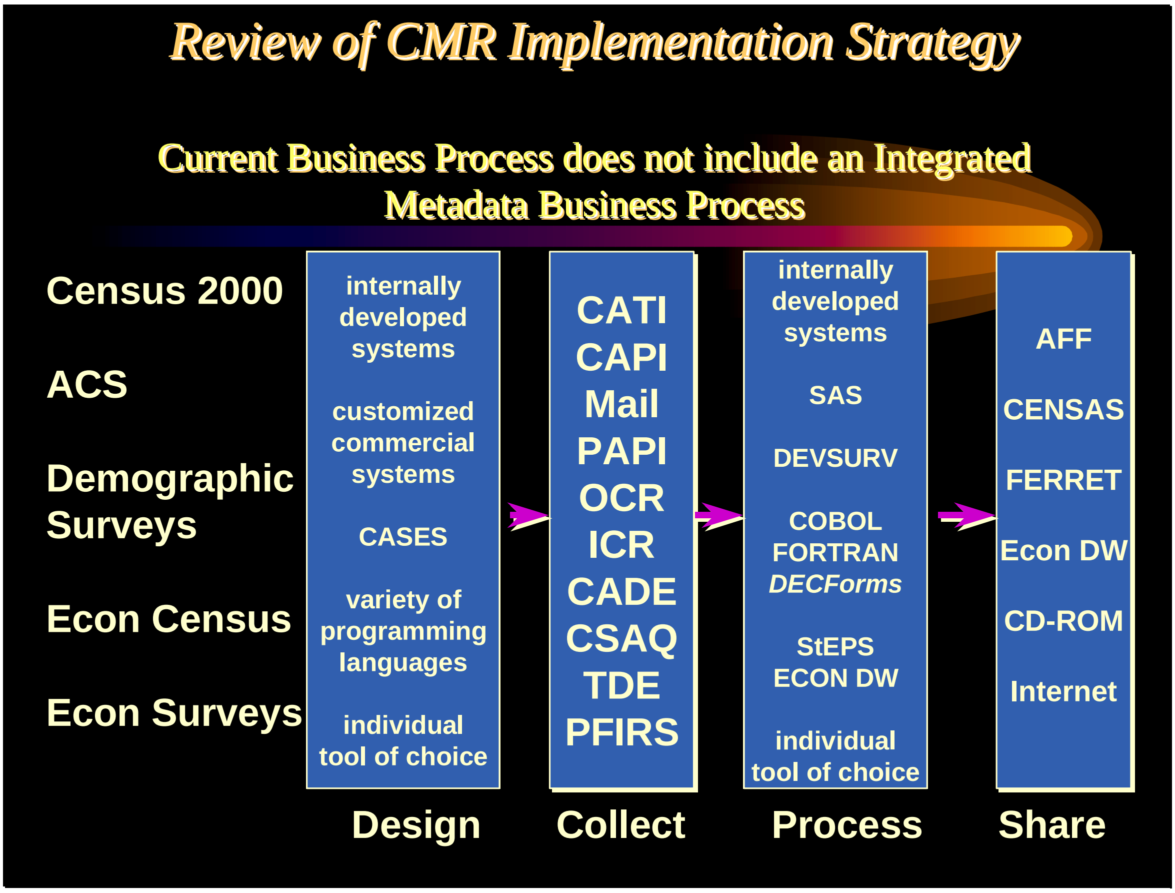
I. INTRODUCTION

1. The U.S. Bureau of the Census (BOC) has been developing a Corporate Metadata Repository (CMR) since late 1998. The CMR is based on a census and survey life cycle model developed in the mid 90s. It was developed in conjunction with work from Sweden, Canada, Australia, and UNECE Metadata Workshops**.** This project was described in a contributed paper at the ISIS 2000 conference in Riga, Latvia. Since our effort to build an Economic Metadata Repository (EMR) is based on the BOC CMR project and actually utilizes techniques and tools developed for the CMR, I will include a brief overview of the CMR project as background for this paper.

2. The U.S. Bureau of Census (BOC), like most other survey organizations, has been purchasing and developing computer solutions for survey processing for many years. This has resulted in a survey-processing environment composed of many disparate solutions, very few of which communicate with each other. The result is that we now have many systems that access or process their own dataset(s) through the use of specific non-shared documentation for those datasets and processes. This leads to a number of very common related complaints:

- It takes a significant amount of time to convert a file used in one system to the format required by another system.
- Very little sharing of documentation or procedures causes the natural proliferation of different systems to solve the same problem.
- The cost to develop a new survey or census is very high if one cannot take advantage of the solutions developed in earlier systems.

[1] Prepared by Samuel Highsmith.



Review of CMR Implementation Strategy

Current Business Process does not include an Integrated Metadata Business Process

FIGURE 1

3. Figure 1 illustrates some of the many systems at the BOC that do not communicate with one another. The diagram depicts the major survey and census groups within the BOC. The Economic Census, 2000 Decennial Census, and Decennial Census supporting American Community Survey (ACS) have all embraced the concept of and plan to use a metadata driven data dissemination effort. We have many methods in use at the BOC to Design Surveys, ranging from internally developed systems to commercially available tools such as CASES from the University of California at Berkeley. We have a wide variety of survey and census data collection tools that use technologies such as Computer Assisted Telephone Interviewing (CATI), Computer Assisted Personal Interviewing (CAPI), mail out surveys, and Computer Self Administered Questionnaire (CSAQ). The processing tools in use are just as diverse, including the Statistics Canada Devsurv, many systems using SAS, and a number of internally developed systems. For data dissemination, we have several web-based solutions such as American Fact Finder, CENSAS, and FERRET. The main point illustrated in figure 1 is that these many systems are performing similar functions without being able to share with each other.

II. BACKGROUND

4. The CMR project is based on a survey and census life cycle that contains detailed metadata elements describing the full survey and census process from survey design, data collection, editing and processing, and data dissemination. The model was initially built using the Erwin modeling tool. It was migrated to Oracle Designer when we transitioned the project from the research area into a production environment. The use of Oracle Designer allows us to generate the data base schema. This technique provides the ability to modify or add new elements to the model and easily incorporate them into the database schema. The CMR architecture is fully web-based with the only tool needed on the user desktop is a web browser. The CMR is designed to be to the survey process what a card catalogue is to a library.

5. Our initial plan for the CMR project was to build a generic set of tools to allow managing and using the metadata stored in the CMR. In order to show how the CMR could be effectively used, we worked with several application areas of the BOC to build applications utilizing the CMR. It became apparent that requirements for applications in different business units are governed by different business rules. We are now completing the core CMR capabilities with standard interfaces and building more or less customized applications that use those interfaces. It has been necessary to build applications that use the CMR core capabilities in order to show value and obtain funding for the project.

III. OVERVIEW OF THE CORPORATE METADATA REPOSITORY

6. Applications utilizing the CMR model include the American Factfinder BOC data dissemination web application, a metadata cleaning operation to support the movement of large Geographic metadata files to the American Factfinder system, a web enabled ISO/IEC compliant Data Element Registry, and the Economic Directorate Economic Metadata Repository (EMR).

7. The American Factfinder (AFF) system uses the CMR metadata model to describe the many hundreds of Economic and Demographic summary data files accessible on that site. For a file to be made available on the AFF, the data providers have to provide a complete detailed description of the file and all data elements. AFF provided a comma delimited text file format detailing exactly how to provide all of the required metadata. We have built for our Geography Division a metadata cleaning operation that receives these files and performs automatic validation of the contents using both database constraints such as range checks and a set of customized metadata business rules for Geographic product metadata. This process uses a tool named Data Quality Inspector, which was originally built by Oracle for the Centers for Medicare and Medicaid Services, formerly named the Health Care Financing Administration. We will be able to automatically provide to the Data Quality Inspector the required metadata on which it operates in validation of the contents of data sets.

8. Virtually every component in our survey and census processing model includes a documentation component. We refer to this as unstructured metadata because it cannot easily be machine processed and used to drive applications. The CMR includes a document management component based on the Oracle Portal product included in Oracle's Internet Application Server. This allows user input and management of documents with a thesaurus and mandatory classification scheme. The immediate benefits include the ability to easily tie unstructured and structured metadata together without having to code a solution. Documents are left in their native format and automatically indexed and converted by Portal. Our structured metadata CMR model also is integrated with the Oracle Portal. This enables full text searches across all structured and unstructured metadata. Our Demographic Current Population Survey has a portal in operation to allow user management of their survey documentation.

9. The EMR is based on the same CMR model but includes additions to the model needed to support Economic specific metadata elements. Metadata can be easily exchanged between the CMR and EMR since they employ the same base model. The CMR is being built with the concept of being a metadata provider for many systems. The methodology for information exchange is based on XML. We currently have an XML based metadata interchange application in place for the Economic Directorate forms design process.

10. In 2000, our Economic Directorate was in the process of a complete redesign for the quinquennial 2002 Economic Census. In the 1997 Economic Census the Economic Directorate had contracted with Fenestra Corporation to build an electronic Computer Self Administered Questionnaire (CSAQ). Fenestra delivered three survey data collection instruments for 1997, which the Economic Directorate deployed and was very pleased with. The Economic Directorate subsequently contracted with Fenestra to build a CSAQ for the more than 650 questionnaires contained in the 2002 Economic Census. Realizing that hand building 650 questionnaires was not practical and would require an automated solution, Fenestra proposed building an electronic metadata repository for the data collection instrument.

11. The CMR project had already designed a model driven metadata repository already in use by our data dissemination Internet site, the American Factfinder. We proposed using our metadata repository for the Economic Directorate. The CMR modeled the entire survey process, of which data collection was only a part. A two-day workshop was held to jointly examine the components of the CMR model and insure they would meet the needs of the Economic Directorate for instrument design. Participants in the workshop included representatives from the Statistical Research Division, Fenestra Corporation, and the Economic Directorate lead to agreement that the CMR would meet the needs for Fenestra's electronic CSAQ application. The outcome of the workshop was general agreement that the CMR model would meet the needs of the Economic Directorate.

12. The CMR architecture actually supported three methods for CMR use by our customers. They could use the centrally managed metadata repository and we would provide all necessary support. This method would mean that the user areas would not have control of their metadata repository and would depend on the IT area for full support. The second method would be for a user area to use our CMR design, but build and support their own metadata repository. They could add new definitions to their copy of the model and generate them to the repository database, but to ensure continuing compatibility could not remove CMR definitions from the model. The third method was designed primarily for those systems that already had existing metadata repositories of some form or fashion. In this case, metadata would be exchanged between the CMR and the other repository using a metadata interchange. Our architecture uses XML as the standard interchange. We will use XML style language to build input and output conversions between other formats.

IV. BUILDING A PILOT APPLICATION USING THE CMR

13. At this point it was agreed that we would build a pilot application using one of the Economic Directorate Census Surveys. The Survey chosen for the pilot was the Annual Survey of Manufactures, which is an annual survey that very closely mimics the behaviour of the Economic Census forms. The pilot application for the Economic Directorate focused on two parts of their Census process. It was a value-added metadata input tool for use in both data collection and data dissemination. The Economic Directorate was already using an Internet based data dissemination tool named American Factfinder to disseminate 1997 Economic Census data. The American Factfinder was the first BOC application to use the CMR model to describe their data files.

14. The major early discovery in the development of this pilot application was that the Economic Directorate already had two existing applications performing a subset of the metadata repository. A DBASE based application already existed that used a comma delimited ASCII input file to build a metadata repository. This DBASE system then provided dissemination files for three different applications: the American Factfinder BOC Internet dissemination website, the Economic Directorate Economic Census CD-ROM data dissemination output product, and the publication data product output. The pilot project was designed to import the DBASE metadata into the CMR, then build an interactive tool which would allow the Economic Directorate to Create, Read, Update and Display metadata.

15. There also existed an Economic Census Reference Input File Control System (REFICS) which contained metadata in the form of parameters for the Census operations. REFICS was a very complicated application with many business rules enabling the Economic analysts to develop comparability between collected, preciously published, and data dissemination product codes. The pilot application would use the same update tool to provide a user-friendlier metadata editing facility for the REFICS system.

16. The methodology employed in this pilot development was recursive rapid application development. The developers met very regularly with the customers to develop requirements were developed and to review the implementation. Some of the keys to success were the usage of model based

code generation rather than custom programming. This allowed quickly making required additions to the model and regenerating the code.

17. The Economic pilot application was begun in summer of 1999 and completed in December of 1999. As a direct result of this successful pilot application, the Economic Directorate decided to build an Economic Metadata Repository (EMR) using the CMR model.

18 The metadata repository is designed as a place to register metadata objects and store information about those objects. The metadata can be created and even stored by other applications. The decision was made to store and manage all reusable 2002 Economic Census metadata in one central Economic Metadata Repository,

19. What we originally planned was to develop generic software that used documented application program interfaces to utilize the CMR. The idea was for the various business areas to develop their own applications taking advantage of the CMR infrastructure. What the Economic pilot application was beginning to make clear was a demand for the CMR team to actually partner with the business areas and develop the necessary applications. The side effect is to severely slow development of the core systems while resources are shifted to application development.

V. BUILDING THE ECONOMIC METADATA REPOSITORY

20. For the 1997 Economic Census all forms design was done on an individual form basis. Analysts from each business unit sent forms specifications to our Administrative and Customer Services Division to be designed. The designed form was sent back to the business unit for validation and correction. This method was very costly and time consuming. For 1997 there were literally hundreds of variations of the same question across forms. With many different analysts designing the forms there was no way to standardize questions or even be sure of how much duplication existed. Three forms were designed by Fenestra for electronic reporting; this was a separate effort from the paper forms design.

21. For 2002, the Economic Directorate planned to redesign their Economic Census processing. They made the decision to begin with engineering a new way to develop paper form surveys. The focus was on standardization of questions across forms, reuse wherever possible, and providing the designers the ability to generate forms with little or no turnaround time.

22. Now that it was agreed that an Economic Metadata Repository (EMR) would be built using the CMR model, some decisions had to be made. The metadata repository is designed as a place to register metadata objects and store information about those objects, but we now needed an architecture showing where the metadata was going to be created, viewed, edited. Areas to consider included:

- What additional metadata elements unique to the Economic Directorate were needed
- What application would be used to create and maintain the metadata
- How would central metadata be provided for application use
- How could cooperating applications change or add to the metadata
- What security needed to be provided and enforced?

23. It was agreed that information that is not sharable with other applications did not have to be returned to or reside in the EMR. This non-reusable metadata included such things as behavior in electronic surveys.

24. The Economic Directorate decided that they wished to store all metadata in the EMR. This required building a metadata input API to be used by other applications to access metadata in the EMR and return new or updated information to the EMR.

VI. ORCHESTRATING AN ARCHITECTURE BETWEEN VENDORS

25. In a series of internal Economic Directorate meetings, a specification detailing the expected appearance and functionality of screens for a Graphical User Interface (UI) was developed and provided to the CMR consulting team. Managers of the Economic Census redesign effort requested that the CMR team develop this UI. Evaluating the very detailed screens provided by this Economic Directorate team lead us quickly to the realization that we would have to hand code many of these screens if we matched the exact look and feel provided in the specification. Hand coding would simply take too long to meet their proposed schedules.

26. We decided to use Oracle Designer to generate sample web pages that would provide the requested functionality but not the exact look and feel of their very specific screen layouts. This generated set of screens was then presented to the developers of the specifications document and users by going thru the specification document and explaining how the generated screens would provide the meet their requirements. This was received very well and we were able to get agreement that code generation providing equivalent functionality to that specified in the specification would be acceptable.

27. At the same time the metadata user interface development was beginning, the issue of how to capture the huge amount of existing information describing the hundreds of paper forms to be used in the 2002 Economic Census. The idea was to input the legacy information describing all forms used in the 1997 Economic Census, validate it, then modify it to fit the 20002 Economic Census Forms. Most of that information either existed in a variety of ad hoc formats or simply in the minds of knowledgeable data analysts. It was further complicated by the fact that each area had developed over time their own unique way to define the content of their form(s). A method to provide this metadata requiring little or no user training had to be provided. It then had to be loaded into database staging tables and validated prior to actually moving the metadata into the production database.

28. The Economic Area was most interested in setting up a tightly controlled environment whereby an area working on one sector could not change metadata content owned by another sector. To further complicate the design, there were a number of roles

29. The method chosen was to have the users provide their metadata using Excel spreadsheets, a product most of our users are very familiar with and comfortable using. Little did we know that by the time this process was complete we would be processing more than 19 different spreadsheet formats. One of the great difficulties in this process was the ability of each area to customize their formats. Many times we would load a set of spreadsheets into the staging tables only to find a tremendous amount of invalid information. The analysts on many of those occasions found a new piece of information to provide, so they simply inserted a new column or re-arranged the order of the columns to match a format that they already had. This was an extremely time-consuming and frustrating process. It is made much more difficult by the fact that we were building a metadata driven application. This metadata driven approach provides huge advantages in the ability to quickly add or change survey forms or their contents. Unfortunately, incorrect or missing metadata has similar far-reaching effects.

30. The CMR project developed a very detailed architecture describing the parts of the system, how they would be developed, and how they would interact with each other. We now needed an architecture describing how the EMR and the GIDS system would exchange metadata.

31. In a series of meetings held with both vendors, the flow of metadata was to be as follows. The EMR would be the storage point for all sharable economic metadata. It would be based on an extended version of the survey and census business process model that is the basis of the CMR. Included in the EMR would be a complete description of the complete contents of each form. The metadata exchange format was agreed to be XML using a customized DTD. The GIDS system would return generated and custom designed sections of the forms as GIDS large object image files. The EMR UI would timestamp and store the generated GLOB in the EMR. Noting that the specification for the EMR graphical user interface is well over 100 pages in length, only a few of the features implemented can be described in this

paper. This detailed specification now contains the business rules for all Economic divisions for producing paper form questionnaires and should prove invaluable for future efforts.

32. The EMR UI supports creating, replacing, updating and displaying metadata. The model was extended to support keeping track of the contents of forms. The functionality included in this includes but is not limited to:

- Roles based security enforcement for all functions;
- The ability to add, modify, and view content for questionnaire forms;
- Automatic rules-based generation of data elements for questions;
- The ability to add, modify, and view complete forms;
- Notification screens listing all points of contact;
- Automatic XML interchange file creation and processing;
- The ability to lock questions and forms upon completion and unlock for required updates,
- Automatic launching of image viewers and external applications.

33. The Economic Directorate developed standard rules for formatting. The UI is highly customized for the Economic Directorate. In an effort to standardize questions which are asked many times and in the past have been asked in many ways, a set of standard questions was designed which are required to be used whenever those questions are asked on a form. These sections are sent to GIDS to be automatically formatted using a set of standard rules governing their layout.

34. The content of forms becomes quite complex when they are completely defined by metadata. Form components that had to be managed by the UI included:

- Headers provide metadata for header layouts to be designed from metadata passed to GIDS;
- Indent Levels provide a code that determines how questions appear on a form;
- Instructions provide sections of instructions detailing how to fill out the form;
- Special Inquiries B Additional questions added for special additional survey requirements;
- Answer groups provide grouping of answers, one example of which is name and address field;
- Write-ins provide free format answer sections.

35. The Economic Directorate plans to offer Computer Self-Administered Questionnaire (CSAQ) as an option for all 2002 forms. The paper content designed and stored in the EMR will be exported completely to the GIDS system. Most of the questions will remain unchanged, but behaviour must be added and historical data included in those areas that will utilize it.

VII. DATA CAPTURE OF ECONOMIC CENSUS 2002 FORMS

36. Probably the largest cost savings provided by metadata for the 2002 Economic Census will be derived from the ability to automate the key form image process using data capture. All questions will have a detailed data element definition and value domain associated with them. The basic data element definitions were generated by the User Interface and the Data Element Registry then used to complete all value domains and definitions.

37. The GIDS system will provide forms-specific layout information including the page number and coordinates for answer blocks. This information is returned via XML file and stored in the EMR. Users will use Excel or the Data Element Registry to edit and validate the data capture metadata.

38. The EMR UI will upon request extract all data capture information for a form from the EMR and send it to the Data Capture subsystem. The dollar savings in key from image will be huge. Blocks on a scanned form will be brought up on a screen for the operator to validate. Previous systems require the key operator to enter a key code identifying the item that is being keyed. This will no longer be necessary, as the key from image code will already know exactly what data element the result field is to

be stored in. Since the value domains for the data elements are known a tremendous amount of editing will be able to be automated. The result should be far more accurate data entry for much less money.

VIII. LESSONS LEARNED

39. The quality of metadata becomes both extremely important and extremely obvious when you build a metadata driven system. Spend as much time at the beginning validating and correcting the metadata content. It will save time in researching problems and then repairing the problems.

40. Hindsight indicates that we should not have used spreadsheets to input the legacy metadata, but should have written a user interface to allow interactive input as well as importing information from other sources.

41. Legacy loading a large volume of metadata takes much longer than anyone expects. The more time spent automating the validation of this load process will save far more time later in the project.

42. Re-engineering the survey process is totally new to the survey analysts. We found that even though we spent a tremendous amount of time working directly with the analysts to define their requirements it was a continuous learning process. This difficulty is actually in writing down correctly all of the business rules that survey analysts have been following for years. New requirements continued to surface throughout the 14-month development time for the EMR user interface.

43. It was essential to have the users actively working with the developers to build the application. It not only helped build a team, but both the developers and the users developed confidence in each other and the willingness to trust and agree to compromises for a common goal.

44. Convincing sponsors on the usefulness of metadata is difficult to do without some form of application to actually show metadata in use.

45. Building a full web enabled application caused a lot of difficulty for our support operations. The Economic Directorate computer support area is more organized around support of computers and operating systems. Building a system that makes heavy use of the network with Unix, VMS, and NT based computers is a fairly new concept. But expecting the entire system to remain up and running when multiple support areas support different parts of the system was difficult. I think this will ultimately lead to organizational change.

IX. FUTURE PLANS

46. It has taken more time that the Economic Directorate expected to build the EMR, the paper forms UI, the data capture system, and the electronic survey collection system. Other than assisting with data dissemination, further use of the EMR for Economic Census processing will have to wait until this cycle is completed.

47. There are many things in store for the CMR over the next two years. Realizing that we have to devote a lot of time to application development, we have had to slow the completion of the core CMR components. Currently we are completing a portal architecture that will allow us to quickly put up requested portal sites and ensure a uniform classification scheme. We are expanding our efforts to perform data cleaning for Geographic products sent to the American Factfinder, are working on metadata driven applications for other demographic products sent to AFF, and are building an application using the Data Element Registry to support the Field Directorate electronic survey data collection process.

48. The system that we are building for the CMR will utilize XML schema as the principal input and output mechanism for the CMR. We will use XSL to allow conversion between a wide variety of formats. We are using XML with a custom Census DTD to support metadata interchange between the EMR and GIDS system.

X. END RESULT

49. Services Division produced 200 final forms and are sending them to the printer in mid-March. The formal approval process for the printed forms has time consuming, but with automation has come the ability to make changes to forms very easily.

50. For the Economic Census 2002 project, all 200 Services sector forms must be completed and sent for printing by mid-March. Beginning in late February, the Manufacturing and Construction Division began using the EMR user interface to correct all of the 1997 and 2002 collected and published metadata using the comparability component developed to implement their business rules. They will be working to complete their paper Census survey forms by June of 2002. The Data Capture system began beta testing in March of 2002.

REFERENCES

- Appel, M. V., Gillman, D. W., LaPlant, W. P. Jr., Creecy, R. H. (1996), "Towards Unified Metadata Systems and Practices", ISIS-96, Bratislava, Slovakia, May 21-24, 1996.
- ANSI X3L8 - Data Representations (1999), "ISO/IEC 11179 Part 1 - Framework for the Specification and Standardization of Data Elements, International Standard, December, 1999.
- Capps, C. (1995), "Overview of the Technical Architecture for FERRET", Census Bureau internal document, Demographic Surveys Division.
- Census Bureau (1997), "Statistical Design and Survey Methodology Metadata Content Standard", Draft, Census Bureau Internal Document, April, 1997.
- Census Bureau (1996), "Table of Contents for Statistical Design and Survey Methodology Metadata Content Standard", Draft, Census Bureau Internal Document, July 2, 1996.
- Gillman, D. W. and Appel, M. V. (1994), "Metadata Database Development at the Census Bureau", Presented at the UN/ECE METIS Working Group Meeting, Geneva Switzerland, November 22-25, 1994.
- Gillman, D. W., Appel, M. V., and LaPlant, W. P. Jr. (1996), "Design Principles for a Unified Statistical Data/Metadata System", Proceedings of SSDBM-8, Stockholm, Sweden, June 18-20, 1996.
- Gillman, D. W., Appel, M. V., and Highsmith, S. N. Jr. (1997), “Building a Statistical Metadata Repository”, Second IEEE Conference on Metadata, Silver Spring, MD, September 16-17, 1997.
- Graves, R. B. and Gillman, D. W. (1996), "Standards for Management of Statistical Metadata: A Framework for Collaboration", ISIS-96, Bratislava, Slovakia, May 21-24, 1996.
- Highsmith, S. N. and Gillman, D. W. (2000), “Building a Metadata Repository at the U.S. Census Bureau”, ISIS-2000, Riga, Latvia, May 29-31,2000.
- LaPlant, W. P. Jr., Lestina, G. J. Jr., Gillman, D. W., and Appel, M. V. (1996), "Proposal for a Statistical Metadata Standard", Census Annual Research Conference, Arlington, VA., March 18-21, 1996.
- Lenz, H.-J. (1994), "The Conceptual Schema and External Schemata of Metadatabases", Proceedings of SSDBM-7, pp160-165, Charlottesville, VA, September 28-30, 1994.
- Rosen, B. and Sundgren, B. (1991), "Documentation for Reuse of Microdata from the Surveys Carried Out by Statistics Sweden", Research and Development Statistics Sweden, June 28, 1991.
- StEPS (1996), "Standard Economic Processing System Document 1: Concepts and Overview", Internal Census Bureau Document, April 16, 1996.
- Sumpter, R. M. (1994), "White Paper on Data Management", Lawrence Livermore National Laboratory document, 1994.
- Sundgren, B. (1991a), "Towards a Unified Data and Metadata System at the Australian Bureau of Statistics - Final Report, December 2, 1991.
- Sundgren, B. (1991b), "Statistical Metainformation and Metainformation Systems", R&D Report Statistics Sweden, 1991:11.

- Sundgren, B. (1992), "Organizing the Metainformation Systems of a Statistical Office", R&D Report Statistics Sweden, 1992:10.
- Sundgren, B. (1993), "Guidelines on the Design and Implementation of Statistical Metainformation Systems", R&D Report Statistics Sweden, 1993:4.
- Sundgren, B., Gillman, D. W., Appel, M. V., 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.
- Wright, G., Alred, S., Dhritiman, S., Ravichandar, D. (1999), “CMR Technical Architecture”, Internal Census Bureau Document, December, 1999.
- Wright, G., Alred, S., Dhritiman, S., Ravichandar, D. (2000), “CMR Software Technical Architecture”, Draft, Census Bureau Internal Document, January, 2000.