STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working Paper No. 11
English only

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Metadata
(Washington, D.C., United States, 28-30 November 2000)

Topic (ii): Metadta modelling and terminology issues

USE OF METADATA FOR THE EFFECTIVE INTEGRATION OF DATA FROM MULTIPLE SOURCES

Submitted by United States Census Bureau [1]

Invited paper

I. ABSTRACT

1. As survey practitioners in statistical agencies continue to expand the collection and dissemination of data electronically, the statistical community has embraced the notion that using data integrated from multiple sources is more powerful than relying on stand-alone data sets. What has been termed integrated statistical solutions (ISS) consists of providing data users with answers to their questions, without the user first having to know the structure of the government or agency or how the data files are organized or formatted. Given the trends in technology and user expectations, the movement toward producing such integrated data applications is certainly inevitable. As a result, the role of metadata to support applications using integrated statistical information has become increasingly important. This paper compares and contrasts alternative metadata-sharing support structures that statistical agencies might employ to enable the integration of data and metadata from multiple sources. It concludes with a brief status update on the Corporate Metadata Repository currently under construction at the U.S Bureau of the Census.

Key Words: Repositories, Registries, Metadata-sharing, FedStats, ISS

II. INTRODUCTION

2. As survey practitioners in statistical agencies continue to expand the collection and dissemination of data electronically, the role of metadata as a driver and in support of understanding the content of our statistical information is becoming increasingly important. We have entered this new century amidst a wave of technological innovation, featuring the Internet and the World Wide Web as the primary means of information dissemination. At the same time, the international statistical community is being challenged with an urgent and critical need for more accurate, timely, relevant, accessible, and interpretable data. Several thought leaders in various statistical agencies around the world are attempting to meet these user needs by working together, like never before, to implement a modernized, customer-driven, cross-program, and cross-agency integrated data access and dissemination service capability.

3. Over the past few years, the statistical community has embraced the notion that using data integrated from multiple sources is more powerful than relying on stand-alone data sets. What has been

[1] Prepared by Mark E. Wallace and Samuel N. Highsmith. This paper reports the general results of research undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform parties of research and to encourage discussion.

termed integrated statistical solutions (ISS) consists of providing data users with answers to their questions, without the user first having to know the structure of the government or agency or how data files are organized or formatted. And, given the trends in technology and user expectations, the movement toward producing integrated data applications is certainly inevitable. However, the question now before us is "How will we realize the vision?" The answer lies in the development of the statistical and spatial metadata infrastructure support to drive the integration of data and metadata from multiple agencies.

III. BACKGROUND

4. Currently there is no assurance that similarly named items will be referring to the same thing. Conversely, there are likely to be data elements with different names in different databases that are actually descriptive of the same things. Further, there is currently a decided lack of rules for identifying which statistics are "integrateable" for various levels of geography, time, and topic. Hence, it is evident that for effective comparison of the meanings and intelligent use of information items in various statistical databases, metadata naming and attributes systems need to be based on international standards, such as the multi-part ISO/IEC 11179, "Specification and Standards of Data Elements".

5. A number of federal agencies have developed or are developing their own metadata registries and are using these systems for operational needs. These include the Environmental Protection Agency, the Bureau of the Census, the Department of Transportation, the Department of Energy, the Bureau of Labor Statistics, the Health Care Financing Administration, the Department of Defense – Health Affairs, and the Department of Veterans Affairs. Operationalizing metadata repositories in these agencies is certainly a good first step. Next, however, to support ISS, agencies must collaborate to develop methods that support the automated sharing of metadata between agencies that have metadata registries.

6. This paper describes some important initial steps that agencies and researchers might take to initiate the development of these metadata-sharing methods. These steps include efforts by a team of several federal agencies, working with partners in the data user and research communities, to develop models that demonstrate the value of, as well as the technical capabilities for, dynamically integrating data from multiple sources.

7. Specifically, over the next several months, this work will involve experimenting with various combinations of hardware, software, data and metadata from the various partners collaborating on this project. There is the potential for determining and validating best approaches for sharing data and metadata from multiple sources across agencies. We hope that lessons we learn from this work will help foster the ushering in of a data and metadata-sharing environment from which a number of new cross-agency topical and geographic-related data access and dissemination tools and products can be introduced.

IV. FIRST STEPS

8. Documenting data sets is a first step in enabling data sharing between agencies. Statistical and geographic databases have been built thus far to support the mandates of single institutions or parts of an institution. All who collect and manage data for activities related to their own responsibilities will need to understand and appreciate the value of those data to others and to collect and structure their data accordingly. To take full advantage of the opportunities offered by new technologies, business, government, and academia will need to develop, support, and fund metadata on a systematic and ongoing basis as well as promote access for all.

9. The Internet has clearly changed expectations and heightened knowledge about the ease of access to information as well as broadened the universe of users. Customers, both internal and external, associated with the various government agencies expect technology to provide easy and rapid access to documented, usable, and interpretable data. Developing metadata repositories based on international

standards will enable us to address the increasing demand for rapid access to documented and usable data. This might eventually be expandable to a virtual government-wide, or even global, level. Metadata would be the integrating mechanism for providing government-wide information.

10. Presently, at the Census Bureau, business processes touch metadata from the survey/census design through the dissemination life cycle (Gillman and Appel 1999). Yet, metadata is not formally cataloged until the dissemination phase where it is required for Bureau dissemination systems and products. We are now beginning to see the advantage in cataloging these data at each phase of the survey life cycle so that metadata doesn't have to be recreated during the dissemination phase. This approach also provides value added during each of the core business processes.

V. INTERAGENCY COLLABORATIVE PROJECTS

11. Various teams composed of staff from all directorates of the Census Bureau in collaboration with the State Data Centers, FedStats agencies, and the research community are researching the potential and validating data integration processes for accessing and integrating both micro and macro level datasets (and their metadata) that are Census Bureau-based and/or remote. These teams also are developing data integration tools to create profiles (predefined and user-specified), and to give customers new data integration functionality, providing information based solutions not currently available in our existing data access tools. It will be through metadata that much of this work will be accomplished.

12. These projects will make use of a collaborative, multi-agency laboratory server environment that is being established to evaluate different tools and approaches, use corporate metadata policies and techniques, and include profiles with some level of graphic display capabilities.

VI. COMPARISON OF VARIOUS METADATA MODEL APPROACHES FOR INTEGRATING DATA FROM MULTIPLE SOURCES

13. In the course of developing our own Corporate Metadata Repository (CMR) at Census, and in working with other organizations to experiment with ways to share and integrate disparate data sets, we have examined the feasibility of implementing three possible metadata models. Our goal is to determine which one(s) might best, over time, bring about the desired outcome of effectively and intelligently integrating data from multiple sources. To this end, we have compared and contrasted a number of factors with regard to each approach.

14. The initial evaluation was performed by a cadre of Census staff from various areas. This group developed the factors used to evaluate the three models and ranked each model numerically. Once their initial evaluation was complete, the metadata model approaches and their preliminary evaluations were shared with representatives of other agencies which are doing work in this area. Their feedback is incorporated in the evaluations below.

15. The factors we examined are performance, scalability, cost, interoperability, control, flexibility, security, short and long-term implementability, maintenance, and reliability. For each metadata model, we assigned two values (on a scale of 1 to 10 with 1 being the worst and 10 being the best) to each factor. The best possible score for the 10 factors examined would be 100. These two values are based on the experiences of a number of agencies now and what we project will be happening in 5 years.

VI.1 Single Source Repository for Metadata (SSRM)

16. This model stores all the agreed upon metadata elements for all participating agencies in one central repository. All applications would use it directly for access and maintenance. This would require all participating agencies to agree on a core set of metadata elements. It is the easiest model to build from the standpoints of interoperability, short-term implementability, control, security, maintenance, and reliability. The difficulties in this model occur in the areas of performance, scalability, cost, flexibility,

and long-term implementability. In fact, this approach is the very methodology used in many of our existing stovepipe systems. This would require construction and continuing support of one central repository to be used by all participating agencies. This approach would not support the concept of unique or different metadata element requirements. All agencies would have to fit the metadata for their various data sets into one standard definition. And, as more agencies participate over time, more resources to support this SSRM would be required by the hosting agency.

- Performance – As usage by multiple organizations grows, reasonable performance of one large application becomes increasingly difficult to provide.
- Scalability – This system has limited scalability due to the increasing cost of providing reasonable performance with a central resource.
- Cost – The initial cost is very reasonable. However, this model becomes very expensive as additional usage and more requirements are placed on it.
- Interoperability – This is a very interoperable system with a single applications program interface for application interface.
- Control – This implementation is the easiest to control.
- Flexibility – Mediocre flexibility due to difficulty in changing metadata elements. Since every organization using this system must provide a strictly agreed upon set of metadata elements, adding new elements or changing existing elements is very difficult.
- Security – This implementation, by being centralized, is the most secure.
- Short and Long-term Implementability – For the short term, this is the easiest model to implement. Over the long term, this approach is less viable since implementability is likely to decrease in proportion to the number of participating agencies.
- Maintenance – Very maintainable.
- Reliability – With fewer points of failure, this model should provide the highest reliability.

VI.2 Distributed Dissemination Metadata Repository (DDMR)

17. This model entails each agency providing an agreed upon core set of metadata elements, but they would not necessarily conform to one international metadata standard. Instead, metadata sharing among the agencies would be accomplished via the use of a standardized software supported metadata interchange. In this model, each organization would build its own metadata repository using the agreed upon metadata elements and underlying model. The metadata elements would be limited to those required by data dissemination applications, meaning that documentation of the survey process not needed for dissemination would not be provided. Metadata added to one repository would normally be replicated to all other repositories using a "token" registration scheme to ensure all registration efforts succeed. Organizations would be able to be selective on which other organizations their metadata would be sent to. They could also choose not to share some or all metadata. Each organization would be responsible for providing their own metadata registry integrity, to include registry backup operations. The end result would be the creation of many unique metadata registries using a common registry format and shared software tools. Since we would need a complete set of shared functionality for this approach, it assumes a common architecture used by all. To implement this model, some organization would have to develop, distribute, and support the application software

- Performance – Since each organization will support and access its own metadata registry, performance is the responsibility of each organization. Access to distributed data residing in other organizations is directly related to the speed of the internet connection deployed by each organization.
- Scalability – This system has somewhat limited scalability due to the technique of distributed data and multiple unique metadata registries. As various agencies share increasing amounts of metadata, the size of the DDMR will increase exponentially.

- Cost – The cost is reasonable, especially since each organization shoulders the cost of its own metadata registry. The only really expensive part of this implementation is the construction of metadata and support of a shared set of software to allow distributing metadata across registries.
- Interoperability – This is a very interoperable system when the shared software is developed and distributed. The unfortunate side effect of this approach is that it could become obsolete when metadata interchange standards come into existence unless the software development effort is tied to the standards development effort.
- Control – Each organization can establish and control its own security, which means control is only as good as each organization's implementation.
- Flexibility – Good flexibility in changing metadata elements.
- Security – This implementation is only as secure as the participating organizations make it.
- Short and Long-term Implementability – In the short term, this system requires fairly sophisticated software development and support by a lead organization willing to take on responsibility for the software, and implementation and support of the model. For the long term, this system will become increasingly difficult to support as the number of participating organizations increases. In addition, the lead organization will be burdened with the responsibility of updating and distributing application software and tools over time.
- Maintenance – Very much dependent on each organization deploying it.
- Reliability – Very much dependent on each organization deploying it.

VI.3 Federation of Unique but Related Metadata Repositories (FMR)

18. This model is based on the concept of a logically central metadata repository (Sundgren, et al 1996). Each agency would maintain their own registry, including the agreed upon set of metadata elements, and, when it becomes available, each agency would conform to one international standard. Recognizing that needs vary across agencies, agencies could, in effect, build their own metadata mart as long as their registries were in compliance with the standard. This way, agencies would be able to extend metadata requirements beyond the core set, add application specific metadata or tune their implementation to meet their application performance requirements.

19. This is the basis for the model currently under construction at the U.S. Census Bureau (Gillman and Appel 1999). The core component is a centrally built and maintained metadata repository like the SSRM. This metadata repository and its supporting tools can be used directly by participating organizations, should they so choose, using the agreed upon metadata elements. However, to support unique organizational requirements, a second technique is for departments to build and support their own metadata repository. In this more loosely coupled setup, the organization can copy the central metadata repository and make any additions required by their applications.

20. Another technique, particularly useful where an organization has already put in place their own metadata repository, is to map the agreed upon metadata elements of the central repository to the components of the FMR and build an interchange. This particular flavor of the FMR approach envisions development and use of a standard metadata interchange format to exchange metadata between repositories. Although this international metadata standard does not yet exist, it appears likely that such a standard based on XML is likely to emerge in the next few years.

- Performance – Performance of the actual central system is relatively easy to provide. Performance of distributed repositories is very much under the control of the separate organization building and supporting their own registry.
- Scalability – This system has virtually unlimited scalability due to the technique of distributed data and multiple unique metadata registries.
- Cost – The cost is reasonable, especially since each organization shoulders the cost of its own metadata registry. The only really expensive part of this implementation is the construction and support of a shared set of software to allow distributing metadata across metadata registries.

- Interoperability – This is a very interoperable system when the shared software is developed and distributed. The unfortunate side effect of this approach is that it could become obsolete when metadata interchange standards come into existence unless the software development effort is tied to the standards development effort.
- Control – Each organization can establish and control its own security, which means control is only as good as each organization's implementation.
- Flexibility – This is the most flexible model to implement.
- Security – This implementation is only as secure as the participating organizations make it.
- Short and Long-term Implementability – For the short term, this is a very easy model to implement. For the long term, this approach continues to be viable.
- Maintenance – Very much dependent on each organization deploying it.
- Reliability – Very much dependent on each organization deploying it.

VII. SUMMARY OF SCORES FOR EACH METADATA MODEL

21. Based on the above evaluations, none of the three models demonstrated overwhelming superiority over the others. Over the long term however, the FMR – which is a hybrid approach – appears to be the best. Below are summary evaluations of each model along with numerical scores.

VII.1 Single Source Repository for Metadata (SSRM)

22. The SSRM represents the most centralized approach to metadata sharing. Because all of the dissemination metadata would be in one repository, interoperability and control are optimized. Maintenance and reliability are also highly rated since all of the shared metadata would be maintained together, and would not be subject to the unique requirements of the individual participating data and metadata suppliers. With a small group of participating organizations, this model should be very implementable. As the system grows over time however, and the number of users increases, maintaining a reasonable level of performance will become expensive and difficult to manage.

23. Also, a centralized approach requires data and metadata suppliers to adjust their metadata standards – at least for the core elements – to fit within the requirements of the single source repository. This limits flexibility and makes implementability less feasible over the long term unless an international standard were adopted and adhered to.

VII.2 Distributed Dissemination Metadata Repository (DDMR)

24. The DDMR model provides the least centralized approach to metadata sharing. Since each participating organization would be responsible for establishing and maintaining its own metadata registry, this model is considered to be infinitely scaleable. Costs associated with developing a virtual web of metadata registries would also be borne by individual organizations which is considered to be a strength. However, control, security, maintenance, and reliability of component registries would remain under the purview of the organization sponsoring the registry, which could become an excessive burden.

25. In addition, the sponsoring organization will need to take responsibility for developing and maintaining the underlying architecture for the distributed environment and fairly sophisticated data access and integration tools which would be common to all registries. Over the long term, this would also become burdensome, and in fact may well become impossible as more and more agencies use increasingly larger replicated repositories.

VII.3 Federation of Unique but Related Metadata Repositories (FMR)

26. The FMR represents a hybrid approach which proposes to take the best functionality from both the centralized and the distributed models, and implement a very flexible model based on international

standards for metadata registries and interchange format. As a result, ratings for most evaluation factors are equal to or better than the scores of the other models. Notable exceptions are, interoperability, control, maintenance and reliability, where the centralized model is strongest. This is due to the fact that metadata registries are developed and maintained individually by each sponsoring organization notwithstanding the logically central repository to which they all contribute. The expectation that there will be variability in these areas is considered a weakness that may be expected to diminish somewhat over time. For this reason, the FMR, while scoring highest of the three approaches over the long term, is not significantly superior especially in the short term. Nevertheless, the FMR model is probably the best choice because of its long term superiority.

VII.4 Numerical Scores

27. As part of our analysis, the three models were subjected to numerical ratings for each of the factors. The highest achievable score was 100. Below are the scoring results by evaluation factor – both short term and long term.

Evaluation Factors	**Score**					
	SSRM		**DDMR**		**FMR**	
	Now	*In 5 yrs.*	*Now*	*In 5 yrs.*	*Now*	*In 5 yrs.*
Performance	4	2	5	5	4	6
Scalability	4	2	6	3	10	10
Cost	7	3	7	7	7	7
Interoperability	10	10	8	4	8	6
Control	10	10	5	5	6	7
Security	7	7	5	5	6	7
Implementability	10	7	6	2	9	9
Flexibility	4	3	7	7	10	10
Maintenance	9	9	5	5	5	7
Reliability	9	9	5	5	5	7
Total Score	74	63	59	48	70	76

VIII. CORPORATE METADATA RESPOSITORY AT THE U.S. CENSUS BUREAU OF THE CENSUS

28. Based on a business process model for survey processing, the BOC began construction of a Corporate Metadata Repository in 1998. This business process model was used to build a statistical metadata repository using an Oracle database. We progressed from building prototypes as a proof of concept, to building pilot applications with BOC stakeholders, to building a production system which will soon be available for BOC users. The history of the project is described in a contributed paper, "Building a Corporate Metadata Repository at the U.S. Census Bureau", presented in May of 2000 at the ISIS 2000 seminar in Riga, Latvia. The components of the current system include:

An ISO/IEC compliant Data Element Registry
A Dataset Registry Component
A Product Registry for Tier 1 Census Products
A Security Component
A published API for metadata input/output
Areas to store Survey and Census design elements
Support for data collection instruments and the links to collected data elements, data sets
Support for post data collection survey and census processing

Support for data dissemination
A CMR tool set which includes
CMR administration, Browsing and reporting capabilities.
CMR interchange which includes an ASCII delimited interchange, XML-based open interchange, upload and download utilities
The ability to output the metadata in a number of formats (XML, Delimited ASCII, FGDC)
Custom tools including BOC developed tools and applications
A metadata driven application to check data quality
An end user maintained basic document management system to organize and maintain unstructured metadata, which is basic Wordperfect, Word, PDF, Ascii text, or other untagged text files.

29. This will be an Inter/Intranet web based system for which the user will need only a standard web browser. The system is designed to allow adding commercial solutions to replace our programmed solutions when they become available. Our applications will interface to an object layer, beneath which is a relational database. One of the most important concepts is the usage of a model to generate both the database and applications. This allows us to add new tables and elements to the model as new requirements become evident with much less development time for re-programming.

30. The Data Element Registry is due for final release in November; the most current version is being demonstrated on day one of this conference by our Oracle consulting team. We also plan to install the Data Element Registry on an external Census machine for examination a possible use by other organizations. This Data Element Registry is currently being implemented at The Federal Aviation Association and is being looked at by other organizations.

31. The final data collection, survey processing, and data dissemination support parts of the system are scheduled for completion late in the first quarter of 2001.

IX. NEXT STEPS

32. Based on an examination of important factors with regard to the various metadata models for sharing data and metadata from multiple sources, it appears that a most viable approach may be to adopt the Federation of Unique but Related Metadata Repositories (FMR) model. We think that this system offers the most flexibility not only now but well into the future. When an international metadata interchange format becomes available, this system will be well positioned to use and take advantage of it.

33. Some of the questions we hope to be able to answer over the next few years are:

i) International standards for metadata registries will exist soon. What will be the impact of new technologies including XML?
ii) How well will we be able to adopt standards within individual organizations, and develop/maintain the necessary vision to support data access and integration capability within and across agencies?
iii) Will we receive the support necessary to continue collaborative efforts among federal agencies, academia and the research community?
iv) How accurate are our predictions regarding the choice of an appropriate metadata sharing model to develop? We are currently researching the feasibility of developing the DDMR and FMR models. If research show that we are incorrect in our assumptions concerning the scalability and implementability of the DDMR, it could prove to be a viable model.

34. We will certainly learn more about metadata sharing efforts such as the FedStats Product Concepts Working Group as we continue to collaborate with various parties in the public and private sectors and in the academic and research communities.

REFERENCES

Capps, Cavan P., Green, Ann, and Wallace, Mark E. (1999), "The Vision of Integrated Access to Statistics: the Data Web", paper submitted to the Association of Public Data Users.

Gillman, Daniel W. and Appel, Martin V. (1999), "Statistical Metadata Research at the Census Bureau", Proceedings of the Federal Committee on Statistical Methodology Research Conference, pp.1-10.

ISO/IEC 11179 (1994-2000), "Information Technology – Specification and Standardization of Data Elements", Parts 1-6, Draft International Standards.

Schneider, Daniel (1999), "Information and Database System Interoperability with Assured Semantic Consistency: The Role of Data Semantics Management Systems – A White Paper for Federal Agency CIOs and IT Architects" (Draft).

Sundgren, B., Gillman, D.W., Appel, M. V., and LaPlant, W. P. (1996), "Towards a Unified Data and Metadata System at the Census Bureau", Proceedings of the Census Annual Research Conference.

Wallace, Mark E., Landman, Cheryl M., Sperling, Jon, and Buczinski, Carla (1999), "Integrated Information Solutions - The Future of Census Bureau Data Access and Dissemination", Proceedings of the Government Statistics, Social Statistics, Survey Research Methods Section, American Statistical Association.

Wallace, Mark E. (2000), "User Driven Integrated Information Solutions - Digital Government by the People for the People", Topic iv: Improving data dissemination strategies, Seminar on Integrated Statistical Information Systems (ISIS 2000).

Wallace, Mark E., Highsmith, Samuel N (2000), "",