

SEMINAIRE

E+; 3=! C

SEMINAR

STATISTICAL COMMISSION AND
 ECONOMIC COMMISSION FOR EUROPE



Distr.
 GENERAL

CONFERENCE OF EUROPEAN
 STATISTICIANS

CES/SEM.43/10
 30 March 2000

ENGLISH ONLY

Seminar on integrated statistical information
 systems and related matters (ISIS 2000)

(Riga, Latvia, 29-31 May 2000)

Topic I: Data warehousing and the development and use
 of statistical databases in a network environment

**BUILDING A CORPORATE METADATA REPOSITORY AT THE
 U.S. BUREAU OF THE CENSUS**

Contributed paper

Submitted by the U.S. Bureau of the Census¹

I. INTRODUCTION

1. 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.

2. Figure 1 illustrates some of the many systems at the BOC which 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

¹ Prepared by Samuel N. Highsmith and Daniel W. Gillman.

available tools such as CASES from the University of California at Berkeley. We have a wide variety of survey and census data collection tools which 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 diagram 1 is that these many systems are performing similar functions without being able to share with each other.

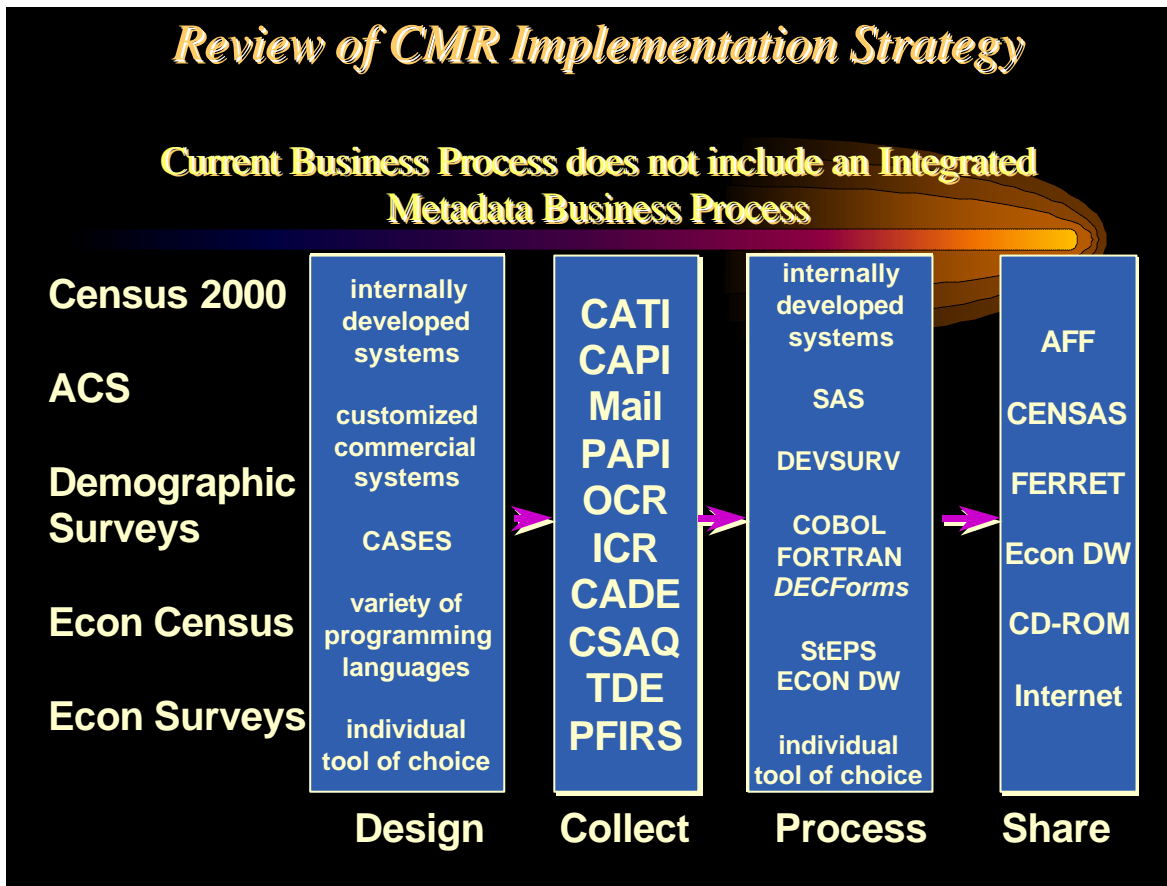


FIGURE 1

3. We have been developing a generic model driven solution to this problem for many years. This solution is now officially designated the Corporate Metadata Repository (CMR) at the BOC. The CMR offers the promise of being able to describe a survey or census throughout the business cycle such that any application capable of interfacing to the CMR will be able to access and immediately use any survey or census information registered in the CMR. Imagine being able to document your survey completely one time and then having any system capable of understanding that documentation easily access and use your survey information with no change to the application. This is our goal.

4. The CMR will include the ability to determine quickly how other parts of the organization solve their survey or census problems. It will enable our users to determine quickly if a survey already exists to perform part or all of what is being planned and/or developed. It supports reuse of whatever solution another area has

deployed. Examples of this reuse range from simply finding and using an existing question to finding and reusing a complete survey operation.

5. There are many things a populated CMR could be used for. For example:
- What if you could perform a search and find all the questions dealing with "age" that that have been asked on demographic surveys within your organization, and then quickly find all the allowable answers to each question?
 - What if you could perform a search and find all the questions dealing with "income" that have been asked on economic surveys within your organization, and then quickly find all the allowable answers to each question?
 - What if you could quickly copy any question and its set of allowable answers to a new survey?
 - What if you could look at one of those variables and quickly determine what questions were asked to produce that variable? Or even what business rules were applied to produce a final result from intermediate values?
 - How would you like to have a data dissemination software tool that can immediately use your new survey output dataset with no programming changes?

6. This paper will describe a way to accomplish the above and much more. Admittedly, the major drawback is that we have actually to enter the metadata, which will allow this system to flourish. I suggest that the only way this "metadata entry" will occur is to provide major value added to the users; it must make their job easier and more productive for such an environment to take hold. Finally, it must only be requested one time. If I enter my metadata in your requested format, the system must be smart enough to reformat it automatically and output it in your choice of formats.

II. BACKGROUND

7. Building a Corporate Metadata Repository at the BOC is a fairly new concept. It is grounded in a significant amount of research and collaboration. The research is based on several things. First, participation by the BOC in work with Sweden, Canada, Australia, and the UN/ECE Metadata Workshop led to the idea of developing a business data model for survey and census processing. There were several potential techniques including fully distributed and centralized solutions which could be applied to the development of a metadata repository (see "Guidelines on the Design and Implementation of Statistical Metainformation Systems"). The CMR at the BOC is designed to be for metadata what a card catalog is to a library. The CMR contains the location and characteristics of whatever is registered within it. Thus the CMR will contain the location of a registered survey dataset, not the data itself. It will contain pointers to the documentation as much as is practical. It does support but not encourage registration and storage of documents within the CMR itself. This is an extremely important concept; the CMR will store the location of information rather than trying to be the central repository of all information.

8. In collaboration with many stakeholders across the Census Bureau and under the guidance of a consultant well versed in metadata repository and data warehouse technologies, a business data model (BDM) describing survey and census processing was developed. The results of this team effort demonstrated that the processes used in our Economic, Census, and Demographic Survey organizations were very similar. The result was successful development of one model describing survey and census processing at the BOC.

9. We simultaneously contracted with a private organization, Metadata Management Incorporated, to build a formal data element registry product based on ISO/IEC 11179. This data element registry was then combined with the BDM into one entity

relationship model using a tool named Erwin. From this model we could automatically generate a database implementing the model.

10. Our research organization held a series of informational seminars to share this vision of a metadata repository with the rest of the Census organization. The feedback from these seminars was very clear. Very few people understood how this metadata repository would help them in their job. It was simply viewed as a somewhat interesting research project and was not taken very seriously in the production areas of the BOC. The preceding development phase covered several years.

11. Now came the stroke of good fortune that this project really needed to get started. The Decennial staff, which had the task of building a data dissemination system for the 2000 Census data, discovered that to deploy successfully literally hundreds of datasets in their internet data dissemination system would require some form of metadata repository. After looking at what our model had to offer, they decided to deploy the CMR model rather than develop their own. They did decide to extend our model to add unique functionality and provide performance gains required by their application. The current American FactFinder application, accessible from the central WWW.CENSUS.GOV web site, is entirely metadata driven from a repository based on the CMR model

12. With the limited success that our informational seminars were having in the production areas of the BOC, it was apparent that we had to show exactly HOW a metadata repository would be useful rather than the theory behind it. Construction then commenced on a prototype application to demonstrate how a metadata repository would actually add value to survey processing at the BOC.

III. BUILDING A PROTOTYPE METADATA REPOSITORY

13. In December of 1996, the Statistical Research Division decided to develop a prototype application of the envisioned CMR. The goals were to:

- Show seamless access to at least three existing data dissemination applications at the Census Bureau,
- Demonstrate the ability to seamlessly search across these three disparate applications,
- Show the advantages of a generalized metadata classification scheme, which we likened to the "Table of Contents" of a book,
- Utilize a completely web-based approach,
- Use an open rather than proprietary solution.

14. The applications to interface with were the American FactFinder application, the FERRET data dissemination tool, and the Economic Directorate Document Management system.

15. Having developed a prototype, we decided to demonstrate the potential capabilities to major stakeholders in the Bureau. We were very fortunate in being able to sign formal agreements with four major parts of our organization. These areas were the Decennial, Demographic, Economic, and Publications parts of the BOC.

IV. PILOT PROJECTS

16. It became apparent that to succeed in building a production system we needed to form partnerships with internal BOC customers and show a real benefit specific to their existing processes. Our Economic Directorate was in the process of completely redesigning 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 two survey data

collection instruments for 1997, which the Economic Directorate deployed and was very pleased with. The Economic Directorate subsequently contracted Fenestra to build a CSAQ for all 450 questionnaires contained in the 2002 Economic Census. Realizing that hand coding the description of 450 questionnaires would be a task so daunting that it begged another solution, Fenestra proposed building an electronic metadata repository covering the data collection instrument.

17. At this point, the Economic Directorate embraced the CMR for storing metadata required by their electronic CSAQ. The primary advantage in using the CMR would be that it modeled the entire survey process, of which data collection was only a part. A two-day workshop involving 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 idea of a pilot application was born.

18. In early development you frequently hear of prototype applications. Indeed, the prototype application previously described in this paper demonstrated the capability of a CMR. A prototype is a throwaway application; you must plan to discard it and use the knowledge gained to move forward. A pilot application is quite another story. The pilot application planned with the Economic Directorate would develop real functionality that would be used in production by the Economic area. That pilot functionality would be supported and migrated to future production applications.

19. In the case of the Economic Directorate the pilot application would focus on two parts of the Census process. It would be a value-added metadata input tool to cover both data collection and data dissemination. The Economic Directorate was already using a CMR based application, the American FactFinder, to disseminate 1997 Economic Census data. 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 an ASCII delimited text input file to build a metadata repository which then was output to three different applications: the American FactFinder, the Cdrom data dissemination output product, and the publication data product output. The pilot project would input the DBASE metadata to the CMR, then build an interactive tool to allow the Economic Directorate to Create, Read, Update and Display (CRUD) metadata. There also existed an Economic Census Reference Input File Control System (REFICS) which contained metadata in the form of parameters for the Census operations. The pilot application would use the same CRUD tool to provide a more user friendly metadata editing facility for the REFICS system.

20. The Economic pilot application was begun in summer 1999 and completed in December 1999. It has been received very well and the analysts are now starting data entry using the product.

V. THE PRODUCTION CMR

21. The next step was to develop an architecture defining the CMR, all interfaces, and all support tools that we proposed to build. The primary requirements for this environment included to:

- Provide an Open Architecture,
- Adhere to [Open Standards](#),
- Adhere to BOC Security Requirements,
- Support [Web Browsers](#) for CMR Web-based Apps,
- Use an Integrated Software Solution,
- Allow integration with [Emerging Industry Solutions](#),
- Use [COTS](#), where possible to keep costs down and custom development to a minimum,

- Use **BOC site-licensed software** or s/w with a high # of BOC seats, where possible to keep costs down,
- Provide an **Open API**,
- Provide an Open standards-based Metadata Interchange,
- Support **Metadata Interchanges** between **CMR** and the **other BOC systems and software**,
- Support CMR accepted **input metadata formats**: XML and AFF,
- Provide an **extensible CMR meta-model** which complies with ISO/IEC 11179,
- Provide a means of **sharing the CMR meta-model** within the BOC agency,
- Provide a means for **integrating unstructured metadata** with the CMR.

22. The following slides are courtesy of the Oracle consulting team that is building the production CMR for the BOC. Figure 2 shows the general design for the production CMR that is currently under construction. This shows the survey and census business processes from survey design to data collection through processing and into data dissemination. Note that there will be a number of metadata interchanges between the metadata held in the business areas and the CMR. At the bottom we show an Economic Metadata Repository which is planned to hold all the Economic Directorate metadata including extensions that we currently do not plan to include in the CMR. With the metadata in one standard format the metadata dissemination depicted will allow reformatting and outputting metadata in a variety of formats.

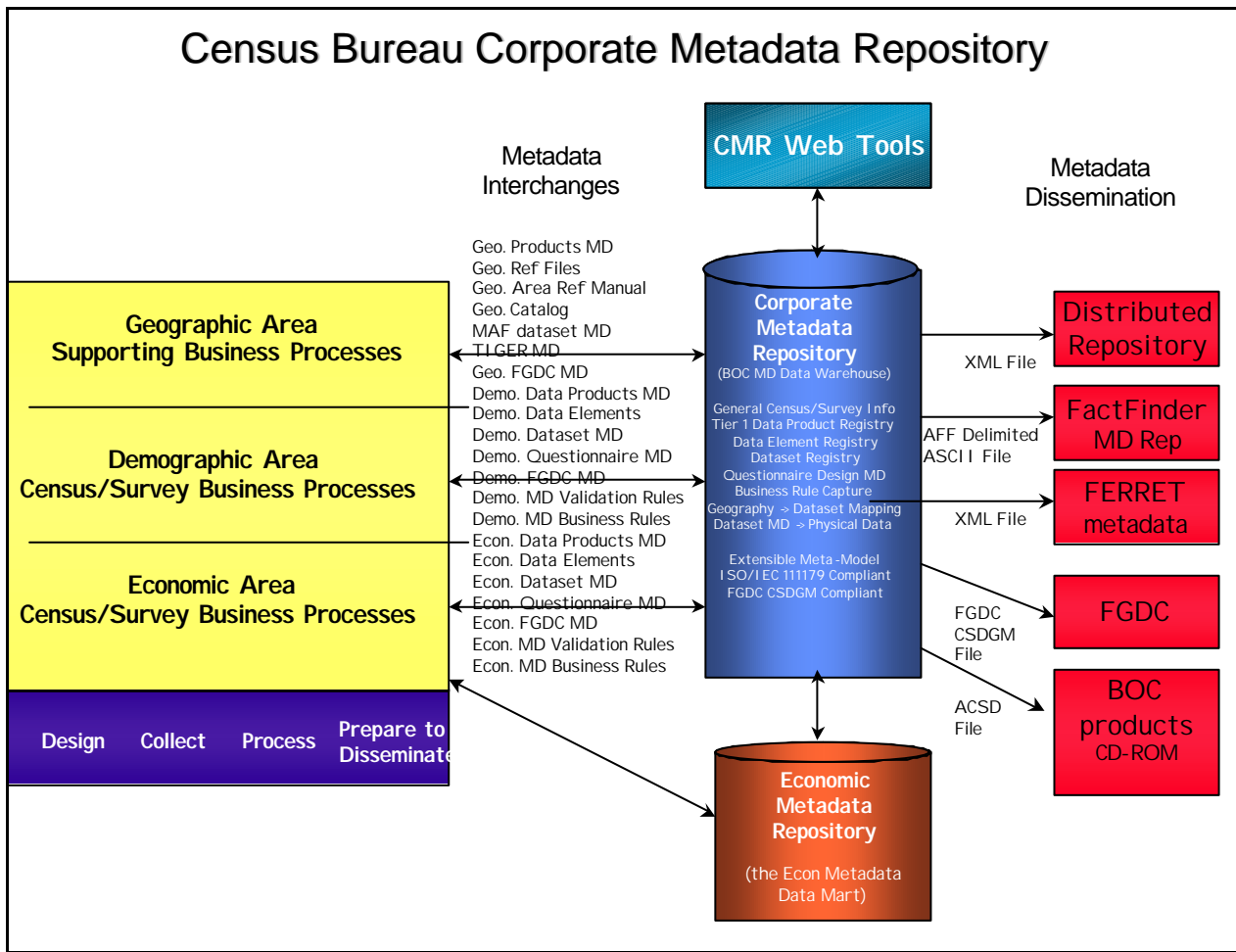


FIGURE 2

23. Figure 3 shows the general CMR architecture that is currently in development. The basic idea is to build a layered product that is as flexible as possible and is based on open, standards based components. Although we may initially have to develop part or all of some components for which there is no current commercial off the shelf software (COTS), we plan to be able to swap those components for COTS products if and when they become available. By providing a published high-level API we plan to be able to add required components to the model, regenerate the CMR and support applications with minimal impact to applications written to the high level API's.

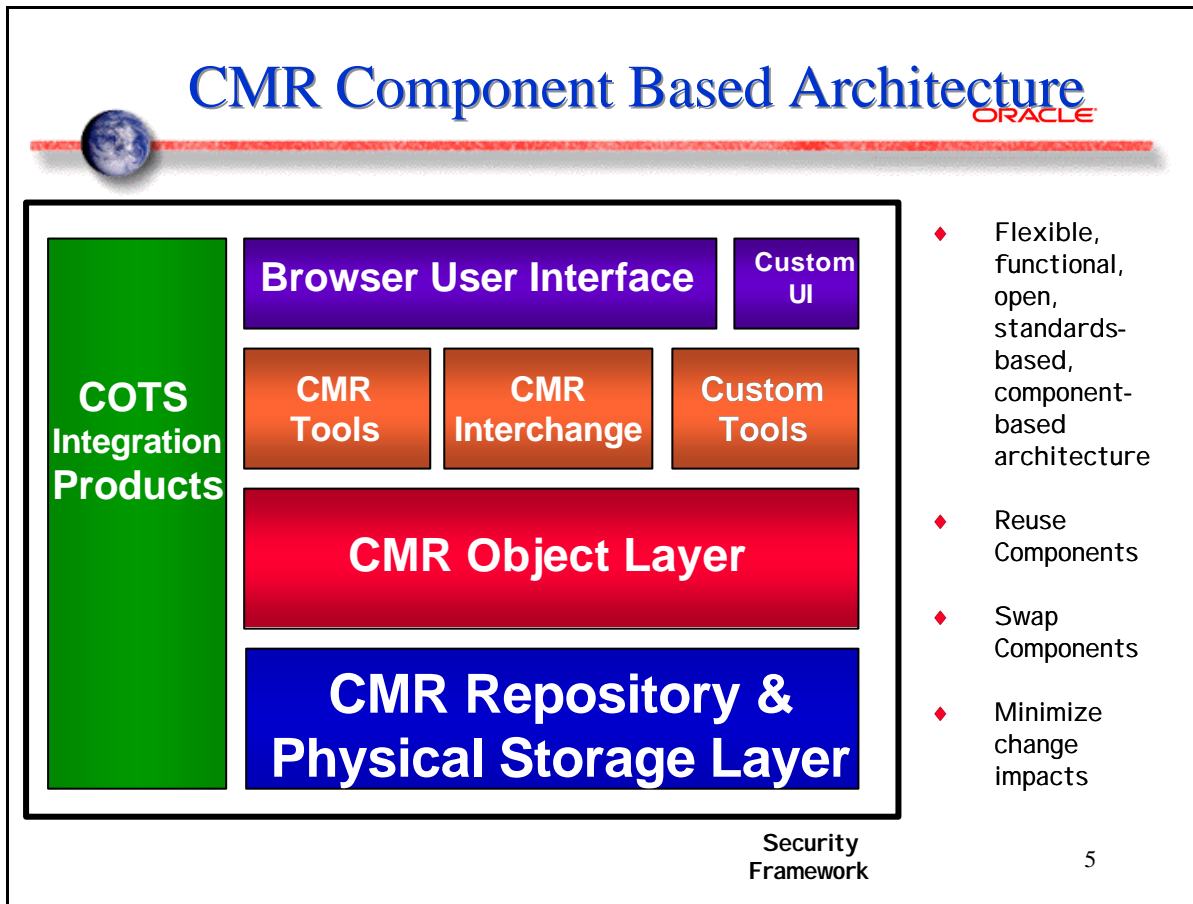


FIGURE 3

24. As reflected below in Figure 4, the CMR physical storage layer will hold the BOC application development, business, and technical metadata. This is an Oracle database generated from the model. Currently documents can be stored elsewhere and referenced by URL's, but ultimately a document management system that can be accessed by the CMR will be required to maintain and control all the documents. This slide is depicting storage of business and technical metadata, including pointers to external registered documents, in a physical repository. We do not envision being able to or wanting to support every need in every part of the BOC. Rather, just as American FactFinder and the Economic Metadata Repository have done, we plan to support the concept of departmental metadata repositories or metadata marts. We will build the software, which will move metadata between the CMR and the various metadata marts. This allows departments to add new definitions and concepts not used or needed by other parts of the BOC.

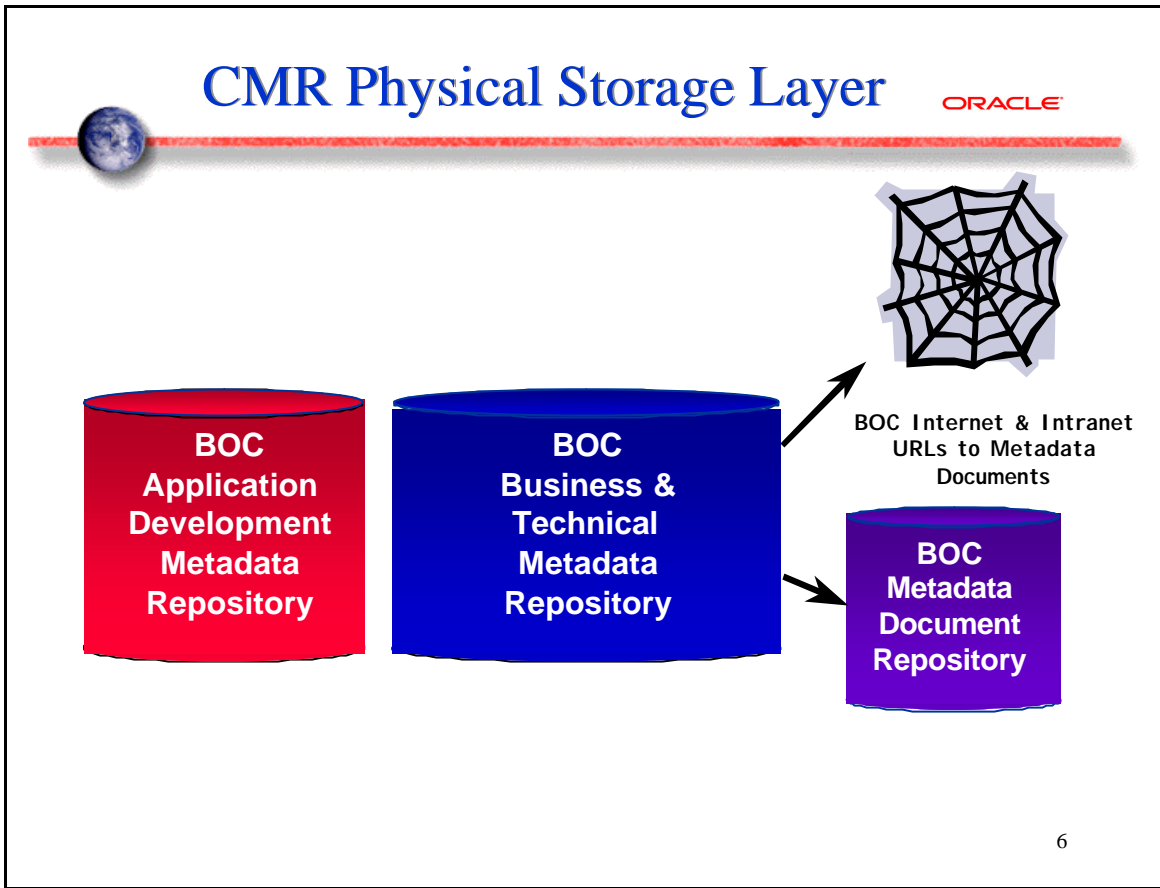


FIGURE 4

25. Figure five shows the methodology we plan for a distributed CMR environment at the BOC. Recognizing that needs vary across departments, we are building and will support a logically central CMR. Users will be able to use the tools and applications provided with the CMR and directly use the CMR to support their metadata requirements. But departments that wish to extend the CMR and perhaps add application specific metadata or tune their implementation can choose to build their own metadata repository mart. This is the approach that the American Fact Finder application has taken. The only real requirement is that all the structures and definitions in the CMR not be removed. Figure five also shows interchange between the CMR and existing applications that already have defined metadata structures. This metadata interchange would be via a mapping operation. The most loosely coupled implementation of a metadata mart would be use of a metadata interchange standard. We believe XML will be able to fill this role and plan to take advantage of Data Documentation Initiative (DDI) under development by the Inter-university Consortium for Political and Social Research (ICPSR).

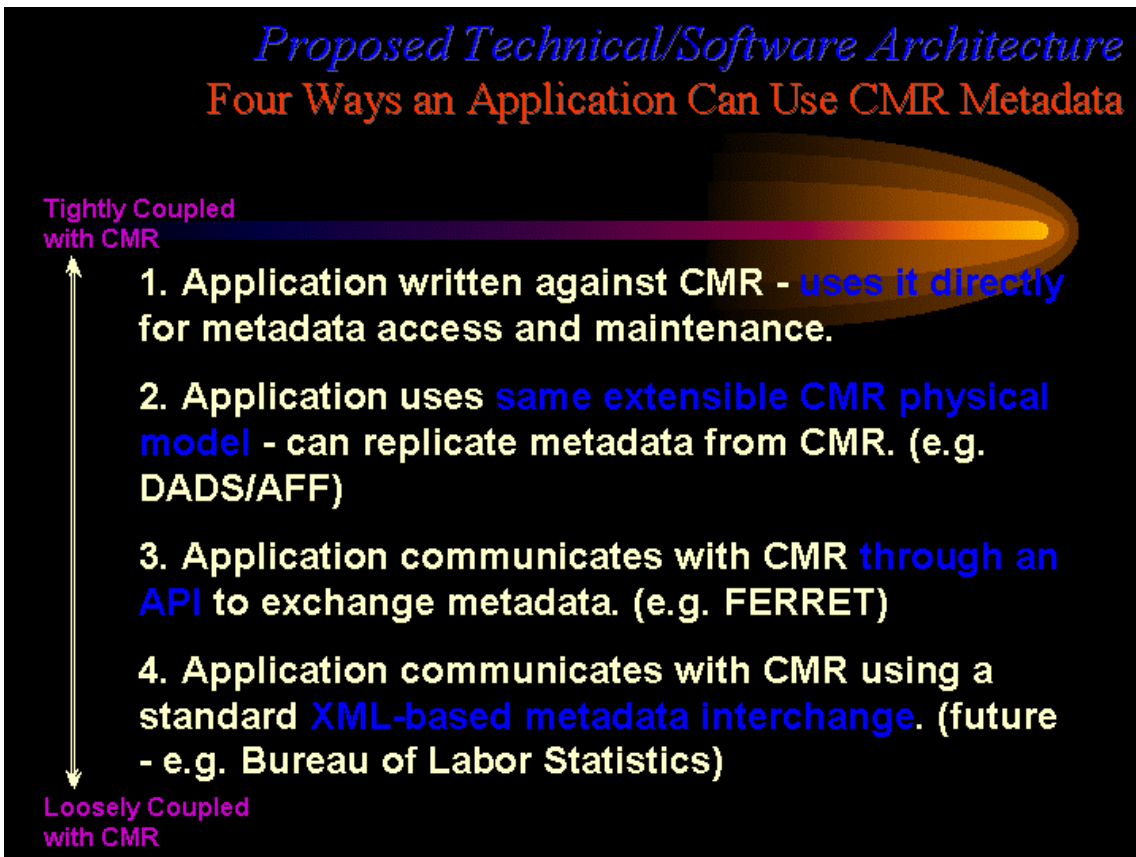


FIGURE 5

26. Figure six shows the Object layer currently under development. The goal is to build an interface layer that contains a published application program interface (API) for applications and a table API for the generated CMR toolset. What this architecture will provide is the ability to build access and security on top of but separated from the physical database structure. This allows using object interfaces to a relational database structure.

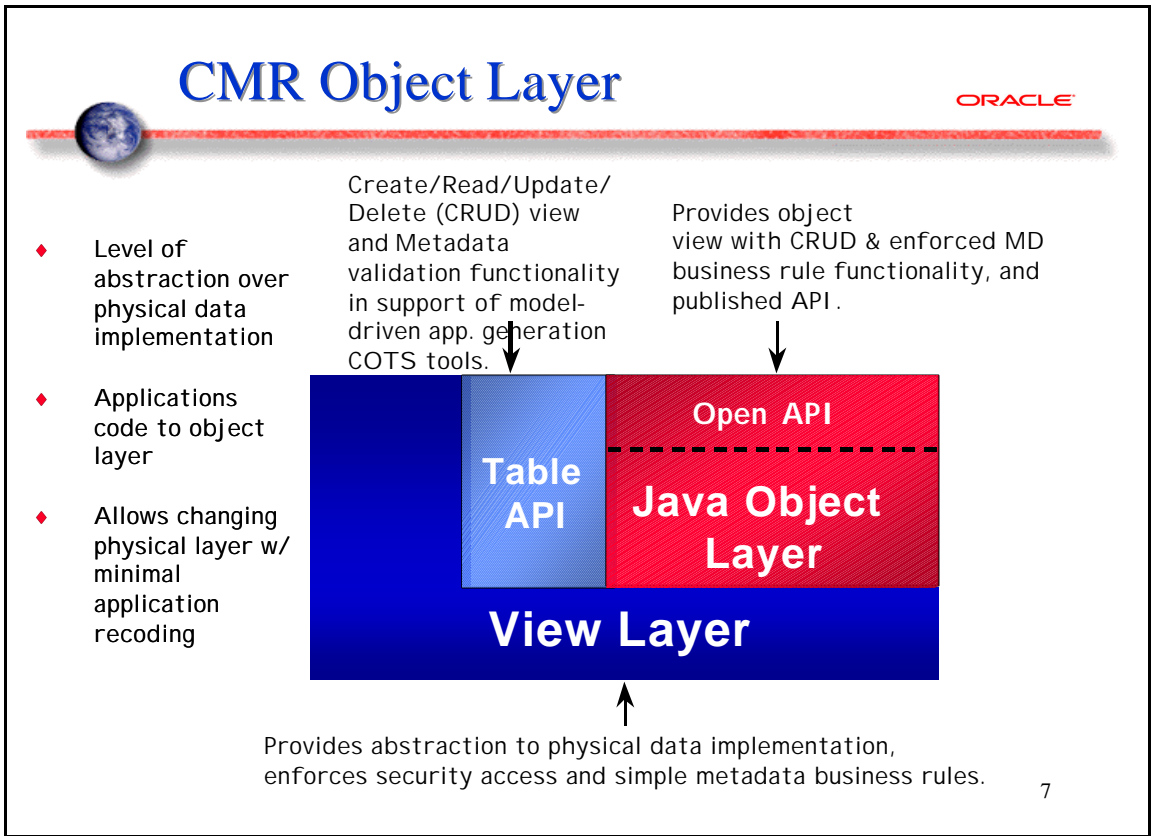


FIGURE 6

27. The CMR tool layer depicted in Figure 7 is extremely important to the success of the CMR. Experience shows that without tools an underlying infrastructure such as the CMR is practically useless or at least difficult for customers to comprehend the benefit of. The CMR tools must provide the ability for the departmental user to administer completely his/her metadata. We are treating metadata as a departmental resource under the control of that department. If the department decides to share their metadata with the rest of the organization that is great and supported. If they choose to keep it under their control and not share that is also supported. The CMR Interchange tools shown are also critical to success. Not only must the CMR interface with and exchange metadata with legacy systems existing at the BOC, the CMR must interface with any other external system. The methodology that we are using is to build an XML-based input/output interface and add XML to external legacy system interfaces where required. We plan to be able to interface to external systems using XML based interchange methodologies as they become available. The final tool interface will be a variety of BOC custom tools developed primarily to satisfy specific departmental requirements. These custom tools might be SAS based or they might be specific tools developed by other vendors to allow other products to take advantage of information contained in the CMR.

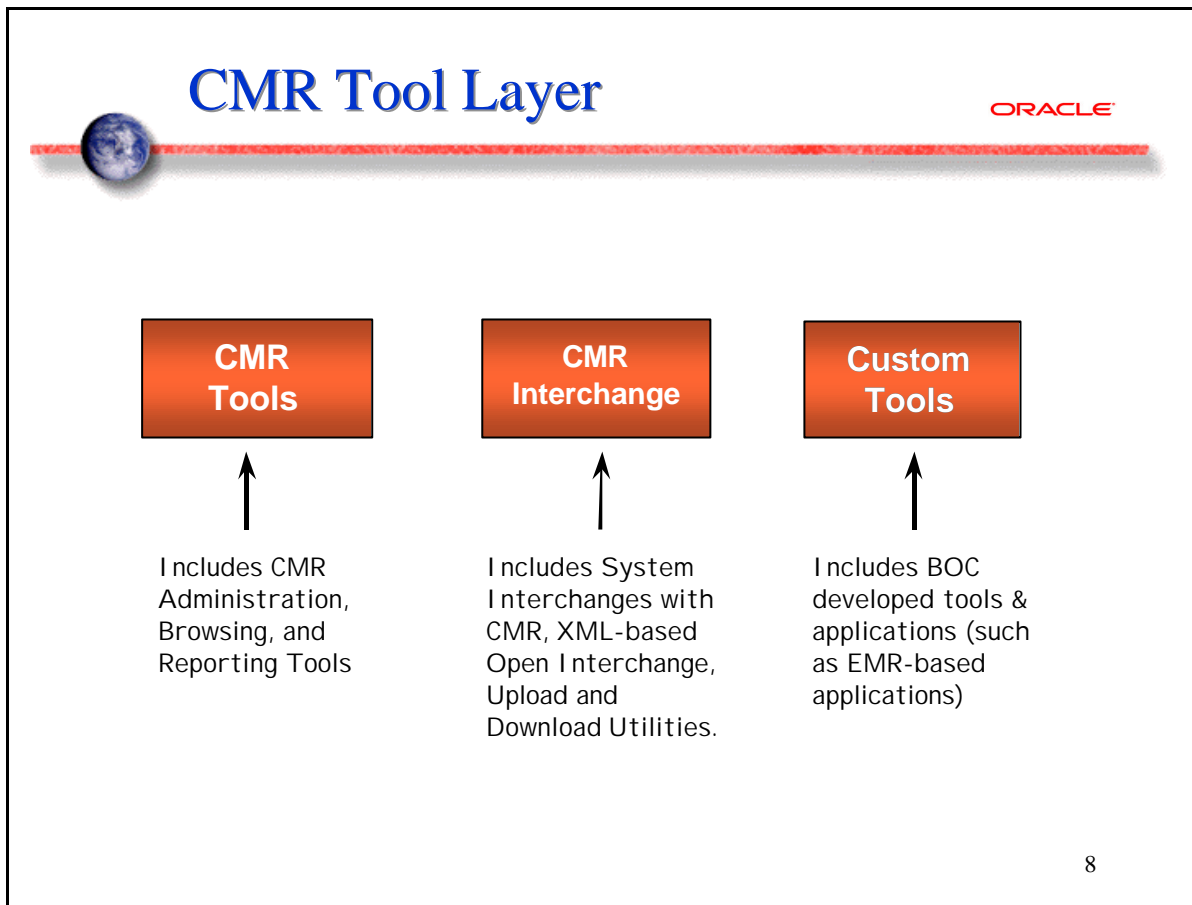
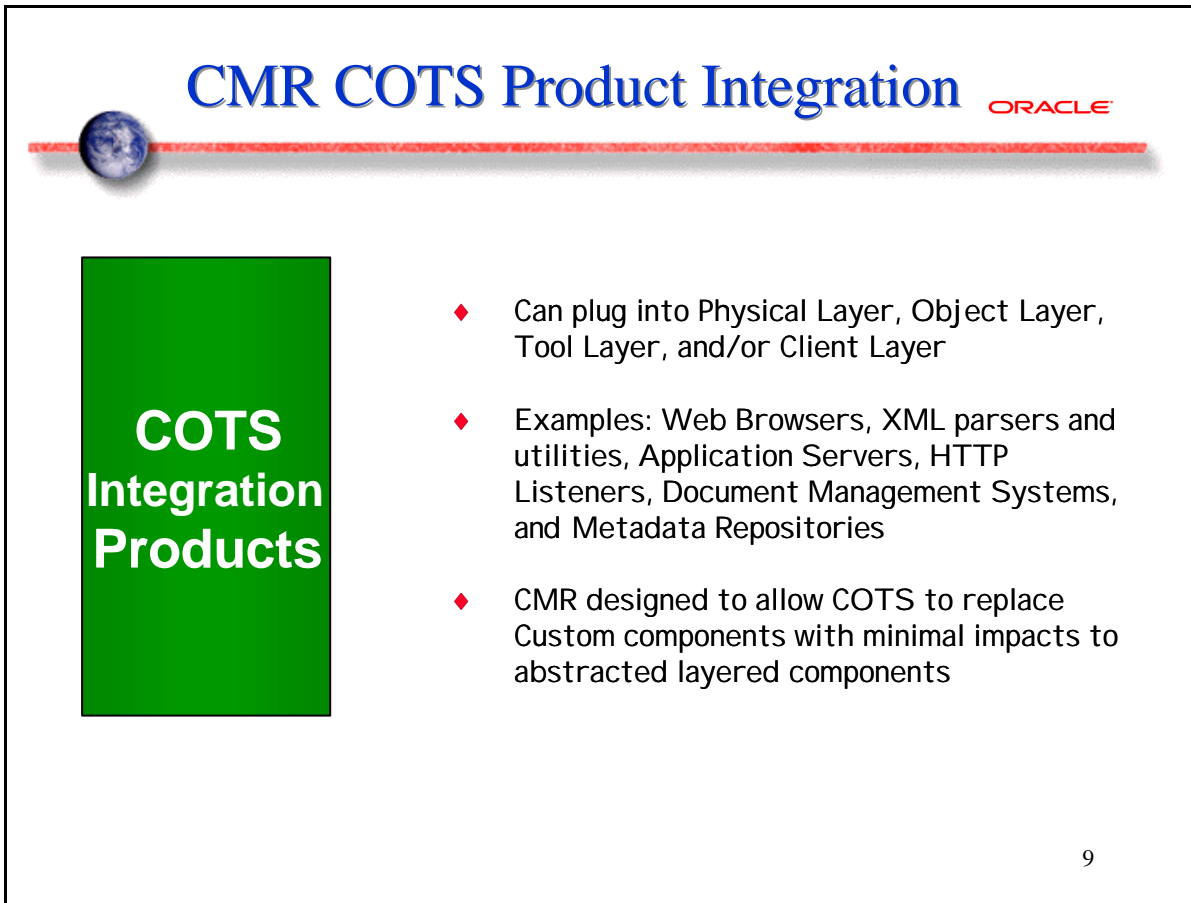


FIGURE 7

28. Last but not least in this production system is the allowance for COTS product integration, shown in Figure eight. This will be a web-enabled application. It must provide the capability to replace components with new COTS solutions as they become available with as little impact as possible to the applications utilizing the CMR. Direct access to the physical layer will be supported but strongly discouraged for ease of migration reasons. Far preferable will be access to the object layer as described above, the tool layer, or the client layer.



The slide features a title 'CMR COTS Product Integration' in blue, with the Oracle logo to its right. A red horizontal line with a globe icon on the left spans across the top. On the left side, there is a green vertical rectangle containing the text 'COTS Integration Products' in white. To the right of this rectangle is a bulleted list of three items, each marked with a red diamond. The bottom right corner of the slide contains the number '9'.

CMR COTS Product Integration ORACLE

COTS Integration Products

- ◆ Can plug into Physical Layer, Object Layer, Tool Layer, and/or Client Layer
- ◆ Examples: Web Browsers, XML parsers and utilities, Application Servers, HTTP Listeners, Document Management Systems, and Metadata Repositories
- ◆ CMR designed to allow COTS to replace Custom components with minimal impacts to abstracted layered components

9

FIGURE 8

29. The final piece is the extremely important CMR User interface layer shown in figure nine. This layer will support both Netscape Communicator and Internet Explorer browsers and application plug-ins. Also supported although not encouraged will be user interfaces which are not web-based. Non web-based applications will be much more useful to internal BOC users than to external organizations.

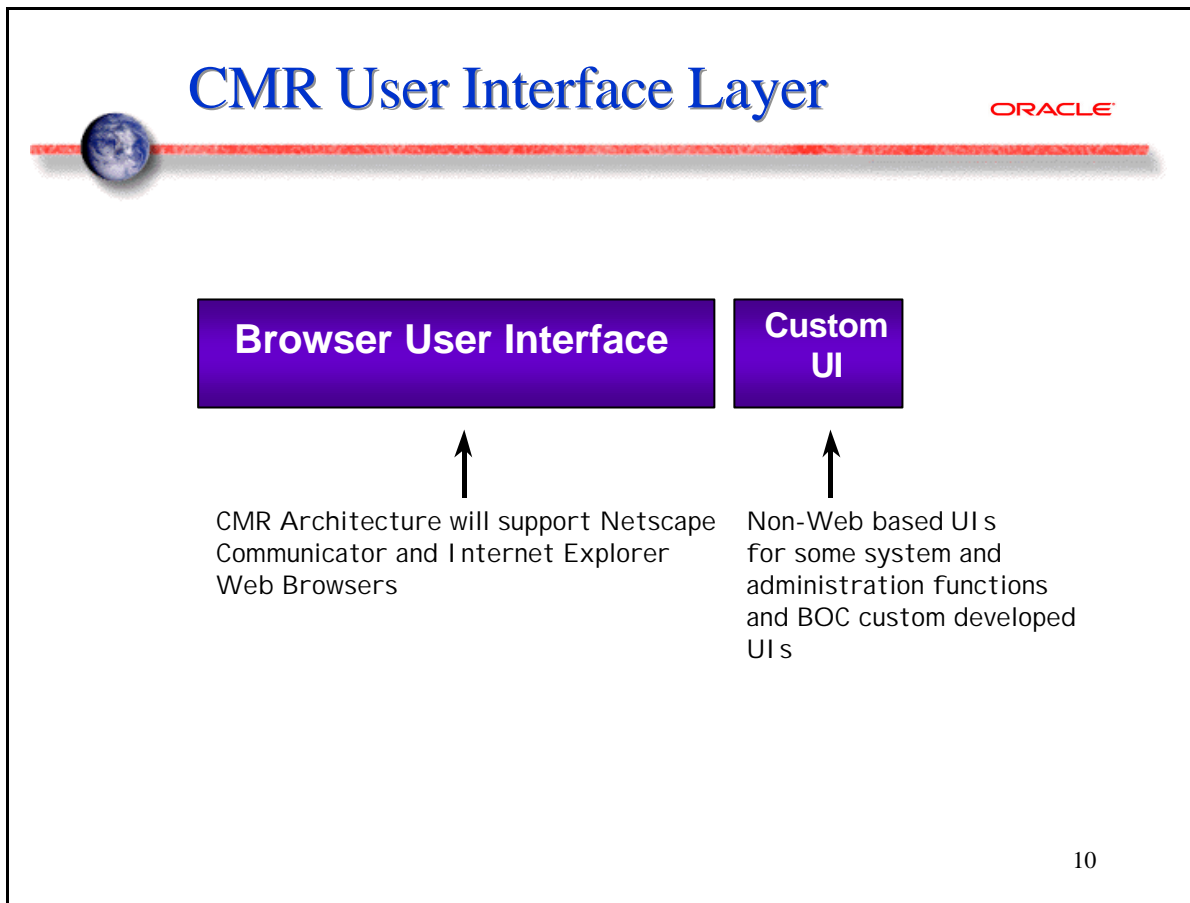


FIGURE 9

VI. CONCLUSION

30. The CMR should be completed and in production use by early 2001. In fact, the Economic Directorate at the BOC is already using it for production use in the Annual Survey of Manufactures. It will provide an infrastructure that will support many data exploration, data manipulation, and data dissemination applications. After you register data sets and their accompanying metadata any application that is able to communicate with the CMR will be able to utilize fully that dataset and all of its' metadata with absolutely no change to the application. In short, define it once and use it forever.

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.
- 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.