PROVIDING DOCUMENT RETRIEVAL THROUGH A METADATA REPOSITORY AT THE CENSUS BUREAU

Gregory J. Lestina, Daniel W. Gillman, Martin V. Appel
U. S. Bureau of the Census, Washington DC 20233

1 Introduction

The information age is not only about new advances in computer technology, but it is about restructuring our businesses to meet the needs of people and take advantage of electronic distribution of information. One organization affected by the impact of the information age is the U.S. government. Since the beginning of the 1990s, the U.S. Government has been trying to streamline or "reinvent" itself. The Information Technology Management Reform Act of 1995 requires the heads of 24 designated federal agencies to each appoint a Chief Information Officer (Williamson, 1996). This and related legislation is intended to support the creation of the National Information Infrastructure, a federal initiative intended to provide the public with inexpensive and timely access to information. The role of the Chief Information Officer will be to add business practices of the private sector to the federal government. Among these practices are business process reengineering, investment analysis, and focusing on the needs of the public. Such practices may take years to develop, but it is necessary if the government wants to compete in the information age.

The U.S. Bureau of the Census (Census Bureau) is also preparing itself for the information age. It has taken advantage of the World Wide Web to distribute statistical datasets, advertise its products, and provide documentation on activities and products. It is also developing a repository for storing statistical metadata. Statistical metadata is the information and documentation needed to describe and use statistical data sets for the lifetime of the data. Statistical metadata can be stored and retrieved in a repository just as the data it describes is stored and retrieved in a database. It is the electronic storage and organization of statistical metadata which will allow statistical agencies to develop automated survey design and processing systems (Gillman, Appel, 1997).

In all organizations there is a need for the effective storage and retrieval of documentation. By automating documentation storage and retrieval, the cost of performing these tasks is reduced significantly. If there is no electronic system it is estimated that employees spend about eight hours per week or 20 percent of their time storing and retrieving documents. With an electronic document management system, employees will spend 5 percent of their time retrieving and storing documents (Popkin, 1995). At the Census Bureau, effective document management is critical for maximizing efficiency for employees as well as satisfying external users' requests.

This paper describes how a statistical metadata repository can be used to help store and retrieve documentation metadata effectively at the Census Bureau. The paper describes repositories currently available for storing and retrieving documents at the Census Bureau. It discusses how a documentation repository can be administered and how a logically central metadata repository can be used to classify and catalog information about documents. It also discusses new technologies available for implementing document storage and retrieval through metadata repository and possible future directions for the Census Bureau in implementing such a system.

2 Document Repositories at the Census Bureau

The Census Bureau calls itself the "preeminent collector and provider of timely, relevant, and quality data about the people and economy of the United States" in its mission statement. It is one of the world's leading statistical agencies and is known primarily for the nationwide census held every ten years. This head count or census is required by the Constitution so that seats in Congress can be reapportioned based on the population. But a decennial census has much more impact on the nation. The statistics collected from the decennial census and the more than 100 current survey programs are used as a benchmark by many governments and businesses for making strategic decisions. For example, local governments use Census data to make decisions about building roads or supporting public schools or providing legal evidence. Marketing and retailing companies use Census data to find customers and learn more about them. Other users of Census data include colleges and universities, libraries, and religious organizations (American Demographics, 1995).

The Census Bureau survey programs are responsible for collecting, processing, analyzing, and distributing data. They also plan and design survey programs. Therefore, the Census Bureau can claim to be competent at all phases of survey work.

As a result, the Census Bureau has accumulated a large amount of memorandums, specifications, computer programs, descriptions, models, and reports about surveys and data throughout the years. These documents are stored in different formats (e.g. paper, html files) and different media (e.g. file servers and metal binders). Many documents are not electronically indexed or classified. Some documents are already indexed in some type of document management system.

Each existing document management system at the

Census Bureau manages the collection of documents based on a particular survey or groups of similar surveys. For example, large amounts of demographic survey documentation are stored in the FERRET Documentation Management System (DMS). This system is a Web-enabled search tool for documents in the Current Population Survey (CPS). A user has the option of moving through a search guide that accesses html documents referenced by a database using CGI scripts. There is also an option of using a WAIS search utility for keyword searches. FERRET DMS was one of the first document repositories with Web access at the Bureau. This system is primarily used by researchers and subject matter experts who need to know survey level and variable level documentation about CPS. Many of the documents found in the FERRET DMS can be located through the Access Tools item on the Census Bureau's web site.

Documents from economic surveys and censuses are indexed and accessed through DOCS Open by PC DOCS, Inc., Burlington, MA. DOCS Open is Open Document Management API Task Force (ODMA) compliant, meaning that it is compatible with all major database systems and other ODMA compliant products. It is also interoperable with most PC applications. At the Census Bureau, DOCS Open is installed on a Windows NT server with a SQL Server 6.5 database. It currently imports documents in mostly WordPerfect 6.1 and Lotus 1-2-3 formats. DOCS Open was installed in mid-1995 and currently has 149 users. It also provides a search tool that searches rows in the database and a keyword search function that is indexed by the Verity search engine. DOCS Open is popular because it is easy to use and the underlying database is compatible with external applications. A web server was recently purchased by the Census Bureau that will allow Web access to DOCS Open. This is expected to be installed by June 1997. As a result, the number of users and the diversity of documentation in the database is expected to increase.

The largest document repository at the Census Bureau is maintained on the Census Bureau World Wide Web server.

Almost all of these documents are in static html, pdf, or ascii format that are available from the Census Bureau Home Page. There are over 35,000 documents available from the Census Web site. They can be accessed through hierarchical search guides, such as a Subjects A-Z lists, or through an indexed keyword search. Another document repository available from the Census site is CenStats. CenStats contains all publications released by the Census Bureau since January 1, 1996, currently there are over 3,500 of these publications available in PDF format.

Most documents available through the Census web site are considered public domain. Documents are stored in directories that represent each subject or category available at the Census Bureau. They are added to the server by authorized Data Disseminators. They must meet internal requirements for appearance and html code elements. Currently, there are documents representing almost all subjects and categories at the Census Bureau.

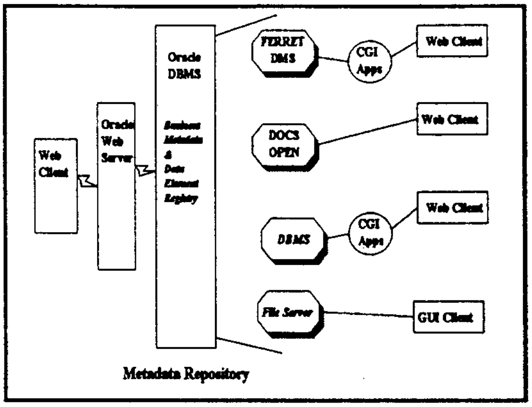

Figure 1. Metadata Repository Architecture (Physical Representation)

The Census Bureau web site is not only a service to Census Bureau data users, but it is intended to be a service for Census Bureau employees to display their documents to the public. Users are generally very satisfied with the web site's accessibility, timeliness of updates, and general "look and feel". It has won several Internet site awards including the 1997 Lycos Award for the top 5% of all web sites, Vice-President Gore's Hammer Award in 1994, Iway Magazine's top 500 sites in 1996, and a Starting Point Choice award in 1996.

3 The Need for a Central Document Repository

The Census Bureau realizes the value of improving efficiency through information technology. It also realizes the value of electronically storing and accessing the information capable of linking the documents and descriptions of Census Bureau data. Work is underway in the development of applications for a logically central metadata repository at the Census Bureau (Gillman, Appel, LaPlant, Lestina, 1996). A logically central metadata repository contains the metadata for survey designs, processing, analysis, and datasets. It creates links to the actual data files, therefore bridging the gap between the data and the users who wish to find them. It provides a central reference for locating information that is currently available on physically separate computer systems at the Census Bureau.

Two relational models and a prototype demonstrating the functionality of the metadata repository have been developed (Gillman, Appel, 1997) (Gillman, Appel, LaPlant, 1997). These models define the structure for storing statistical metadata. For example, there are more than 60 entities representing the business functions of the Census Bureau. One of these entities is called Documentation. It is the structure where metadata for documents are stored. The metadata will provide the necessary information to allow the electronic access to documents that are located on separate servers. Figure 1 shows the proposed physical architecture of how the metadata repository interoperates with distributed systems at the Census Bureau.

4 Implementation Issues

According to the Gartner Group (Popkin, 1995), the design and planning of a repository accounts for about 65% of the initial costs. This includes the classification or indexing of document metadata in the business data model and the development of applications that interoperate with document repositories. The other 35% of the initial costs are derived from the implementation of the repository. Implementation issues include preparing documents and registering them into a central repository, the access of documents from a central repository, administration and version control, and security.

The Gartner Group also estimates that in order for a documentation repository to be effective in an organization, at least 75% of the useful documents must be accessible through the repository. As a result, there are many issues concerning the registering and indexing of documents. For example, there may be thousands of documents to consider when starting implementation. Documents come in many different formats (e.g., html, WordPerfect, MS Word, GIF, ASCII text, paper) and there may be no standard mechanism for registering (the process of submitting document metadata to the central repository) or indexing them into the repository. To begin solving these issues, documents can be first classified simply as either electronic or paper. These types of documents can be classified further into categories such as "required" and "not required." Required documents are those that are necessary to perform the daily business functions of the organization. The "not required" documents are documents that support the "required" documents and are not necessary to perform a daily business function. Generally, a document with any "required" element is considered a "required" document. By classifying documents in this way, users and implementers may find it easier to organize their workload.

Once documents are sorted, they are eligible for registration. Such a process involves entering identification information about each document into a template, and then submitting the entries to the central repository. This process can be very time-consuming, so in addition to categorizing documents, it may be necessary to enter only two or three critical items such as title, URL, and contact name. Documents can also be registered as they are used so that less time is used retrieving or searching for a document. Unfortunately, there is a trade off. The less information that is registered about a document, the less accurate or less useful the metadata. Therefore, management needs to determine what indexing attributes are necessary when registering documents. The Census Bureau is currently researching ways of registering documents and is building Web-based tools for registering documents.

Another implementation issue for document repositories is how users access documents. The design of an access or search mechanism is critical for the usefulness of the document repository. There are at least three factors that control the effectiveness of search mechanisms: 1) the accuracy and completeness of indexing document metadata; 2) the utility of the browser interface to the repository; and 3) the effectiveness of a search engine for indexing and retrieving static documents. The accuracy of the search results depends on the accuracy and completeness of attributes in the repository. The browser interface is a combination of a template or Web-based CGI forms that represent the way a user perceives the functions of the organization. For example, a user perceives the Census Bureau as producing data for surveys and censuses. A Web-based browser interface would allow the user to select "Surveys" from the form and provide more detail with the subsequent linking page. Each page presents a logical progression of the organization. The main or home page of the browser interface could have a keyword search utility to locate documents indexed through the search engine. The Census Bureau is currently researching these techniques and is experimenting with search engines, search guide techniques, and browser interfaces for its central repository.

The security or access privilege of documents is another important issue. The implementers and designers need to decide not only who has privileges for accessing documents and what documents are accessible, but also account for documents that change security during their existence in the repository. The Census Bureau home page allows public access to the documents stored on the Census Bureau world wide web server. Otherwise, most documents are protected from external users by a firewall. The FERRET DMS and DOCS Open have more sophisticated security access. Security specifications for users and documents are stored in the database. Users are classified by security groups. They are identified as they access the system and are allowed access to documents and directories assigned to the security group. DOCS Open provides security attributes for viewing, editing, copying, and deleting documents.

Versioning is another critical issue. Frequently documents need to be updated or deleted from the repository. There needs to be an attribute in the document repository to allow for the document version number as well as an attribute representing a deleted document. For example, a document may be registered that is an update to an existing document. The most current document would be added to the repository. However, the previous and current document would be distinguished as two versions of the same document.

After the initial registration of documents, there must be a "registration authority" or someone responsible for making sure attributes in the repository are properly completed and specified and monitoring document metadata registration and deletion. A database administrator is responsible for maintaining user and document security privileges, adding and removing users, updating repository schema, and providing recommendations for application interoperability with the repository.

5 The Role of Distributed Objects

As we approach the end of 1997, information technology has progressed so that the currently used architectures for metadata repositories are almost obsolete. In 1992 there was a two-tiered architecture for accessing a server across a network. For example, a user issued commands from a GUI (Graphical User Interface) client (tier one) to a file server located across a network (tier two). In 1994, client/server architecture became more sophisticated by the addition of the World Wide Web. A web client (tier one) would submit requests to a CGI application located on an HTTP server (tier two). The CGI application would then submit SQL queries to a relational database(tier three) and return the results to the client. Repositories at the Census Bureau currently use this architecture.

As system architectures evolve, so does their performance. The original two-tier architecture shows weaknesses in the areas of scalability, performance, and management of system resources (Chen, 1997). The current three-tier model is now becoming outdated. The revised three-tier model introduces a business objects layer, that is, a logical layer of business objects or reusable code that represents the business rules of the organization and manages transactions (see Figure 2). For example, a user would submit a request to a business object from a Web client. The business object executes the business process, for example, accessing or registering a document. Objects are pieces of reusable code that can run on different platforms, move across networks, work with other objects, and are not owned by any one application or network. If a user needed to search for a document, he or she would get a Document Access Object instead of writing and testing a CGI application. For registering documents, the user would access the Document Registration Object. Business objects communicate and interoperate with one another by means of an Object Request Broker (ORB). Therefore, a Documentation Registration object would need to communicate with an ORB to know what machines and other objects to interrelate with in order to register a document.

The Common Object Request Broker Architecture (CORBA) is the leading standard definition on how objects move and interrelate in a 3-tier environment (see Figure 3). CORBA is the work of the Object Management Group (OMG) consortium which represents over 650 companies in the information technology industry. At the end of 1996, most of the major systems vendors released products implementing the CORBA 2.0 standard. By the end of 1997, it is expected CORBA ORBs will have major significance in the industry (Orfali, Harkey, Edwards, 1996).

The Java programming language is envisioned to be the start to managing distributed objects on the World Wide Web. It is the object-oriented answer for Web interoperability and portability. Java applets can be run on any hardware platform, can be moved to any network or operating system, and are designed to be executed and display results through a Web browser. Because of the object-oriented nature of Java, it may be the ideal tool for interacting with the CORBA standard (see Figure 4). Using this model, Java applets would communicate with business objects through the CORBA ORB. Such a model may represent the future of Web-enabled database access in the next one to two years.

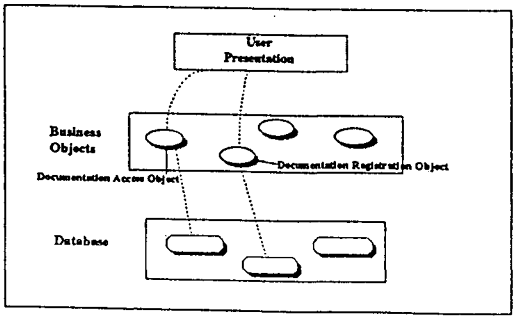
User Presentation
Business Objects
Documentation Registration Object
Documentation Access Object
Database

Figure 2. Three-Tier Architecture With Business Objects Layer (Logical Representation)

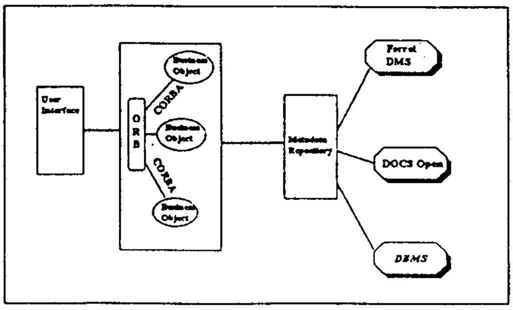

Figure 3. CORBA ORB Interacting With Metadata Repository

6 Proposed Direction for the Census Bureau

Work towards making documentation accessible through a metadata repository at the Census Bureau has just begun. The Census Bureau is in the process of developing Web based applications to register, retrieve, and update metadata in the repository. It is also planning to develop metadata administration processes that help users classify metadata objects, discriminate between similar metadata concepts, and ensure objects are registered completely. To accomplish these goals, the Census Bureau needs to create an informal committee to establish guidelines for registering and retrieving documentation metadata, using section 4 of this paper as a guide. The committee would need to make recommendations about goals for preparing and registering document metadata from various subject matter areas, goals for accessing documents and document metadata through search guides and search engines, and goals for administering security, version control, and metadata registration.

The committee would need to contain no more than seven members who have diverse levels of experience. There needs to be members who are technically familiar with the database systems at the Census Bureau and how they can interact with the Web. Section 5 of this paper provides a guideline for developing applications using the latest developments in technology. As mentioned in Section 5, these new developments are easier to maintain and are capable of running on different computer platforms. There also needs to be members who are familiar with users' requirements for registering, searching, and administering metadata through the repository. These members would help define how registration templates and search mechanisms would look to the user. All members would have an understanding of the vision for metadata access and retrieval at the Census Bureau.

In addition to an informal committee deciding on implementation issues, a separate team of technical experts needs to be available to prepare simple examples or prototypes of templates, search guides, and search engines to illustrate the committee's findings. These examples would demonstrate the actual connectivity and browser screens users would encounter and would be prepared as they are discussed in the committee. From these examples or prototypes, technical experts can refine screen designs incrementally with user's input. Such an approach has proven useful in other system development projects (Gibbs, 1997)(Edwards, Harkey, Orfali, 1996) and would be most effective in developing new systems at the Census Bureau.

The committee would report to an information technology team leader who would set deadlines for decisions and prototype requirements. The information technology team leader would ultimately be responsible for setting the vision for document metadata access and registration at the Census Bureau.

7 Conclusion

The efficient access and storage of documentation metadata through a metadata repository is critical for improving business operations at the Census Bureau. A metadata repository provides the capability for classifying documents, so that similar documents from different surveys and programs can be located from a single source. A metadata repository also provides links between variables and documents as well as products and documents.

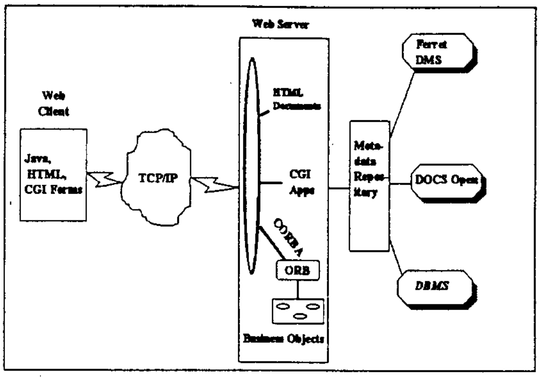

Figure 4. Using Java Applets to Access Legacy Databases Through the Web

The following are examples of why the storage and access of documentation metadata is so important: A researcher or computer programmer needs to obtain the latest specifications or needs further information to continue the development of his or her project. An electronic document metadata repository would provide an electronic search mechanism for locating documents that he or she probably never knew existed or never knew where to find. As a result, the number of days to complete a survey design or planning project may be reduced considerably.

A manager working on census data collection needs to know what areas have been collected and what areas still need follow up. He or she may also need the procedures for handling these particular areas or may need to know what personnel are best qualified. A document metadata repository allows the manager to locate a number of documents available pertaining to "data collection followup" or "personnel qualifications" from a single source. Previously, a manager may have had to contact several sources to get the information he or she needed.

Implementing the registration and retrieval of document metadata through a metadata repository can imply organizational and strategic changes. The ideas presented in this paper require a conversion of work process from a simple clerical storage of documents to an electronic method of access and registration of document metadata. However, it is to the Census Bureau's advantage to anticipate changes and then react to these changes, rather than become a victim of obsolescence. It is critical then that the Census Bureau invest the resources in improving the efficiency of document management. Efficient document management will enable lower costs throughout the organization.

8 References

Anderson, M., Austin, T., Bair, J., Baylock, J., Brown, D., Casonato, R., Comport J., Dresner, H., Kirwin, W., Popkin, J., Scherberger, K., Whitten, D., Lett, B., "The Electronic Workplace: Fulfilling the Promise", Gartner Group, May 30 1996.

Chen, M., Ph.D., Developing Client/Server Applications, Print Services, George Mason University, 1997, pp 1-10 3-Tier Architecture.

Edwards, J., Harkey, D., Orfali, R., The Essential Client/Server Survival Guide, Second Edition, John Wiley & Sons, Inc., 1996

Gibbs, W. Wayt, "Taking Computers To Task," Scientific American, July 1997, pp.82-9.

Gillman, D. W., Appel, M.V. (1997), "The Statistical Metadata Repository: An Electronic Catalog of Survey Descriptions at the U.S. Census Bureau". Presented at International Association for Social Science Information Service and Technology (IASSIST) , May 1997, Odensk, Denmark.

Gillman, D.W., Appel, M.V., LaPlant, W.P. (1996), "Metadata Management at the U.S. Bureau of the Census: A Standards Based Approach". Presented at the 9th Annual DAMA International Symposium, the Metadata Conference, March 1997, Dallas, Texas.

Lestina, G.J., Appel, M.V., Gillman, D.W., LaPlant, W.P. (1996), "Technical Development of the Proposed Statistical Metadata Standard". Presented at the American Statistical Association Joint Statistical Meetings, August 1996, Chicago, Illinois.

Malhotra, Yogesh, "National Information Infrastructure: Myths, Metaphors And Realities", @BRINT, HTTP://www.brint.com/papers/nii/issues.htm, 1995 "No Alternative to the Census," American Demographics, November 1995

Popkin, J., "Implementing an IDM System: Tactical Decisions", Gartner Group, August 7 1995

Williamson, Mickey, "Roundtable: Government Reform. Rethinking the Way Government Works", CIO, July 1996