

SEMINAIRE

E+; 3=! C

SEMINAR

STATISTICAL COMMISSION AND  
 ECONOMIC COMMISSION FOR EUROPE



Distr.  
 GENERAL

CONFERENCE OF EUROPEAN  
 STATISTICIANS

CES/SEM.43/10 (Summary)  
 3 March 2000

Original: ENGLISH

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 Census<sup>1</sup>

**Summary**

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. 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 in such a way 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

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

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.

3. This paper will describe a way to accomplish the above and much more.

4. The CMR is grounded in a significant amount of research and collaboration. 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 development of a metadata repository. The CMR at the BOC is designed to be for metadata what a card catalog is to a library. It contains the location and characteristics of whatever is registered within it.

5. 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.

6. 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.

7. Now came the stroke of good fortune that this project really needed to get started. The Decennial staff, given the task of building a data dissemination system for the 2000 Census, 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](http://WWW.CENSUS.GOV) web site, is entirely metadata driven from a repository based on the CMR model

8. In December 1996, the Statistical Research Division decided to develop a prototype application of the envisioned CMR. The applications to interface with were the American FactFinder application, the FERRET data dissemination tool, and the Economic Directorate Document Management system.

9. We used the prototype application to demonstrate value to the program areas of BOC. The result was a series of signed memoranda of understanding to jointly pursue building the CMR.

10. At this point, the Economic Directorate embraced the CMR for storing metadata required by their electronic CSAQ. The idea of a pilot application was born. 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 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.

11. 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 were 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.

12. The CMR, supporting tools and interfaces 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 fully utilize that dataset and all of its' metadata with absolutely no change to the application. In short, define it once and use it forever.