
DATA DNA: THE NEXT GENERATION OF STATISTICAL METADATA

Cynthia M. Taeuber, Daniel W. Gillman
Laura Smith

A Discussion Paper Prepared for the Brookings Institution Urban Markets Initiative

March 2007

URBAN MARKETS INITIATIVE SUMMARY OF PUBLICATIONS*

2007

Utility Payments as Alternative Credit Data: A Reality Check

2006

Give Credit Where Credit is Due: Increasing Access to Affordable Credit Using Alternative Data

Downtown Detroit in Focus: A Profile of Market Opportunity

Measuring the Informal Economy – One Neighborhood at a Time

Tools to Avoid Disclosing Information About Individuals in Public Use Microdata Files

Fulfilling the Promise: Seven Steps to Successful Community-Based Information Strategies

Neighborhood Housing Markets and the Memphis Model: Linking Information to Neighborhood Action in Memphis, Tennessee

The Affordability Index: A New Tool for Measuring the True Affordability of a Housing Choice

2005

Federal Statistics: Robust Information Tools for the Urban Investor

Market-Based Community Economic Development

Using Information Resources to Enhance Urban Markets

2004

Using Information to Drive Change: New Ways of Moving Markets

* Copies of these and Brookings metro program publications are available on the web site, www.brookings.edu/metro/umi.htm, or by calling the program at (202) 797-6131.

ACKNOWLEDGMENTS

The authors wish to thank Andrew Reamer, Katherine Wallman, Susan Schechter, Brian Harris-Kojetin, and Brand Niemann for their contributions to this work.

The Urban Markets Initiative thanks the Ford Foundation and the Rockefeller Foundation for their generous support. UMI also thanks its founding funder, Living Cities: The National Community Development Initiative. Living Cities is a partnership of leading foundations, financial institutions, nonprofit organizations, and the federal government committed to improving the vitality of cities and urban communities.

ABOUT THE AUTHORS

Cynthia M. Taeuber (cmtaeuber@comcast.net) is a former senior Census Bureau chief and also former associate at the Jacob France Center at University of Baltimore. Daniel W. Gillman (Gillman.Daniel@bls.gov) is a research mathematical statistician with the Office of Survey Methods Research at the federal Bureau of Labor Statistics. Laura Smith (lsmith@brookings.edu) is a senior research assistant with the Urban Markets Initiative at the Brookings Institution. For more studies by the Urban Markets Initiative, go to www.brookings.edu/metro/umi; the authors may be contacted directly with comments on this work.

ABSTRACT

Statistical metadata is commonly defined as data about data. Metadata documents information about a statistical dataset's background, purpose, content, collection, processing, quality, and related information that an analyst needs to find, understand, and manipulate statistical data. As such, the metadata for a statistical dataset broadens the number and diversity of people who can successfully use a data source once it is released. It is the purpose of this paper to discuss issues related to the development and use of statistical metadata and to describe resources to standardize and automate statistical metadata. While there are many types of metadata – this paper is concerned only with statistical metadata.

This paper describes components of a complete statistical metadata system as well as critical elements of basic information for a statistical metadata system. It also reviews the tools that are available now or that could reasonably be developed to create and structure metadata for better access and understanding of datasets by diverse users. Currently lacking in the field of data collection are incentives for shifting the creation of statistical metadata from a costly burden to a benefit; this paper addresses possible incentives and suggests ways to integrate metadata into existing and developing datasets. Finally, this paper describes implications of the tools and related cautions for the National Infrastructure for Community Statistics (NICS).

TABLE OF CONTENTS

I. INTRODUCTION

As shorthand, statistical metadata is commonly defined as data about data. Metadata related to a particular statistical dataset identify the data and describe its content and quality so that data users can retrieve, use, and process it appropriately.¹ Metadata describe information about a statistical dataset's background, purpose, content, collection, processing, quality, and related information that an analyst needs to find, understand, and manipulate statistical data. The information in a statistical metadata system is essentially a reference library about a dataset. As such, the metadata for a statistical dataset broadens the number and diversity of people who can successfully use a data source once it is released. This paper discusses issues related to the development and use of statistical metadata and describes resources to standardize and automate statistical metadata. While there are many types of metadata—this paper is concerned only with statistical metadata.

Private and public statistical organizations produce extensive statistical data from surveys, opinion polls, and administrative records. The information is distributed as data files, and, at least from federal statistical agencies, with documentation or metadata that provides information an analyst needs to use the statistical data appropriately. There are several issues. Formal metadata is not available at all for many datasets, especially administrative records. It is available for federal surveys but there are no generally agreed-upon conventions or formats amenable to automating the relationships within and among datasets.

Uniform ways to describe and manage diverse information are needed, and standardized metadata for every dataset in a system is essential for the interoperability of that system. Interoperability refers to “the ability of information and communication technology (ICT) systems and of the business processes they support, to exchange data and to enable the sharing of information and knowledge.”² Going further, if statistical metadata is standardized and then structured to be machine readable so that the content, context, and relationship among variables is linked, it becomes possible to efficiently assess information from diverse sources³ with ontology tools, that is, a formal and reusable library of terms, their definitions, and related concepts for machine processing, one of the types of web-based tools we discuss below.

The Conference of European Statisticians, under the aegis of the United Nations Statistical Commission and the Economic Commission for Europe proposed guidelines for statistical metadata including a distinction among three types of metadata:

- Type I - Assists search and navigation on a website (e.g., a search engine, topic links, a site map);

1 Management of Statistical Metadata at the OECD (v2.0,6/9/2006), see page 3: www.oecd.org/dataoecd/26/33/33869551.pdf

2 Brand Niemann, “Semantic Interoperability Community of Practice Enablement (SCOPE) for Enterprise Architects,” ArchitecturePlus Seminar: Semantic Interoperability, Ontology and Their Potential for Federal Information Sharing, January 18, 2005.

3 Ibid, slide 23.

- Type II - Information to describe statistical data, so data users can understand and evaluate the appropriateness of the data for its intended use and then analyze results; statistical metadata includes, for example, documentation of definitions, relationships among variables, specifications, procedures, classification schemes, and instructions; and
- Type III - Assists post-processing (e.g., downloading data; statistical tools for analysis).

The first type of metadata is well established and will not be considered further in this paper. There are, however, organizational issues with this type of metadata and improvements are needed so that data users can find data more easily. The GovStat Project is addressing these.⁴

The third type of metadata for processing data from a website has many well-established tools already. There is, however, a need for further development of automated statistical tools that are commonly used for data analysis by both experts and those who use data infrequently. Such tools include automatic calculators for frequently used statistics such as percentage change and adjustment for inflation. Users of survey data could be greatly helped with automatic computation of confidence intervals and statistical testing of the significance of comparisons. The complication is that easy-to-use tools carry a danger of being misused by those who are unfamiliar with the concepts and assumptions behind the computations.

The second type of metadata, information to describe statistical datasets, is the focus here.

This paper will:

- Describe components of a complete statistical metadata system as well as critical elements of basic information for a statistical metadata system (Appendix A lists critical elements separately);
- Review the tools that are available now or that could reasonably be developed to create and structure metadata for better access and understanding of datasets by the diverse users of the National Infrastructure for Community Statistics (NICS);
- Consider incentives for shifting the creation of statistical metadata from a costly burden to a benefit; and
- Describe implications of the tools and related cautions for NICS.
- Provide a glossary of technical terms (Appendix B).
- Provide an example of the types of questions for which an analyst uses statistical metadata (Appendix C).

⁴ Sheila Denn, Stephanie W. Hass, and Carol A. Hert, "Statistical Metadata Needs During Integration Tasks," 2003, see: http://www.siderean.com/dc2003/301_Paper50.pdf; and William Kules and Ben Shneiderman, "Designing a Metadata-Driven Visual Information Browser for Federal Statistics," Proceedings of the 2003 National Conference on Digital Government Research, pp. 117–122, <http://www.dgcr.org/dgo2003/>.

II. WHAT INFORMATION IS IDEAL AND WHAT IS MOST CRITICAL?

Statistical metadata, as we refer to it here, is a part of an information infrastructure that helps users decide whether a statistical dataset is appropriate to the question being addressed, and provides guides for users on how to locate and then manipulate and analyze the data with statistically valid methods.

The discussion below focuses not on the many ideas about appropriate architecture for managing statistical metadata, but rather, on the components of basic information that is essential to conduct community-level research properly.

As local and state governments consider release of administrative records for statistical purposes, we present the guide below so that data owners will know the types of documentation to preserve over time and what knowledge is fundamental and critical about every data set.

Statistical metadata helps researchers gauge the quality of the data and determine whether it is sufficiently reliable for their purposes. Denn, Haas, and Hert have studied the needs for statistical metadata as users integrate data from different sources. They found that of the integration tasks they observed, the most important uses of metadata were to note discrepancies and why there are differences among variables, to manipulate statistics, and especially to make comparisons across geography, time, concepts, index values, and among sources. Common user problems included difficulty in relating the technical terms agencies use to more familiar language, knowing definitions of variables and making comparisons, needing help to interpret the data, understanding the geography related to the dataset, and finding information about the currency of the statistics and when they are updated.⁵ David Stevens of the Jacob France Institute at the University of Baltimore also notes the dangers of forcing a fit of definitions and that statistics (especially administrative records) are not necessarily updated on a standard or announced schedule.

Even though federal statistical agencies commonly create and maintain historical metadata for their surveys and the statistical files they produce, how they do the documentation is not standardized across agencies. Recently, the U.S. Office of Management and Budget (OMB) conducted a review of Statistical Policy Directives Nos. 1 and 2 covering standards for statistical surveys and publication of statistics, and has issued for public comment proposed principles and guidelines for statistical surveys.⁶ This includes guidelines for documentation of survey information.

For the purposes of NICS, we can consider the OMB recommendations as an example of “best practice” for surveys. Another best practice example is the OECD’s basic metadata principles (all statistical data must be accompanied by metadata and the metadata must be consistent) and

⁵Denn, Haas, and Hert, *Ibid.*, p. 9.

⁶ U.S. Office of Management and Budget “Proposed Standards and Guidelines for Statistical Surveys,” http://www.whitehouse.gov/omb/infoereg/proposed_standards_for_statistical_surveys.pdf. See also the U.S. Office of Management and Budget’s Federal Register Notice and current Statistical Policy Directive No. 1, Standards for Statistical Surveys and Statistical Policy Directive No. 2, Publication of Statistics at: <http://www.whitehouse.gov/omb/infoereg/statpolicy.html#pr>. See especially Section 7.3 for guidelines.

their guidelines for a Metadata Common Vocabulary (MCV). OECD refers to “reference metadata” that describes the content and quality of the statistical data and it includes conceptual metadata, methodological metadata, and quality metadata.⁷ Standardization of the type of information provided and how it is documented is the first step towards the objective of automatically creating metadata and frequent information updates.

OMB’s proposed principles and guidelines apply to surveys better than to administrative records. Federal statistical agencies routinely document surveys, but document federal administrative records with less regularity. It is rare that state and local administrative records, such as building permits, tax assessor files, and public assistance statistics, have formal metadata attached. Rather, that information tends to be passed among employees orally as needed.⁸

While there are some comments related specifically to administrative records below, it would be useful for NICS to provide guidance on the core elements for statistical metadata applicable to administrative records at the national, state, and local levels.

The list below is our concept of a complete metadata system, the ideal. It implies that the metadata is maintained historically. The list categorizes aspects of statistical metadata as: (1) Characteristics of the Data; (2) Quality of the Data; (3) Dissemination of the Data; (4) Papers and Presentations; and (5) Training and Assistance.

We use double asterisks (**) to indicate critical items for a basic metadata set. A ** at the heading label means that all sub-bullets are “critical” if the overall category has the ** marking. Appendix A provides the list of critical elements separately.

1. Characteristics of the Data

We use statistical metadata to understand the content, scope, and purpose of the statistical data we are analyzing and to understand its limitations and possibilities for integration with other information. This provides the information necessary for a key concept of the scientific method—the ability to replicate results within measurable error limits. As such, we need a clear understanding of the target population (or “universe”), what the purposes of the survey are, where and when the data were collected, and how the data were collected. We also need an historical understanding of changes in the data set and the relationship of particular variables with apparently similar topics from other data sets.

1.1 Overview of the data set

1.1.1 Source and historical background: survey name, organizational sponsor(s) of a survey or administrative data set, organization name(s) that conducted data collection.

⁷ Management of Statistical Metadata at the OECD (v/2.0,6/9/2006), see page 3: www.oecd.org/dataoecd/26/33/33869551.pdf

⁸ Tom Kingsley, e-mail to author, September 5, 2005.

1.1.2 **Objectives** - purposes for which information is required, stated within the context of the program or research problem that gave rise to the need for information; how the information is used.

1.1.3 **Uses** - decisions to be made based on collected information and how information will support decisions.

1.1.4 **Users** - organizations, agencies, and groups expected to use the information.

1.1.5 **Type of Respondent**, such as housing units, persons (self/proxy), or establishments.

1.1.6 **Model and its assumptions** if the data are estimates or projections.

1.1.7 **Data release version and type** – whether preliminary or final, and whether this is a pilot study with a small number of cases or restricted geographic area.

1.2 Guidelines and the process for collecting and processing the data

1.2.1 ****Forms or questionnaires** – the documents that show the way respondents are asked questions, the categories for responses, and sometimes, the rules for who can be in the universe

1.2.2 **Rules for data entry** - procedures, and training given to person entering data on the form (e.g., manuals for interview rules)

1.2.3 **Data capture** - Method of data capture, accuracy rate, quality control measures

1.2.4 **Keying/scanning specifications** – The rules for data entry and processing

1.3 Population Universe, Population Coverage

1.3.1 ****Define the target population** - all the people, establishments, or other units in the data set

1.3.1.1 If administrative records, define the program participation rules and the means of collecting the data (program information provided by a respondent? through interviews with a case manager? Is information keyed and are there any quality control measures?)

1.3.1.2 If a survey, describe the sampling frame used to identify this population.

1.3.1.3 If applicable, information on eligibility criteria and screening procedures.

1.3.2 **Description of the survey design**, including the:

1.3.2.1 Results of small-scale field tests of survey procedures,

1.3.2.2 Methods used to implement the design and collect the data (such as mail, telephone, or personal interviews),

1.3.2.3 ****Sampling frame** (i.e., the sources of information such as lists, directories, and records, that cover the universe and information about any exclusions),

1.3.2.4 ****Size of the sample and the rules for selection from the universe and determination of the size,**

1.3.2.5 Sampling unit used if there is multi-stage or multi-phase sampling,

1.3.2.6 Method of estimating sampling variances, and

1.3.2.7 Disposition of sample cases (e.g., numbers of interviewed cases, ineligible cases, and nonresponding cases).

1.3.2.8 Sampling scheme (e.g., simple random sampling versus stratified sampling or sampling by probability proportional to size)

1.3.2.8.1 Inclusion probabilities and weights for each sample unit

1.3.3 ****Eligibility criteria, such as residence rules for households surveys**

1.3.4 **Universe definitions such as Household/family definition for household surveys and establishment for economic surveys**

1.3.5 **Coverage** - Measurements of the completeness of coverage of the target population and the sampling frame, that is, the extent to which all elements on the list used are members of the target population and provide measures of the extent to which units are missed and duplicated on the frame.

1.4 ****Time Frame of data set(s)**

1.4.1 **Time coverage and frequency** of the dataset.

1.4.2 **Variations in timing** - what is known about cyclical, seasonal, or other variations over time in the dataset.

1.5 **Reference period of questions** – defines over what time period the respondent should consider in their answers

1.6 Information for Using the Data

1.6.1 **Wording of questions** or information on the form of administrative records

1.6.2 **List of data elements, the range of their possible values, and their definitions** and, for the search function, their plain-English synonyms; and any changes in the definitions over time (e.g., race and ethnicity)

1.6.3 **List of data elements by data set, year of availability, lowest geographic area, and population or housing universe**

1.6.4 **Description of manipulations and adjustments**, including indexes, derived data constructed by combining information from other variables on the file (example: poverty index), and whether data are seasonally adjusted with information about the method of adjustment

1.6.5 **Unweighted frequency counts** to check tabulations from public use microdata records

1.6.6 **Variance estimates** - Explanation of how to calculate estimates of variances that are specific to the survey

1.6.7 **Record layout**, that is, the description of the data elements on the file and their physical location

1.6.8 **Code lists** used, including classification schemes for variables (e.g., the North American Industry Classification System versus Standard Industrial Classification), and recoding rules

1.6.9 **Top coded values**, if any – the categories for numeric answers, or if open-ended, if there are limits on the top value the processed data allows

1.6.10 **Unit response rates** (weighted and unweighted) for surveys and participation rates for administrative records, and how the rates are calculated

1.6.11 **Contact** for questions – names, telephone numbers, and email addresses.

1.6.12 **Errata and Notes**, including geography and data corrections

1.7 Geographic scope

1.7.1 **Geographic coverage** - areas included in data set (specific areas present in the data set)

1.7.2 **Definition of geographic components and hierarchy** - description or maps of geographic boundaries and how geographic entities are interrelated

1.7.3 **History of changes in geographic boundaries** and how handled

1.7.4 **Maps** of geographic boundaries (outlines of areas)

1.8 Comparisons

1.8.1 **Time series comparisons** – explain important changes such as the history of revisions within the data set, the character of revisions, and the effect of revisions on the data series; and legislative/program changes that would affect time series comparisons

1.8.2 **Comparability of similar data elements among data sets**, such as among states, with related surveys

1.8.3 **Procedures for adjusting dollar amount** (for example, which series from the Consumer Price Index was used or should be used for this data set?)

2. Quality of the Data

To evaluate the data for their purposes, and to understand its biases and level of precision, users draw on information about known data anomalies and a description of the sources of error

(both sampling and nonsampling) associated with the survey, how errors were calculated, and edits to the original data to account for errors. They need to know, for example, coverage as well as response rates at the unit level and for items on the questionnaires. The components below refer to the basic framework adopted throughout the statistical community to provide information about the data's accuracy, relevance, timeliness, coherence, comparability, and accessibility.

2.1 Data Limitations

2.1.1 **Statistical precision** of survey results, at least for the major estimates. This could include estimates of sampling variances, standard errors, or coefficients of variation, or presentation of confidence intervals.

2.1.2 **Nonsampling errors** - For both administrative and survey data, provide reporting errors, response variance, interviewer and respondent bias, and errors in processing the data that may affect the data, any measures of bias,⁹ and methods to deal with such problems.

2.1.3 **Edit and imputation rules** such as for nonresponse to an item and how nonresponse is handled in the database (e.g., left blank? edited? If edited, what are the edit rules for using available information and assumptions to substitute values in the data set?).

2.1.4 **Confidentiality edits** – describe the statistical techniques used to ensure that information about individuals is not released.

2.1.5 **Weighting scheme** for survey data, including adjustments for nonresponse and benchmarking and how to apply them.

2.2 Advanced Methodology

2.2.1 **Evaluations** of the accuracy of the data - studies

2.2.2 **Data quality** - Provide research that measures data quality and explain measures to gauge the quality of the data

2.2.3 **Quality of address reporting, household composition**

3. Dissemination of the Data

Data producers release information to the public and data users need to understand the avenues for access and when they can get it. They also need to be advised if there are revisions to a previously released data set and the procedures the producer uses to protect the confidentiality of the data. Documentation needs to be provided for both summary tabulations of the data and Public Use Microdata files (PUMS).

3.1 Data dissemination and release schedule

3.1.1 **How to obtain data** – for example, whether the data are available through the internet, in a publication, or in some other format, along with any restrictions on uses and distribution of the data

3.1.2 **Data products, type** – description of the choices of data products
3.1.3 **Data release schedule** – are data released periodically on a regular schedule? Or as they are available? Is there a central notification point so that data users can find out about new releases?

3.1.4 **Timeliness** - length of time between data availability and the event or phenomenon it describes (context of value and use).

3.2 **Confidentiality procedures** – what are the edit procedures to protect the confidentiality of the data and to what extent does that affect the results and analyses?

⁹ Bias is defined as the deviation of the average survey value from the true population value.

3.3 **Sponsor/legal authority** - agency(s) or organizations responsible for sponsoring the data collection, processing, and dissemination under U.S. or state codes or contracts.

3.4 ****Additional documentation for Public Use Microdata Sets¹⁰**

Describes construction of the information and how to access and manipulate the data.

4. **Papers and presentations**

Professional papers and presentations related to the data set, including analysis of policy questions, research about the quality of the data, and decision memoranda help data users deepen their understanding of issues related to the dataset.

5. **Training and Assistance**

Training introduces data users to basic concepts, terminology, examples, and helpful hints and solutions.

5.1 **User Training**

5.1.1 A **“Wizard”** to walk the data user through the steps of a software application

5.1.2 **How to use specific data sets** – Training on what affects analyses, such as the survey design or administrative rules, products that are available, caveats, comparisons with similar data, and data limitations

5.1.3 A **listserv** to provide alerts about data problems, education about data sets, share with many people immediately and create a community of problem solvers. Web-based systems can include communities of practice that allow users to share challenges and solutions, exchange experiences involving real-world applications of data, and gain access to experts.

5.1.4 **Organized constituencies** (for example, Association of Public Data Users)

5.1.5 **Data security** - Educate data users about the physical and statistical security of data, especially matched data sets.

5.2 ****Contact for further information and assistance** — specifics of who and how.

¹⁰ Public Use Microdata Samples (PUMS) are computer-accessible files containing survey records for a sample of housing units, with information on the characteristics of each housing unit and the people in it. PUMS files allow users to prepare tabulations according to their own specifications. Identifying information is removed to protect the confidentiality of the individual respondents.

III. WHAT WEB-BASED TOOLS ARE AVAILABLE?

Some data sets have metadata available already, although the format, content, and vocabulary differ greatly among datasets. Other datasets have no formally organized metadata attached. What web-based tools are currently available that can handle large and small data sets? What existing metadata models can support the heterogeneous attributes of the hundreds of data sets? How can information from different sources be integrated?

Figure 1 is a summary of selected web-based tools now available or under construction. See the NICS website section on metadata for further discussion of each listed below (<http://www.nicsweb.org/metadata/>). Below the figure, we discuss resources that seem most relevant to NICS (shown in bold font in Figure 1) in terms of each tool, the institutional framework for each tool, the purpose and outputs of the tools, and which tools may be useful to NICS. There is also an explanation of ontology tools (schemes to represent knowledge).

Figure 1. Summary of Metadata Resources by Type

(**Bold font** indicates the resource is discussed in more detail in the section that follows this figure.)

Resource	Resource Description	Type of Resource				
		Standards	Projects	Conferences	Software	Products
The Resource Description Framework (RDF) - http://www.w3.org/Metadata/Activity	PICS work led to the development of the Resource Description Framework (RDF), which provides a more general treatment of metadata. RDF is a declarative language and provides a standard way for using XML to represent metadata in the form of statements about properties and relationships of items on the Web.	X				
Extensible Markup Language (XML) - http://www.w3schools.com/xml/xml_watis.asp	XML stands for Extensible Markup Language; XML is a markup language much like HTML and was designed to describe data	X				
Semantic Interoperability Projects of the European Interoperability Framework's Interchange of Data between Administrations (IDA). See pp. 26–27: http://europa.eu.int/idabc/servlets/Doc?id=22108	"Semantic interoperability is a necessary component in achieving full interoperability since it is concerned with ensuring that the precise meaning of exchanged information is understandable by other parties."		X			

Resource	Resource Description	Type of Resource				
		Standards	Projects	Conferences	Software	Products
TKME - http://geology.usgs.gov/tools/metadata/tools/doc/tkme.html	TKME is an editor for formal metadata, or structured documentation conforming to the Content Standard for Digital Geospatial Metadata developed by the Federal Geographic Data Committee (FGDC).				X	
M3CAT - http://www.intelec.ca/technologie_a.htm	The multistandard and multilingual metadata cataloguing M ³ Cat is a tool for the creation of geospatial metadata				X	
Nesstar Explorer - http://www.nesstar.com/	NESSTAR is a Semantic Web application for statistical data and metadata that aims to streamline the process of finding, accessing and analyzing statistical information.				X	
Global Justice XML Data Model (GJXDM) – http://it.ojp.gov/topic.jsp?topic_id=43	The Global JXDM is a comprehensive product that includes a data model, a data dictionary, and an XML schema. The Global JXDM is an XML standard designed specifically for criminal justice information exchanges.	X				

Resource	Resource Description	Type of Resource				
		Standards	Projects	Conferences	Software	Products
National Information Exchange Model (NIEM) – http://www.niem.gov/	The National Information Exchange Model (NIEM) is a Federal, State, Local and Tribal interagency initiative providing a foundation for seamless information exchange.		X			
Data Reference Model (DRM) - http://www.whitehouse.gov/omb/egov/a-5-drm.html	The DRM categorizes government information into greater levels of detail. It also establishes a classification for Federal data and identifies duplicative data resources. A common data model will streamline information exchange processes within the Federal government and between government and external stakeholders. The DRM provides a standard means by which data may be described, categorized, and shared.	X	X			

Resource	Resource Description	Type of Resource				
		Standards	Projects	Conferences	Software	Products
Statistical Data and Metadata Exchange (SDMX) - http://www.sdmx.org/about/index.aspx	An exchange between the BIS, ECB, EUROSTAT, IMF, OECD, UN, and the World Bank. They have joined together to focus on business practices in the field of statistical information that would allow more efficient processes for exchange and sharing of data and metadata within the current scope of our collective activities.	X	X			
Dublin Core Metadata Initiative - http://dublincore.org/ and http://archive.dstc.edu.au/RDU/reports/Sympos97/metadata.html	The Dublin Core Metadata Initiative is an open forum engaged in the development of interoperable online metadata standards that support a broad range of purposes and business models. DCMI's activities include consensus-driven working groups, global conferences and workshops, standards liaison, and educational efforts to promote widespread acceptance of metadata standards and practices.	X				
MARC- http://www.loc.gov/marc/marc.html	The MARC formats are standards for the representation and communication of bibliographic and related information in machine-readable form.	X				

Resource	Resource Description	Type of Resource				
		Standards	Projects	Conferences	Software	Products
Data Documentation Initiative (DDI) – http://www.icpsr.umich.edu/DDI/index.html	The Data Documentation Initiative is an international effort to establish a standard for technical documentation describing social science data.	X				
FGDC – http://www.fgdc.gov/metadata/metadata.html			X			
METS - http://www.loc.gov/standards/mets/METSOverview.v2.html	METS, a Digital Library Federation initiative, attempts to build upon the work of MOA2 and provide an XML document format for encoding metadata necessary for both management of digital library objects within a repository and exchange of such objects between repositories (or between repositories and their users).		X			
Meta Content Framework - http://www.w3.org/TR/NOTE-MCF-XML/MCF-tutorial.html#sec1	The Meta Content Framework (MCF) is a tool to provide information about information by attaching properties to objects					X
DataFerrett - http://dataferrett.ce.nsus.gov/index.html	The DataFerrett is a data mining tool that accesses data stored in The DataWeb through the internet.				X	

Resource	Resource Description	Type of Resource				
		Standards	Projects	Conferences	Software	Products
GovStat Project – http://www.ils.unc.edu/govstat/	The GovStat Project seeks to create an integrated model of user access to and use of US government statistical information that is rooted in realistic data models and innovative user interfaces.		X			
Statistical Knowledge Network (SKN) - http://ils.unc.edu/govstat/papers/SKN_version6.doc	The Statistical Knowledge Network's (SKN) purpose is to support the transmission, sharing, understanding, and use of statistical expertise and information. The SKN supports a consortium of people, organizations, and resources devoted to government statistical information.		X			
Statistical Interactive Glossary (SIG) - http://eprints.rclis.org/archive/00004996/01/Brown_GovStat.pdf	The GovStat Statistical Interactive Glossary (SIG) was developed as part of a larger resolution by the GovStat Project to deliver help in an online environment (Marchionini <i>et al.</i> , 2003). The SIG was designed to provide useful explanations to users as they browsed a statistical report or table.					X

Resource	Resource Description	Type of Resource				
		Standards	Projects	Conferences	Software	Products
Environmental Information Exchange Network - http://www.epa.gov/neengprg/info/index.html	The Environmental Information Exchange Network (Exchange Network) is a new approach for exchanging environmental data between EPA, states, and other partners. Using the Internet and standardized data formats, the Network exchanges information between nodes, or portals maintained individually by participating partners.		X			
"Guidelines for Statistical Metadata on the Internet" – United Nations Statistical Commission - http://www.unece.org/stats/publications/metadata.pdf	The guidelines were prepared by Statistics Norway with the assistance of a working group composed of Canada, USA, EFTA, Eurostat, OECD, UNSD and the UN/ECE secretariat. Other ECE member countries and international organizations participating in the international work on statistical metadata organized under the umbrella of the Conference also contributed to the preparation of this material.		X			

Resource	Resource Description	Type of Resource				
		Standards	Projects	Conferences	Software	Products
US Census - http://www.census.gov/srd/www/metadata/ASA96TOC.HTML	The Bureau of the Census is developing a Statistical Metadata Content Standard to define the necessary metadata to describe all aspects of survey design, processing, analysis, and data sets.		X			
XML - http://www.w3.org/XML/	Extensible Markup Language (XML) is a simple, very flexible text format derived from SGML (ISO 8879). Originally designed to meet the challenges of large-scale electronic publishing, XML is also playing an increasingly important role in the exchange of a wide variety of data on the Web and elsewhere.	X				
XML Schema - http://www.w3.org/XML/Schema	XML Schemas express shared vocabularies and allow machines to carry out rules made by people. They provide a means for defining the structure, content and semantics of XML documents in more detail. XML Schema was approved as a W3C Recommendation on 2 May 2001 and a second edition incorporating many errata was published on 28 October 2004	X				

Resource	Resource Description	Type of Resource				
		Standards	Projects	Conferences	Software	Products
XSL - http://www.w3.org/Style/XSL/	XSL is a family of recommendations for defining XML document transformation and presentation. It consists of three parts: XSL Transformations, XML Path Language, XSL Formatting Objects	X				
XSLT - http://www.w3.org/TR/xslt	XSLT is designed for use as part of XSL, which is a stylesheet language for XML. XSLT is also designed to be used independently of XSL. However, XSLT is not intended as a completely general-purpose XML transformation language. Rather it is designed primarily for the kinds of transformations that are needed when XSLT is used as part of XSL.	X				
Common Warehouse Metamodel (CWM) from Object Management Group (OMG) - http://www.omg.org/cwm/	The Common Warehouse Metamodel (CWM™) is a specification that describes metadata interchange among data warehousing, business intelligence, knowledge management and portal technologies.	X				

Resource	Resource Description	Type of Resource				
		Standards	Projects	Conferences	Software	Products
ISO 19115 (Geospatial metadata) - http://www.iso.ch/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=26020	ISO 19115:2003 defines the schema required for describing geographic information and services. It provides information about the identification, the extent, the quality, the spatial and temporal schema, spatial reference, and distribution of digital geographic data.	X				
ISO/IEC 11179 - http://metadata-standards.org/11179/	An international standard for describing the semantics of data, and managing and registering that information. The 6 parts are framework, classification, metamodel and basic attributes, definitions, naming, and registration.	X				
METIS - UNECE / Eurostat/ OECD Statistical Metadata Work Session - http://www.unece.org/stats/archive/04.01.e.htm	A series of meetings for statistical organizations (national and international) to discuss advances in statistical metadata theory, practice, and experience.			X		
Common Metadata Framework - http://www.unece.org/stats/cmf/	A project to organize and synthesize the vast pool of information from the statistical metadata community into a framework of best practices for national statistical offices to follow.		X			

Resource	Resource Description	Type of Resource				
		Standards	Projects	Conferences	Software	Products
OMG's Ontology Definition Metamodel - http://www.omg.org/ontology/	Defines a family of independent metamodels, related profiles, and mappings among the metamodels corresponding to several international standards for ontologies.		X			
Semantics of Business Vocabulary and Business Rules (SBVR) - http://www.omg.org/docs/bei/05-08-01.pdf	SBVR is designed to support interchange of business vocabularies and rules among organizations. SBVR is conceptualized optimally for business people and designed to be used for business purposes independent of information systems designs.	X				

Source: Typology devised by Laura Smith (Brookings Institution, Urban Markets Initiative) and compiled by Andrew Reamer, Laura Smith, and Cynthia Taeuber.

Federal statistics are a good place to start with the issues of standardization and automation of metadata. The FedStats website (<http://www.fedstats.gov/>) links data users to statistics and associated metadata from more than 100 federal agencies and hundreds of related websites. The federal statistical community understands the need to standardize the components and attributes of metadata to make it easier to find and use data across agencies. We see evidence of this in the proposed standards and guidelines for metadata in surveys.¹¹ Many statistical agencies tell data users how their data compare with data and concepts from other agencies.

Efforts to automate standard metadata and make it available across data sets are related to the opportunities presented by the growing use of the Internet in recent years. The National Science Foundation (NSF) has supported research to expand the ability of government to better use technology. Significant support for integrating information has come most recently as a result of the security interests of the Department of Homeland Security (DHS).

The proposed metadata principles and guidelines for OMB's Statistical Policy Directives 1 and 2 would provide in-depth and rigorous information about a dataset to extend and expand the

¹¹ Office of Management and Budget Statistical Policy Office, <http://www.whitehouse.gov/omb/inforeg/statpolicy.html#pr>

existing standards for automated metadata. For example, the Dublin Core is cited as a “higher level metadata standard” yet it consists of only 16 elements.¹² As the elements indicate, the Dublin Core is general and was developed to meet the needs of librarians. The International Organization for Standardization, ISO/IEC 11179, specifies a likewise limited set of data elements needed to share data.

The ISO¹³/IEC¹⁴ 11179 - Metadata registries - standard is a metadata specification devoted to data semantics. It also contains a model and an overview of a procedure for registration, hence the "registries" in the name. However, the main focus is the semantics of data. Here data is not restricted to structured data; any kind of data may be described.

The standard is divided into six parts, each of which describes an aspect of the standard. A short description of each part follows:

- Part 1 - Framework -- an overview of the standard and the methodology behind data semantics
- Part 2 - Classification -- presentation of a model for managing a classification scheme, especially as it relates data elements (variables) to each other
- Part 3 - Metamodel and basic attributes -- presentation of the full model for data semantics, classification, and registration
- Part 4 - Formulation of data definitions -- principles for writing good data definitions
- Part 5 - Principles for naming and identification -- provides a naming convention for each of the principal parts of data semantics
- Part 6 - Registration -- procedures for registration

The last published version of the standard is the 2nd edition, completed in 2005. All the latest published parts of ISO/IEC 11179 are freely available on the web¹⁵. The 1st edition of the standard, published in 2000, was superseded by the 2nd. It was called Standardization and specification of data elements. The change in focus away from just data elements in the 1st edition necessitated a change to Metadata registries.

The basic unit for describing data in ISO/IEC 11179 is the data element (variable). The model specified in the standard shows how one should describe a data element. It is concept based and follows the general framework of the terminological theory of data¹⁶.

However, the standard does not address statistical data per se. It contains a general description of data, and does not go any further than that. Even the idea of a data set is not described in the standard.

12 The elements of the Dublin Core are: Coverage; Description; Type; Relation; Source; Subject; Title; Audience; Contributor; Creator; Publisher; Rights; Date; Format; Identifier; and Language.

13 International Organization for Standardization

14 International Electrotechnical Commission

15 Information Technology Task Force (ITTF) under ISO and IEC (http://isotc.iso.org/livelink/livelink/fetch/2000/2489/Ittf_Home/ITTF.htm).

16 Metadata Standards and Their Support of Data Management Needs. Paper #7 at <http://www.unece.org/stats/documents/2006.04.metis.htm>

The ISO 19139 - Metadata – Data set implementation- allows for documentation of geographic and non-geographic data and is written in XML.¹⁷ Kules and Shneiderman note that federal statistical agencies are generally not funded to catalog the metadata they produce in any machine-readable standard.

DataFerrett's Metadata Tool¹⁸

DataFerrett provides access to microdata and aggregate data from the Census Bureau, the Bureau of Labor Statistics, and other agencies.¹⁹ It is a web-based tool that handles large and small data sets along with the heterogeneous attributes of the various data sets.²⁰ As well as allowing users to find data across various data sets, it includes the metadata provided by the supplying agency that is responsible for the data collection. DataFerrett allows metadata to be corrected, updated, and maintained historically.

The DataFerrett website provides the Metadata Interface File (MIF) documentation at: http://www.thedataweb.org/mif_usersguide.html. The MIF includes the name of the data collection, the name of each dataset within that collection, and critical metadata items: the time period for the dataset, including whether it is a continuing data set or whether there is a stop date; the name, description, synonyms for the variable name, and values of each item in the dataset; whether there is an associated weight for an item from a sample survey; confidentiality edits; recodes; allocation flags; top coding; geography level; and the security level (public data or sponsor only). Supplementary information can be included if it is provided by the supplying agency. The MIF is an ASCII file that is used to populate the DataFerrett metadata database.

There is a search engine for concepts and definitions of variables. Data users can find variables from a list of those in a data set as well as through a keyword search. Data users can click on a hyperlink to see descriptions and the technical documentation for data sets as provided by the collection agencies; likewise, a user can view the definitions of variables as they are supplied by the agency. Once a set of variables is selected, the user can simply highlight a variable name to read the variable's description, the question text, the answer categories (values), the universe, and information about related variables IF the metadata has been supplied.

The DataFerrett tools of particular value to NICS include identification of all datasets in its system with information about a topic of interest (e.g., housing vacancy), automatic access to definitions and metadata related to a dataset, and the ability to update the metadata and maintain it historically.

Networked Social Science Tools and Resources (NESSTAR) Explorer

17 Jeff Partridge, "Developing a Metadata Template for CDC," http://www.cdc.gov/phinf/05conference/05-11-05/4C_Patridge.pdf.

18 Ferrett stands for "Federated Electronic Research, Review, Extraction, and Tabulation Tool." See <http://dataferrett.census.gov/>

19 Every record is a unit of analysis.

20 All records are added up to get totals for each category in the universe.

NESSTAR Explorer is an effort to create a “data web” to make it easy to publish, locate, and access statistical data. It is similar to a normal Web Browser. NESSTAR is “a Semantic Web application for statistical data and metadata that aims to streamline the process of finding, accessing and analyzing statistical information.”²¹ The Semantic Web “is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation.”²² The glossary of this paper also includes further information for “semantic web.”

NESSTAR Explorer is a data publishing web tool that includes very limited metadata. “Users use the system pretty much as they use the Web: if they know where some information is stored they can point” their client application to it (for example by typing the object URL in a location bar or by clicking on a hyperlink). The client will access the remote statistical object and display it to the user. The users can also perform searches to find objects with particular characteristics such as: “find all variables about political orientation.” This is similar to using a search engine such as Google to find all HTML pages that contain a given keyword.”²³

It uses data archives that are compliant with the Data Document Initiative (DDI) specification. The DDI is a way to document social science data and metadata in standardized Extensible Markup Language (XML) to make it easier to process by computer.²⁴

XML is a recommendation of the World Wide Web Consortium (W3C). It is a text-based language for marking up the content of documents and data sets. Marking up is the process of placing tags (essentially, names) around sections of data in a file. The tags are user defined, and provide a way to embed some metadata along with the data.

The sections of data within an XML document named by tags are called elements. Elements are structured in a hierarchy. There is much more detail to these technical issues, and the interested reader can learn more at the W3C web site for XML - <http://www.w3.org/XML/>.

Marked up text is easy to process by computers. The structure of an XML marked up document is well-defined (through a schema language, called XML-Schema), so a computer can easily tell if the mark up is valid. Also, XML uses text as its format; it is readable by a human. This eases the burden of processing XML.

XML is related to HTML, the mark up language for the Web. The main difference is that HTML tags are pre-defined whereas XML tags are defined by the user. HTML is used for formatting

21 NESSTAR Technical overview: http://www.nesstar.org/Release-free/Technical_overview.pdf. Also see <http://www.nesstar.com/>. NESSTAR is a wholly owned subsidiary of the UK Data Archive and the Norwegian Social Science Data Service.

22 James Hendler, Tim Berners-Lee, and Ora Lassila, “The Semantic Web,” *Scientific American*, May 2001.

23 NESSTAR Technical overview: http://www.nesstar.org/Release-free/Technical_overview.pdf.

24 The Data Documentation Initiative is an international effort to establish a standard for technical documentation describing social science data. Data Documentation Initiative (DDI) Homepage: <http://www.icpsr.umich.edu/DDI/>; see <http://www.icpsr.umich.edu/DDI/codebook/faq.html>.

data (especially documents). However, another XML-like language, XSL (eXtensible Stylesheet Language), is used for formatting XML documents.

The metadata in NESSTAR Explorer is specific to a dataset and does not provide a function to compare features of the metadata among datasets. While NESSTAR uses machine-readable standards for metadata entries, it is too limited for NICS. We would have to explore whether its boundaries could be extended to allow more depth to the metadata content and whether it could be developed to allow comparisons of the metadata among sets of data.

To this end, NICS has been involved with DDI activities and updates, including its major revision currently underway.²⁵ The DDI aims increase its focus on the role of metadata in the survey life-cycle, thereby strengthening the metadata base on which NESSTAR operates, Through NICS involvement, we hope to provide input and support for such efforts.

National Information Exchange Model and the Federal Enterprise Architecture Data Reference Model

Federal agencies are moving towards Enterprise Architecture and away from information silos and technology that cannot be used across agencies. The Department of Justice (DOJ) and the Department of Homeland Security (DHS) have a partnership to develop an XML-based core data model,²⁶ the National Information Exchange Model (NIEM).²⁷ NIEM is an inter-agency initiative that exchanges data among justice and public safety agencies as well as by agencies beyond the justice community. For NICS, NIEM is a proven system that demonstrates that by using XML, data can be exchanged among different computer systems. The user sees a web interface and can access information.

NIEM is an expansion of a limited exchange model called "Global JXDM" that was developed by DOJ. The Global Justice XML Data Model (GJXDM) includes a data model, a data dictionary, and an XML schema (the rules for encoding information on the World Wide Web). It is a tool that allows data to be structured so it can be shared and understood among different systems. Individual systems can remain whatever they are. The data that comes out of a system is transformed or annotated so it can be interpreted by another system.

At the heart of GJXDM is the data dictionary, a critical part of documenting statistical systems such as in NICS. Jim McKay describes the data dictionary as "...a spreadsheet containing identification of data elements, and the meanings or definitions of those data elements, all of which are unique. The data model builds relationships between the data elements, and the result, in simple terms, is that disparate systems connect via the unique identifiers."²⁸ The developers "removed the redundancies and duplications and resolved semantic differences. Currently, Global JXDM consists

25 DDI Version 3.0 has been released for Public Review and evaluation through mid-April 2007. Visit <http://www.ddialliance.org/ddi3/index.html> for more information and to provide your feedback.

26 "Universal core data types" cover the interests of all the partners while "core data types" are of interest to two or more partners.

27 See http://www.niem.gov/pdf/20050307_press_release_dhs_doj_global_jxdm_exec_briefing.pdf and <http://www.niem.gov/implementation.php>

28 Jim McKay, "XML Out of the Shadows," Government Technology, June 2005, <http://www.govtech.net/magazine/story.php?id=94099>

of a well-defined and organized vocabulary of 2,754 reusable components out of which there are 400 Complex Types, 150 Simple Types, and 2,209 Properties that facilitate the exchange and reuse of information from multiple sources and multiple applications."²⁹ After spending a considerable amount of time to find common data elements, the developers then had to develop common definitions for similar concepts that could be used across agencies. This experience would be valuable for NICS to draw upon and develop guidance for doing the same thing for NICS, especially where surveys are concerned. The content of administrative record files are likely to change more often than surveys, but both change and that reality needs to be built into the plans from the start.

GJXDM has already proved its efficiency and ability to save money. Important to NICS is the comment by Bureau of Justice Assistance (BJA) Director Domingo Herraiz: "The best news about this model is that there's no secret to duplicating its success. We're receiving reports from numerous states on improved information sharing and the cost of efficiency of implementation."³⁰ Herraiz made the analogy of travelers with different native languages sharing information through another language, such as English or French, which is common to all of them.³¹ For example, Wisconsin has employed GJXDM so multiple agencies can obtain driver and vehicle records and check for stolen vehicles and wanted persons as well as do criminal history searches. In Ohio, 900 separate police departments can now exchange data and information.

NIEM is broadening the scope of the GJXDM and exchanging new data. They are also developing technology to simplify and safeguard access to the information by trusted users whose identities have previously been validated. Some express their concern that the system will become too complex. The Georgia Tech Research Institute is developing a "subschema generator" tool to handle a large model. NIEM is developing rules for data that states are willing to share, taking into account the differences in state laws about access. They have found that education and training are barriers to bringing data owners onboard.³² Additionally, NICS should plan to track changes in state laws and policies.

FEA/DRM

To accomplish these efforts, NIEM partners are collaborating to develop and implement common XML standards for exchanging data through the Federal Enterprise Architecture Data Reference Model (FEA DRM)³³ described below. The partners are also developing an XML profile of NIEM to implement the FEA DRM. They expect to publish a paper on the concept of NIEM operations in September 2005 (see: <http://www.niem.gov/library.htm>).

²⁹ National Information Exchange Model, see <http://www.niem.gov/aboutniem.htm>

³⁰ Jim McKay, *Ibid.*

³¹ Domingo Herraiz, "The Pathway to Success in Information Sharing: Where the Global Justice XML Data Model Is Today," *Police Chief Magazine*, June 2005, http://www.policechiefmagazine.org/magazine/index.cfm?fuseaction=display&article_id=611&issue_id=62005

³² Jim McKay, *op. cit.*

³³ See <http://www.niem.gov/implementation.php>

To further data exchange, the U.S. Office of Management and Budget (OMB) has established interagency working groups to develop the Federal Enterprise Architecture (FEA) Data Reference Model (DRM).³⁴ The vision of the DRM, which classifies federal data, is to improve the ability of decision makers between and within government agencies to get the right information to the right place at the right time. The purpose of the DRM, then, is to help federal agencies to use standard approaches to (1) find the right data through data descriptions; (2) exchange data by describing the requirements of the exchange and the characteristics of the data; and (3) understand the context of data by applying standard approaches to metadata to describe, share, and categorize data as formal taxonomies that classify and define the relationship among data elements. A formal taxonomy defines the category and links to a data element³⁵ and, as stated in the official description of the DRM, “requires an approach to the common categorization, exchange, and structure of data.”³⁶ Taxonomies provide a scheme for classifying the content of metadata and to organize unstructured information, such as word-processing documents and PDF files. This includes tagging data according to its security and privacy attributes, a huge advantage in keeping track of a critical requirement in managing multiple datasets with different rules.

The description of the scope of the DRM says that, given “the decentralized nature of data management in the federal government, the varied existing statutes, policies and directives, the DRM only requires that agencies implement a standard template for building their information architecture. Use of the DRM, in and of itself, does not mandate or create information sharing. Departments and Agencies will still decide what data to share based on common business needs.”³⁷ NICS will want to work closely with the DRM team to ensure that system requirements for statistical metadata are included.

The GovStat Project

Ideally, NICS needs a repository of information that is machine-readable and standardized in terms of format, content, and vocabulary. The GovStat Project moves towards achieving those goals. It is a joint effort of the University of North Carolina Interaction Design Lab and the University of Maryland Human-Computer Interaction Lab that is funded by the National Science Foundation. The Project has created interfaces for data users to better access federal statistical information. Their objective is a unified Statistical Knowledge Network (SKN)³⁸ that integrates heterogeneous information across federal statistical datasets, provides help finding and comparing information, and providing alternatives for finding and viewing information. The SKN project has identified the need to be able to make comparisons of methods, concepts, scope, time periods, and geographic coverage.³⁹

34 FEA DRM Schema Specification (Draft Version 0.1), <http://web-services.gov/lpBin22/lpext.dll/Folder17/Infobase6/1/50c/688/6f9?fn=main-j.htm&f=templates&2.0>

35 Michael Daconta, “Formal Taxonomies for the U.S. Government,” <http://www.xml.com/pub/a/2005/01/26/formtax.html>

36 Overview of the Data Reference Model, http://colab.cim3.net/cgi-bin/wiki.pl?DataReferenceModel_09_2004/OverviewOfThe_DRM_Vollv1

37 The Federal Enterprise Architecture Data Reference Model: A Synopsis,” NSF Collaborative Expedition Workshop #43, August 16, 2005, http://colab.cim3.net/file/work/Expedition_Workshop/2005_08_16_DesigningTheDRM_forDataAccessibility/McCaffery_DRM_Synopsis_2005_07_15.doc

38 See: http://ils.unc.edu/govstat/papers/asist-03ont-gloss_files/frame.htm

39 Carol A. Hert, “Current Directions for GovStat Metadata Efforts,” Slide 9, December 5, 2003, http://ils.unc.edu/govstat/papers/FutureDirections_files/frame.htm

The SKN includes a Statistical Interactive Glossary (SIG)⁴⁰ designed to help data users understand statistical terms and related concepts. It allows users of federal government statistical websites to see definitions of statistical terms while browsing statistical websites. The SIG covers a limited set of terms and related concepts (ontology) that non-expert data users come upon frequently in various datasets. While the terms are too limited for this project, there is no reason that NICS would have to be similarly confined.

The definitions are written in plain English for data users with only a basic level of statistical literacy. NICS could use their work on plain-English definitions as an example to encourage those who will create metadata for other datasets to do likewise. It is likely, however, that in the end, NICS will take metadata that is not written as plainly as is desirable. This project may be helpful to NICS in developing a common thesaurus of metadata terms that data providers can draw on as they write metadata (also see the related objectives of the IMF Metadata Repositories Project below).

The GovStat Project has developed principles for the SKN and SIG that are useful for NICS. For example, while the SIG sometimes points to more advanced and related resources, new principles help to minimize interruption to the user's work task by incorporating these resources into the context of the work the data user is doing.

The Statistical Data and Metadata Exchange (SDMX)

The Statistical Data and Metadata Exchange (SDMX) initiative (www.sdmx.org) is a cooperative project of the Bank for International Settlements, the European Central Bank, the International Monetary Fund (IMF), the Organization for Economic Co-operation and Development, the Statistical Office of the European Communities, the United Nations Statistical Division, and the World Bank.

The SDMX initiative sets standards to aid the exchange of aggregated statistical data and its associated metadata. Of interest to NICS, the system handles data and the associated metadata regardless of the content, presentation, or compilation techniques and can handle administrative and survey data.

Like NICS, an objective of SDMX is to automate the collective sharing and exchanging of economic and financial statistics from various sources along with the associated metadata. To do this, they developed the SDMX, a standardized metadata system that supports a large number of data sets provided by national statistical agencies and central banks to disseminate information about statistical standards and practices for the data sets of members through the Dissemination Standards Bulletin Board (DSSB).⁴¹ The Statistics Office of the International Monetary Fund (IMF) says:

40 See: http://ils.unc.edu/govstat/papers/asist-03ont-gloss_files/frame.htm; and <http://ils.unc.edu/govstat/papers/brown-asist-abstract.doc>.

41 See: http://www.sdmx.org/data/DOC01_Framework_V01_website.pdf; and

<http://72.14.207.104/search?q=cache:RV4RoMnWveAJ:www.sdmx.org/Data/IMF%2520Metadata%2520Rep%2520Project.doc+smdx+metadata&hl=en>

“The use of a standard presentation format for statistical metadata on the DSBB enables data users worldwide to gain access to information in a readily recognizable and comparable form. In recognition of this, in March 2003 the IMF launched an enhanced DSBB website that transformed the existing set of static pages into pages that are dynamically generated on request. This has greatly increased the DSBB’s flexibility to meet specific user needs by enabling users to order up a set of web pages containing information gleaned from a search and query operation in formats tailored to their purpose.

“Despite the success with which these enhancements have been received, the existing DSBB metadata query facility lacks the capability and compatibility to interact fully with other sources of statistical metadata available at the national and international levels. This is because a common vocabulary, an internationally agreed model articulating the manner in which information is stored, and a standard format for rendering metadata and macroeconomic time series data has not yet been sufficiently developed.”⁴²

SDMX is primarily for time series data although it can also provide some support for cross-sectional data, and eventually for hierarchical classifications. Its standards support the requirements for batch exchanges, generic processes for any type of metadata, and automated processes that transform metadata into “application-specific processing formats, other standard formats, and presentation formats.”⁴³ It can use both HTML and the XML format for the exchange of data and metadata between providers and consumers across the Internet.

To support data exchange and queries across diverse repositories of metadata, the IMF has a Metadata Repositories Project. The goal of this project is to identify and use commonalities among metadata sets to develop standard format, structure, and vocabulary for open metadata repositories. SDMX maintains a framework for the standardization of concepts, terminology, and key families within the statistical domain.⁴⁴ The project includes development of a common thesaurus of metadata terms that promotes standard names, common vocabulary, and definitions for core statistical concepts.

Ontology Tools

Ontologies are schemes to represent knowledge. To use ontology tools that assess and link information from diverse sources, NICS will need machine-readable standard metadata that has a finite vocabulary that can be clearly classified, relationships among variables that can be specified, and a hierarchical relationship of subclasses among classes.⁴⁵ Michael Denny described ontologies as “...a way of specifying the structure of domain knowledge in a formal logic designed for machine processing. The effect on information technology (IT) is to shift the burden of capturing the meaning of data content from the procedural operations of algorithms and rules to the representation of the

42 Statistics Department, International Monetary Fund, “The IMF Metadata Repository Project: An Activity Aligned with SDMX Standards,” <http://72.14.207.104/search?q=cache:RV4RoMnWveAJ:www.sdmx.org/Data/IMF%2520Metadata%2520Rep%2520Project.doc+smdx+metadata&hl=en>

43 Statistical Data and Metadata Exchange Initiative, “Framework for SDMX Standards (Version 1.0),” p. 8, 2004, www.sdmx.org/

44 Statistical Data and Metadata Exchange Initiative, “Metadata Common Vocabulary,” http://www.sdmx.org/Data/SDMX_MCV_release1_200404.pdf

45 Deborah McGuinness, “Ontologies Come of Age,” in *The Semantic Web: Why, What, and How*, MIT Press, 2002, p. 6.

data itself.”⁴⁶ In his survey of ontology tools, Denny found 96 commercial and open-source ontology editors and classified their characteristics into 13 categories.⁴⁷

Some of the tools automatically extract metadata from the original text documents, a clear benefit to NICS as this should ease updating metadata. Web services such as Yahoo and Google use ontology-based approaches to find and categorize information on the Internet. The Federal Enterprise Architecture uses ontologies as a framework for identifying the logical relationship of information.

The construction of ontologies, as Denny describes, is an iterative process that builds from core concepts. Interpretation of information comes as a result of how the text is classified, restricted, and entailed⁴⁸ (that is, inferring the meaning and relationship of text from other text). Current entailment systems have relatively low accuracy (less than 60 percent).⁴⁹

Denny surveyed users and asked them about enhancements they would like to see in ontology editors. “Users would like tool features that make building full-blown ontologies easier and more foolproof, especially for domain experts rather than ontologists.”⁵⁰ Brand Niemann notes that “A commercial Integrated Development Environment (IDE) for ontology building does not appear to exist but several are in development.”⁵¹ Because the ontology editors offer a range of functions, Denny concludes that, “In the absence of an IDE for ontologies, tried and true or otherwise, the practical approach today is to rely on several ontology building tools to fashion different aspects of an ontology and manage the development process.”⁵² Nevertheless, it is difficult to know all the concepts, relationships, and computational models that are needed to construct an ontology.

Topic-specific initiatives

Some initiatives are specific to a topic, including education, health, and the environment. The Department of Justice programs were noted above. Below are selected examples of federal-state partnerships that developed statistical administrative records with metadata.

- Education Data Exchange Network (<http://www.ed.gov/about/inits/ed/pbdmi/eden/workbook.doc>) - This network is a federal-state-industry partnership that provides a standard format and content for data from states about the performance of education programs, schools, and students. Their goals are to improve access to data by the public, data quality, and timeliness. Metadata is limited and includes, for example, some definitions, how confidentiality and missing items are handled, edits, and a detailed record layout.

46 Michael Denny, “Ontology Tools Revisited,” July 14, 2004, <http://www.xml.com/pub/a/2004/07/14/onto.html>

47 The specific software and categories are organized by Denny at: http://www.xml.com/2004/07/14/examples/Ontology_Editor_Survey_2004_Table_-_Michael_Denny.pdf

48 Ido Dagan, Oren Glickman and Bernardo Magnini, “The PASCAL Recognising Textual Entailment Challenge,” in the Proceedings of the PASCAL Recognising Textual Entailment Challenge, April 2005. See: <http://www.pascal-network.org/Challenges/RTE/Introduction/>; and <http://en.wikipedia.org/wiki/Entailment>

49 Rajat Raina, <http://hunch.net/index.php?p=100>. Also see: <http://www.cs.biu.ac.il/~glikmao/rte05/>.

50 Denny, op.cit. Also see Niemann, SCOPE, op.cit, slide 26.

51 Niemann, SCOPE, op.cit, slide 26.

52 Denny, op.cit.

- Environmental Information Exchange Network (<http://www.epa.gov/neengprg/index.html>) - This federal-state-Tribe partnership exchanges environmental data. The Exchange Network works to improve data quality, better integrate data across the various sources, and improves availability of data. The data formats are standardized so data can be exchanged across the Internet through the Environmental Data Standards Council. As noted on its website, "Data Exchange Templates (DETs) and schemas, data standards, and data Trading Partner Agreements (TPAs) are also used to ensure data integrity by clearly defining data needs and establishing standards for transmission."
- Connecting for Health (<http://www.connectingforhealth.org/workinggroups/datastandardswg.html>) - This is a public-private partnership to achieve a national network and infrastructure and to create tools to share health information so as to improve patient care and reduce medical errors. Their working group on data standards focuses on identifying common standards and definitions and making them ready for an electronic standards-based model of data transmission and exchange.
- Environmental Public Health Tracking Network (EPHT) (http://www.cdc.gov/phin/05conference/05-11-05/4C_Patridge.pdf) - The Federal Geographic Data Committee (FGDC) has established standards for documenting digital geospatial datasets⁵³ as required by Presidential Executive Order 12906. The EPHT Network facilitates data searches by determining common elements among the data sets and agreeing on standard information to document. They are currently using FGDC standards, an extension of ISO 19115 and they used the freeware tool TKME. As described by Jeff Patridge, EPHT developed requirements for a metadata tool and metadata registry and has promoted the creation and use of metadata among network members.⁵⁴

53 Elements: Dataset title; Contact info; Status; Attributes; Purpose; Citation; Spatial domain; Distribution; Access constraints; Time period of dataset; Keywords, and Metadata reference.

54 Jeff Patridge, "Developing a Metadata Template for CDC, http://www.cdc.gov/phin/05conference/05-11-05/4C_Patridge.pdf

IV. DIFFERENCES IN METADATA NEEDS BETWEEN SURVEYS AND ADMINISTRATIVE RECORDS

Little work has been done to define differences in metadata requirements between surveys and administrative records, a prime source of potential information from federal and local sources and a major interest of NICS. Brand Niemann of the Environmental Protection Agency suggests that one approach may not meet all needs and that we should determine differences in the needs for metadata tools and resources as a 2 x 2 matrix. In Figure 2, we consider some differences.

Figure 2. Differences in Metadata Resources

Source	Surveys	Administrative records (AR)
Federal	Under OMB's Statistical Policy Office, there is a guide to the type of information to include as part of a metadata system, but no guides for machine-readable format. The potential exists for the latter through the FEA DRM as described above.	There are federal-state partnerships to integrate, maintain, and provide access to ARs released for statistical purposes across states (e.g., crime, educational performance, the environment, disease registries and vital statistics, unemployment) with prescribed formal but minimal metadata (e.g., record layouts and edit rules) that do not necessarily meet the standards of federal surveys. Other ARs are not a part of the statistical system and up to now, have not been thought of as potential sources of statistics (although they could be). Thus, the creation, content, and format of such ARs are ad hoc, not maintained, and formal metadata is usually not available to researchers or inadequate.
State, local	There are few state surveys and even fewer local surveys. The few that exist (e.g., Oregon) have limited public metadata. They could use the same standards as federal surveys.	States and local areas have ARs that are not part of a federal-state partnership (e.g., drivers' licenses, public assistance, building permits) but the creation, content, and format of metadata is usually, at best, informal, and idiosyncratic. More often, metadata is not organized nor is it maintained.

V. SUMMARY OF CURRENT SITUATION, IMPLICATIONS FOR NICS, AND INCENTIVES FOR DEVELOPING METADATA

Current Situation and Implications for NICS

It is a basic tenet of NICS that community data is a valuable asset and that all levels of government, private organizations, and businesses, should have access to competent analyses of the data for planning, evaluating, and providing services to their people. As such, the development of statistical metadata is not an esoteric exercise. Rather, statistical metadata is the foundation for the appropriate analyses of data that inform us about, for example, community development, security, the environment, health, equity issues, and economic growth.

One step in meeting the objective of a national infrastructure for community statistics is to develop a metadata infrastructure and governance that provides incentives and tools for participation to owners of state and community data. As we have seen from the review above, there are tools and resources to draw on. Nevertheless, there are significant challenges for NICS in regards to metadata creation and maintenance.

A measure of the success of NICS would be widespread development and implementation of metadata standards and automation tools that facilitate better data sharing across communities. What options are available to NICS now that would start us on that path? What are the entry points? What avenues should NICS support and help to develop? A NICS metadata infrastructure needs to bear in mind, as discussed below, technology, standards, outreach, funding and human resources, as well as a governance mechanism and other activities that lead to the creation, use, maintenance, processing, and distribution of metadata.

1. *Technology* –

Metadata resources and approaches to creating metadata for local communities are available to the NICS program and much of it is publicly available. In Figure 1 above, we identified metadata resources, including software languages, metadata entry tools, standardization and classification tools, search tools, networking tools, repository services, and standards for metadata. While much is available, current technology needs to be adjusted to meet the detailed and extensive information of the statistical system.

There are examples of tools that address some needs, such as the Global Justice XML Data Model ability to coordinate and communicate among systems that were developed differently. The DataFerrett has a search engine for concepts, definitions, and datasets that include specified variables. The federal team working on the Data Reference Model (DRM) is alert to the more detailed information in statistical systems and may eventually provide these systems a convenient means of formatting their metadata for automation so that the information can be shared across systems. NICS should continue to work with DRM to encourage a system that meets the needs of

analysts for detailed information. Others will be able to use what has been developed without direct cost to NICS.

The team from the GovStat project has developed valuable information, including an interactive glossary and interfaces that improve access to various datasets. They have given much effort to plain English definitions for data users with only a basic level of statistical literacy. As these aspects are expanded, they will be useful for the NICS program.

2. Metadata standards for content and automation and policies for revisions –

In Appendix A, we provide critical elements in a statistical metadata system as principles and guidelines to those creating metadata. This is a foundation from which NICS members can draw and a step towards the standardization of information. The statistical system does need, however, to develop standard formats for metadata to automate it and make it operable across systems.

Technology is less of an issue than the fact that statisticians have not discussed conventions for formatting statistical metadata to be machine-readable so that heterogeneous systems can communicate, share, and process information among the systems. Information technology people refer to this as “semantic interoperability.” To have that requires agreement on how to search for information, give it context, and how to characterize it so that information can be combined across sources.

A barrier to communication among those from different fields who need to work together is their respective jargon. Those who are knowledgeable about statistical metadata use jargon that is different from that those in the information technology field use. NICS may be able to help by finding people who can explain concepts in a common language, a bridge towards progress in advancing current tools and systems.

Current automated systems for metadata meet the limited needs of librarians for metadata better than for the detailed metadata needs of statisticians. Not only are statistical metadata physically extensive, the search requirements of analysts are complex. Appendix C shows an example of how analysts use statistical metadata. Below are some of the types of questions for which analysts use metadata:

- What are the options among data sets? Because there are multiple data sets and sources, data users first need to be aware that there are options and then need information to help them to decide which dataset is best for their particular purpose. Metadata that answer questions about the timing of the data set, the geographic areas available, and the subjects and universe available eliminate some datasets from further consideration.
- Which data set has the characteristics that are appropriate to the problem for which the data user is trying to find an answer? What is the purpose of the survey or administrative data set? What is the survey design or the time series? Are the data cross sectional or longitudinal? What is the sample size? For example, the purpose of the Current Population

Survey is to publish monthly employment and unemployment statistics and it has the most detailed questions on that topic. That generally makes it the first choice among surveys about workers – but the sample is only large enough to provide data for the nation, and multi-year data for states and very large metropolitan areas. If the geographic area needed is below the state level, the CPS is not an option. The data user might turn to the decennial census or the American Community Survey.

- Technical definitions of topics, coding rules, and data processing edits clarify differences among data sets. For example, how residency status is defined is a critical factor in whether it is valid for a data user to compare information from different data sets.
- Accuracy of the data tells the data user how far out on the limb they dare go with their analyses and inferences. A data user needs the sample design (survey size), the questionnaire (more detailed questions on the subject yield better measurements), and data collection methodology (who provides the information? How good is the training for interviewers? Is there follow-up for nonresponse? Has there been research on data quality?).

3. Outreach, promotion, and networking about metadata development and maintenance –

NICS is developing a website about metadata that will provide basic education and training for those who know enough to go to that site. Mechanisms for active outreach and networking, however, are critical if we are to have any hope of widespread action.

4. Funding and human resources –

It is one thing for technicians to agree with the need for metadata that operates across systems. It is another for them to have the political backing for an ongoing budget to create and maintain what many politicians would see as the mind-numbing detail and jargon of the numbers and information technology worlds, worlds they may not want to know about let alone include as a budget priority. NICS will be challenged to change that view.

5. Governance and other activities that lead to the creation, use, maintenance, processing, and distribution of metadata -

A national governing structure is needed to make metadata creation a standard practice and part of the mainstream functions and cost of the development of data files. To build a national metadata infrastructure towards which communities can contribute and use requires a clear set of specifications, a detailed action and business plan, and sufficient dollar and people resources to create and maintain the infrastructure. Some of the tasks, but by no means all, can be accomplished through committees and volunteers as coordinated through NICS. NICS can turn for help to existing federal committees and national organizations with mutual strategic goals that already have authority and funding to do some tasks.

NICS can draw on the governance experience of those who have worked for several decades towards a National Spatial Data Infrastructure (NSDI), a national effort to compile geospatial data for local use. As with metadata, compiling geospatial data is not at the top of the

political agenda even though both are foundations for work on key national issues. The NSDI has struggled to find a governance structure so they can go beyond data sharing to building a national infrastructure. They have ample experience that demonstrates the need for governance by “a collaborative leadership structure that reflects the needs and contributions of all parties.”⁵⁵ Likewise, NICS needs to establish a governance structure so that the different levels of government as well as organizations can participate in decisions. It is not simple to have a structure that provides representation to dozens of federal agencies, 50 state governments, about 3,100 counties, more than 18,000 municipalities, and thousands of private organizations and businesses.

Butler and his colleagues identified why some past governance attempts have failed, including lack of commitment, the priority of individual needs over concessions that meet the needs of the majority, lack of authority to share resources, inadequate funding, and resistance to a governance structure when they have been able to act individually in the past. For a similar model, they point to the Federal Highway Administration and the funding of state departments of transportation to build functionally equivalent roads and their sharing of information on appropriate building materials for local conditions.

Incentives for making metadata available

Why should NICS participants prepare, provide, and maintain metadata? After all, metadata creation and maintenance is resource intensive and very detailed work. What is the business case to support it? What incentives are available for shifting the creation of metadata from a costly burden to a benefit? NICS has developed a system of incentives and we list a few below.

- **Save money:** More (but not all) policymakers recognize that data are strategic assets. Standardized metadata and organized, uniform ways of presenting it in a machine-readable format, saves money. If content requirements and format that meet exchange standards are available, the resources a data provider needs to create their own system are reduced. This is what happened when the Department of Justice made GJXDM available. This message will not work for those policymakers who see data as a liability because they lose control of the message they wish to present.
- **Expand uses of data and reduce collection:** Standardization of metadata across datasets makes it easier for data producers to use other datasets in conjunction with their own. This may reduce the data items that must be collected from sources and add value added for analyses.
- **Federal standards are available as a model:** For surveys sponsored by federal statistical agencies, OMB suggests that agencies use voluntary consensus standards (as defined in OMB Circular A-119) for data exchange and metadata management. OMB has proposed a minimum set of specific elements agencies should include in their documentation of statistical surveys. For circumstances where use of a voluntary consensus standard is

55 Al Butler, Alan Voss, Dennis Goreham, and John Moeller, “The National Geospatial Coordinating Council, A Dramatic New Approach to Build the NSDI,” GeoWorld, October 2005, p. 38.

impractical, agencies are asked to follow procedures defined in OMB Circular A-119 for developing, adopting, and/or using the appropriate government-unique standard. Agencies are asked to coordinate with OMB to ensure that any new or modified standards are consistent with guidelines defined in the Federal Enterprise Architecture Data Reference Model for the machine processing of metadata. NICS may wish to coordinate with OMB to facilitate the avoidance of conflicts and help ensure mutual interoperability with any standards that the NICS community of practice develops.

- **Reward** Federal agencies fund the creation of many datasets, some that are surveys and some that are created from administrative records on topics of policy interest to the agency. They could require that machine-readable metadata be supplied as part of the project. There is precedent for this. For example, to receive grants, the Department of Justice requires state and local agencies to conform to the standards of GJXDM if they are using XML. Two issues to consider are how quality can be encouraged if not enforced, and whether metadata creation and maintenance would be a specific budget item.
- **Improve performance and preserve vital information:** As more state administrative records are converted for use as statistical datasets to develop state policy and monitor performance, a requirement to create uniform metadata in a structured way would contribute to the analysis of collective data resources. For example, the FEA DRM provides a framework for agencies to speak the same language about information they need for policy and to create agreements for data exchange and integration. Currently, these data are processed and analyzed primarily through trusted academics based on personal contacts with administrators in state agencies. Rarely, if ever, is metadata created as part of the project. NICS may want to provide outreach and training to states as to how they can gain efficiencies if they document, update, and maintain information about the datasets. This might be accomplished through state budget offices, as they are likely to review funding state priorities and are therefore most likely to understand the costs and benefits of metadata creation. NICS may wish to consider an analysis of social networks around data creation within states to accomplish this goal.⁵⁶

⁵⁶ Noshir Contractor, "Analysis of Social Networks in Digital Government," presentation at Research Symposium of the National Infrastructure for Community Statistics (NICS), Brookings Institution, June 30, 2005.

VI. ARCHITECTURE ISSUES

NICS cannot be effective if it cannot be implemented. In this section we briefly describe issues for an architecture for NICS that statistical offices may follow. As we will argue below, an agreed upon architecture that constituents follow is essential for NICS to be successful.

There is another reason to argue for a common architecture, though, and this is much more important. All the surveys that statistical offices conduct differ. They differ within the office, of course, but they differ across offices as well. Basic concepts, designs, and processing are usually not the same, even for surveys that try to measure the same or similar ideas, such as labor force surveys.

The differences among surveys are in their components and descriptions. These descriptions, as we have said earlier, are statistical metadata. However, the descriptions types, e.g., concept definitions, database designs, or edit procedures, are the same. The types of statistical metadata we use to describe a survey are the same for all surveys, independent of who conducts it.

This is the crucial point. It means we can use a common metadata model for organizing our statistical metadata. This doesn't have to be the case, but there are at least three important reasons to do it:

- Each business unit wishing to manage statistical metadata doesn't have to build its own metadata model. This shortens development time and eases the cost of implementing systems
- If an organizational unit wishes to compare part or all its survey descriptions with that of another unit inside or outside its statistical office, then it knows how these other descriptions will be organized, vastly reducing the time and cost of making a comparison
- If a user wishes to find data, the best source for that data may not be known in advance. Comparing the data produced from candidate sources is eased significantly by using common descriptor types. If, for instance, a variable is described using the same types of components across all surveys, then comparing variables is reduced to comparing their components.

In all the cases above, we reduce the comparison problem to comparing statistical metadata in like components. If there is no common metadata model, then the components themselves won't be the same, let alone the metadata stored in them. In this way, we accept that each survey NICS cares about may be different, but all are described using the same template, model, or schema.

Another component of the architecture is the transfer mechanism for moving statistical data and metadata from one system to another. This requires a large investment in resources to build, especially if each program office within each statistical office builds it differently.

Just as with the case of a metadata model, a data and metadata transfer mechanism that is common across all surveys vastly reduces development time and cost. A single system may not be

the most efficient solution for a specific survey; however, the efficiency for the statistical organization is great. A single way of transferring data from production to dissemination greatly reduces the cost of moving the organization's data to the public.

For a long time, statistical organizations recognized the need for a single Internet source for all their data dissemination. With NICS, organizations could look to a single source for data from many (if not all) sources. Again, the cost of developing a system is reduced because many organizational units (the statistical offices themselves) share the cost.

As we see, for production of statistics, for the transfer of data, and for their dissemination, the costs of systems development are reduced when they are shared. Thus, taking the whole organization or multiple organizations into account when designing systems is cost effective.

Below, we will examine the individual steps described above: production, transfer, and dissemination. The main focus of the discussion will be on the standards that can be used to build the models and systems required.

Why are standards important? There are several reasons for this. Most standards represent a consensus among interested parties who need to solve a problem. As long as the consensus process is open (open to any interested party), transparent (the stage of development of any project can be determined), and subject to due process (there is a well-defined process for appeal, e.g., for patent problems), the standard can be viewed as authoritative. Implicit in this is the assumption that the people involved in the process have the requisite expertise. While there is no way to know this, the parent organizations have a vested interest in seeing that this is so. The truest test of the worth of a standard is its uptake—the fact that there are organizations building systems based on the standard. It shows that other people feel the solutions presented solve business problems.

Given that organizations are implementing standards, why should others? The reasons are fairly simple. Standards undergo a substantial peer review process, more than an academic paper in a refereed journal. The results can be trusted. They may not represent the bleeding edge of research, but most organizations are loath to build systems that way anyway. Secondly, standards save time in development costs. The developers of standards have already thought about how to efficiently solve a problem. Most have already implemented the results themselves. Again, the results can be trusted.

Production

The statistical production process has received much attention in the statistical metadata arena. The work of individual researchers, groups, implementers, and standard organizations are too numerous to mention. A consensus does appear to be in reach, however. The community of practice defined by the participants in the METIS work sessions on statistical metadata has settled on three standards: ISO/IEC 11179, the Neuchâtel Classification and Variables models, and the Data Documentation Initiative (DDI).

ISO/IEC 11179 specifies how to describe the semantics, i.e., the meaning, of data. The standard was not developed for the statistical community, per se, but statistical data are fundamentally data, and thus can be described in a way common across various subject areas.

Each datum is not described separately, but data that share similar semantics, for instance a column of data in a database, are described together through a construct called a data element. The standard divides data elements into two main parts: representational and conceptual.

- The representational part concerns itself with the allowed values, the data type, format, and unit of measure (if necessary) of a data element.
- The conceptual part describes the set of objects (the population) about which data are collected and the characteristic of that population that is measured.

ISO/IEC 11179 is incorporated into both the Neuchâtel combined model and the DDI. The differences between the Neuchâtel models and the DDI are their scopes. Neuchâtel is focused on data, the concepts used in a survey, and how they fit together. DDI describes more of the process, but does not quite go into the detail of Neuchâtel.

DDI is currently undergoing a major revision from being predominantly a specification for describing a data set to a specification for describing the production of data and data sets as well. The Neuchâtel Variables model version 1.0 is being released as this report is written. The model links concepts from survey designs to the variables a survey produces.

In combination these standards represent a thorough description of a survey and its data, including tables. The sample design, questionnaire, database design, processing steps (editing, coding, and imputation), and analyses. The DDI, in particular, has concentrated on a detailed life cycle model of a survey to check that its description is comprehensive. The Neuchâtel models, on the other hand, go into deep detail about data, classifications, and survey concepts. The classification model, for instance, contains a way to manage concordances between versions of classifications to support time series comparability.

Transfer

There are many standards available for transferring data in statistical offices. These include XML, ASN.1 (Abstract Syntax Notation), and others. ASN.1 has lost favor to XML in recent years. For one thing, XML is easier to understand and implement. In addition, all recommendations coming from W3C gain much immediate attention when they are released, in part due to the success of XML.

Specifications for transfer that are for specific subject matter have been built on XML. Many are of interest to statistical agencies, especially SDMX. The purpose in developing SMDX was to create a standard mechanism to transfer statistical data from national statistical offices to the

international offices that helped to fund⁵⁷ its development. The project has evolved to the point that an initial version of SDMX is now an international standard: ISO 17369. It is a mechanism for transfer of statistical data and metadata using one common framework.

The framework for SDMX is based on shared models called data (or metadata) structure definitions. These are defined once and used over and over. They can be shared across agencies. Several efforts are underway to build the structure definitions for many kinds of survey data and metadata.

SDMX is based on XML and the electronic data reporting standard GESMES (General Statistical Messages). Several agencies are using it, including the seven developers, the Federal Reserve Bank, and the Food and Agriculture Organization. Even though other specifications and home-grown solutions may be effective, SDMX represents an efficient standard that all statistical offices can follow. It was built with the needs of statistical offices in mind. In addition, updated versions of the standards are planned, and statistical offices can submit desired changes. For now, SDMX is the closest the statistical world has for a standard transfer mechanism for data and metadata.

Dissemination

Many statistical data dissemination systems have been built and are in production. Most of them were built with a single survey in mind, or if they happen to manage data from several surveys, there is no link among them. The NESSTAR system, designed for use with the DDI, is an exception. It uses the DDI as its metadata scheme, and that scheme is, of course, common across all surveys and data sets. The company formed to build, update, and promote the software no longer exists, however, and the future of the system is in doubt.

NICS needs a system like NESSTAR. The standards discussed earlier for production and transfer take into account the concepts used in surveys and data. These concepts need to be made available, and are brought to the data dissemination system through metadata transfer. Concepts are required to make the comparisons and harmonization across data sets, and that is the promise of NICS. NESSTAR is designed to handle them thanks to the DDI, the underlying metadata model.

This is not to say the current version of NESSTAR is sufficient. It is not. Much work is required before a system could be made to work. Statistical agencies must work together to build the appropriate system. Building systems for one survey at a time most likely will not work.

Acknowledgements: The author thanks Andrew Reamer, Katherine Wallman, Susan Schechter, Brian Harris-Kojetin, and Brand Niemann for their comments.

57 SDMX came out of a project funded by International Monetary Fund, World Bank, Bank of International Settlements, European Central Bank, Eurostat, Organization for Economic Cooperation and Development, and the UN Statistics Division.

APPENDIX A

CRITICAL ELEMENTS IN A SYSTEM OF STATISTICAL METADATA

The critical elements below are excerpted from the ideal list of elements in a statistical metadata system as provided above. The critical items are the pieces of information any analyst of statistical data needs to make informed decisions about the appropriateness of using that data set to answer particular questions.

1. Characteristics of the Data

1.1 Overview of the data set

- 1.1.1 **Historical background:** survey name, organizational sponsor(s) of a survey or administrative data set, organization name(s) that conducted data collection.
- 1.1.2 **Objectives** - purposes for which information is required, stated within the context of the program or research problem that gave rise to the need for information; how the information is used.
- 1.1.3 **Uses** - decisions to be made based on collected information and how information will support decisions.
- 1.1.4 **Users** - organizations, agencies, and groups expected to use the information.
- 1.1.5 **Type of Respondent**, such as housing units, persons (self/proxy), or establishments.
- 1.1.6 **Model and its assumptions** if the data are estimates or projections.
- 1.1.7 **Data release version and type** – whether preliminary or final, and whether this is a pilot study with a small number of cases or restricted geographic area.

1.2 Guidelines and the process for collecting and processing the data

- 1.2.1 ****Forms or questionnaires**
- 1.2.2 **Rules for data entry** - procedures, and training given to person entering data on the form (e.g., manuals for interview rules)
- 1.2.3 **Data capture** - Method of data capture, accuracy rate, quality control measures
- 1.2.4 **Keying/scanning specs**

1.3 Population Universe, Population Coverage

- 1.3.1 ****Define the target population** - all the people, establishments, or other units in the data set
 - 1.3.1.1 If administrative records, define the program participation rules and the means of collecting the data (program information provided by a respondent? through interviews with a case manager? Is information keyed and are there any quality control measures?)
 - 1.3.1.2 If a survey, describe the sampling frame used to identify this population.
 - 1.3.1.3 If applicable, information on eligibility criteria and screening procedures.

- 1.3.2 **Description of the survey design**, including the:
 - 1.3.2.3 ****Sampling frame** (i.e., the sources of information such as lists, directories, and records, that cover the universe and information about any exclusions),
 - 1.3.2.4 ****Size of the sample and the rules for selection from the universe and determination of the size**
- 1.3.3 ****Residence rules**

1.4 **Time Frame of data set(s)

- 1.4.1 **Time coverage and frequency** of availability of the data set.

1.4.2 **Variations in timing** - what is known about cyclical, seasonal, or other variations over time in the data set.

1.5 **Reference period of questions**

1.6 **Information for Using the Data**

1.6.1 **Wording of questions** or information on the form of administrative records

1.6.2 **List of data elements, the range of their possible values, and their definitions** and, for the search function, their plain-English synonyms; and any changes in the definitions over time (e.g., race and ethnicity).

1.6.6 **Variance estimates** - Explanation of how to calculate estimates of variances that are specific to the survey

1.6.7 **Record layout**, that is, the description of the data elements on the file and their physical location

1.6.8 **Code lists** used, including classification schemes for variables (e.g., the North American Industry Classification System versus Standard Industrial Classification), and recoding rules

1.6.9 **Top coded values**, if any

1.6.11 **Contact** for questions – names, telephone numbers, and email addresses.

1.6.12 **Errata and Notes**, including geography and data corrections

1.7 **Geographic scope**

1.7.1 **Geographic areas included** in data set (specific areas present in the data set)

1.7.2 **Definition of geographic components and hierarchy**

1.7.3 **History of changes in geographic boundaries** and how handled

1.7.4 **Maps** of geographic boundaries (outlines of areas)

2. **Quality of the Data**

2.1 **Data Limitations**

2.1.1 **Statistical precision** of survey results, at least for the major estimates. This could include estimates of sampling variances, standard errors, or coefficients of variation, or presentation of confidence intervals.

2.1.2 **Nonsampling errors** - For both administrative and survey data, provide reporting errors, response variance, interviewer and respondent bias, and errors in processing the data that may affect the data, any measures of bias,⁵⁸ and methods to deal with such problems.

2.1.3 **Edit and imputation rules** such as for nonresponse to an item and how nonresponse is handled in the database (e.g., left blank? edited? If edited, what are the edit rules for using available information and assumptions to substitute values in the data set?).

2.1.5 **Weighting scheme** for survey data, including adjustments for nonresponse and benchmarking and how to apply them.

3. **Dissemination of the Data**

3.4 **Additional documentation for Public Use Microdata Sets**

Describes construction of the information and how to access and manipulate the data.

5. **Training and Assistance**

⁵⁸ Bias is defined as the deviation of the average survey value from the true population value.

5.2 ****Contact for further information and assistance** — specifics of who and how.

APPENDIX B
GLOSSARY OF TERMS RELATED TO STATISTICAL METADATA
Compiled by Cynthia Taeuber and Laura Smith

For more definitions of statistical terms, see the Glossary in:

http://www.whitehouse.gov/omb/infoereg/proposed_standards_for_statistical_surveys.pdf . Some of the definitions below were taken from this source. Other definitions are based on discussions with experts in the field including Brand Niemann (Environmental Protection Agency) and Andrew Reamer (Brookings Institution).

Bias - the deviation of the average survey value from the true population value. Bias refers to systematic errors that affect any sample taken under a specific design with the same constant error.

Coding – refers to converting text to numbers or other symbols that can be counted or tabulated in machine processing.

Confidentiality – involves techniques to protect data about individuals from disclosure.

Coverage – the extent to which a survey’s list from which it draws a sample (“the sample frame”) lists all members of the target population once. “**Coverage error**” is the discrepancy between the frame and the actual population included in the survey.

Cross-sectional sample survey is based on a representative sample of respondents drawn from a population at one point in time.

Editing and Imputation – techniques that use available information and some assumptions to derive substitute values for inconsistent or missing values in a data file.

Enterprise Architecture – “An organization’s framework of technology hardware, software, and related policies” from www.ffiec.gov/ffiecinfobase/html_pages/gl_01a.html

Estimates - a numerical value for a population based on information collected from a survey and/or other sources.

Friend of a Friend (FOAF) – a machine-readable modeling of social networks based on RDF.

Interoperability Framework – provides organizational, semantic, and technical standards and principles for heterogeneous systems so that it is possible to communicate and share and process information among the systems. Semantic interoperability, for example, requires agreement on how to search for information, give it context, and how to characterize it so that information can, for example, be combined across sources (see: <http://xml.coverpages.org/ni2004-12-06-a.html>).

Longitudinal survey – follows a representative sample of a population over time and involves repeated measurements of characteristics.

Measurement error - the difference between observed values of a variable recorded under similar conditions and some fixed true value (e.g., errors in reporting, reading, calculating, or recording a numerical value). **Response bias** is the deviation of the survey estimate from the true population

value that is due to measurement error from the data collection. Potential sources of response bias include the respondent, the instrument, and the interviewer.

Model – formal assumptions and mathematical relationships that generates a set of observations. A **metamodel** provides standard rules for building models so that data from different sources can be aggregated.

Nonresponse error - the overall error in estimates caused by differences between respondents and those who do not respond to a survey. Nonresponse error consists of both sampling variability and nonresponse bias (that is, when the observed value from a survey deviates from “truth” about a population because respondents differ in important ways from those who do not respond to the survey).

Ontologies – the term is used by computer information specialists in several ways, but generally, it refers to a formal structure of knowledge for machine processing, that is, reusable libraries of terms, their definitions, and related concepts

Precision of a survey – is a measure of the difference between a sample result and the result if a complete census had been taken under the same conditions.

Public Use Microdata File (PUMS) - includes the detailed responses for a sample of individual respondents from a complete data collection. PUMS files use various techniques, such as aggregation, limited geographic detail, elimination of unique identifiers, and coding, to avoid disclosure of information about individuals.

RDF – Resource Description Framework (<http://www.w3.org/RDF/>). RDF is a method or convention for formatting metadata for the web so it can be merged with the metadata associated with other datasets. It is an application of XML that allows coding, exchange, and sharing (“reuse”) of structured metadata across applications. In RDF, information is a set of statements, each with a subject, verb, and object, and everything is identified with a Uniform Resource Identifier (URI). RDF is a way to keep track of, to integrate, heterogeneous information from various sources.

Sampling Error – the error that occurs because not everyone who should have been in the sampling frame was interviewed as part of a survey. It is the error associated with the variation in samples drawn from the same frame population. The variance equals the square of the sampling error.

Schema – (1) the rules for encoding information; and (2) a model of the relationships among categories in a data base. For example, see Slide 59:
http://www.olsug.org/Presentations/May_2005/Workshops/RDF_Workshop05.pdf

Semantic Web – a machine-readable format that is compatible with the Web and that adds definition “tags” to information that “enables computers to discover data more effectively and allows new associations to form between pieces of information.” See: Susie Stephens,
http://www.olsug.org/Presentations/May_2005/Workshops/RDF_Workshop05.pdf

Target population - any group of potential sample units

Taxonomy –classification of a body of information that includes definitions and clarification of the relationships among the parts. When associations among the categories are defined, it is possible

to automate techniques such as queries and inferences. The categories may be a collection of heterogeneous items that have some relationship to each other, or a class of items with homogeneous attributes (e.g., Persons; housing units).

Unit nonresponse - occurs when a respondent fails to respond to all required response items (i.e., fails to fill out or return a data collection instrument).

Universe - data covering all known units in a population (i.e., a census).

Weights - relative values associated with each sample unit that are intended to correct for unequal probabilities of selection for each unit due to sample design. Weights most frequently represent the relative portion of the population that the unit represents. Weights may be adjusted for nonresponse.

APPENDIX C

USING STATISTICAL METADATA TO DECIDE WHAT DATA SOURCE TO USE FOR HOUSING VACANCY

Example prepared by Cynthia M. Taeuber⁵⁹ and Susan P. Love⁶⁰

Problem: The Housing Vacancy Survey (HVS) provides a vacancy rate as does the American Community Survey (ACS). Is there any meaningful difference in the two rates?

There are more sources for a vacancy rate than the HVS and ACS, but for this example, we will use these two. If you look at the data for a particular year and geography, you will find that the estimates differ significantly. Why?

The answers are in the documentation (metadata) for each survey. The exercise points to the need for a standard format for metadata and the value of automated search capability. Through trial and error, we find the information below. Presentation and detail of the information is not standard between the two surveys even though they are both released by one agency, the Census Bureau.

- **Definition of "vacancy"** --I look at the questionnaires and a "fact sheet" about differences between the ACS, HVS, and CPS
 - ACS--<http://www.census.gov/acs/www/SBasics/SQuest/SQuest1.htm>;
http://www.census.gov/acs/www/Downloads/2004/usedata/Subject_Definitions.pdf
 - HVS--<http://www.census.gov/hhes/www/housing/hvs/qtr205/q205def.html>; and the fact sheet--<http://www.census.gov/hhes/www/housing/homeownershipfactsheet.html>
- **Survey purpose** -
 - ACS--<http://www.census.gov/acs/www/SBasics/What/What1.htm>
 - HVS--<http://www.census.gov/hhes/www/housing/hvs/overview.html>
- **Sample Size** --
 - ACS--<http://www.census.gov/acs/www/SBasics/SSizes/SSizes03.htm>
 - HVS--<http://www.census.gov/hhes/www/housing/hvs/faq.html> and the fact sheet --
<http://www.census.gov/hhes/www/housing/homeownershipfactsheet.html>
- **How the data are collected** --
 - ACS--<http://www.census.gov/acs/www/SBasics/DataColl.htm>
 - HVS--<http://www.census.gov/hhes/www/housing/hvs/datacollection.html>
- **Residency Status**
 - ACS--<http://www.census.gov/acs/www/AdvMeth/CollProc/CollProc1.htm>
 - HVS--<http://www.census.gov/hhes/www/housing/homeownershipfactsheet.html>

Can the housing vacancy statistics from the American Community Survey (ACS) replace the quarterly reports from the Housing Vacancy Survey (HVS) on residential vacancies and

59 Contact: cmtaeuber@direcway.com

60 U.S. Census Bureau, susan.p.love@census.gov

homeownership, the source that has been used for the past 50 years? Based on the metadata, the short answer is no -- estimates about vacancies from both surveys are needed and one cannot replace the other. Even though the information about vacant units seems, at first glance, to be similar between the two surveys, the estimates for rental and homeowner vacancy rates, and estimates of tenure and vacancy status differ substantially for good reasons.

Here are main points we learn from the documentation:

- The rental vacancy rate for the nation from the HVS has been an economic indicator for five decades and is used widely in the federal statistical system and by the housing statistics user community. The Bureau of Economic Analysis, for example, depends on the HVS for the rental vacancy rate and additional measures from the HVS to prepare quarterly and annual estimates of the housing services component of personal consumption expenditures in gross domestic product and the rental income component of national income.
- The data collected for vacant units, the interviewing time frame, and the estimation methods all differ between the HVS and the ACS.
- The ACS is known to produce a depressed vacancy rate because of its data collection design. The ACS does not classify vacant units as year-round versus seasonal units but applies the simplified decennial census definitions (<http://www.census.gov/acs/www/UseData/Def/Vacancy.htm>). The HVS, but not the ACS, collects information on duration of vacancy. The survey designs differ. HVS collects data in one week by personal interview so that vacancies are identified immediately. The ACS collects data over three months in three stages: by mail, telephone, and in the third month, by a personal visit to a subsample of one-third of units that did not respond by mail or telephone. The first two stages collect data only from occupied units. It is not until the last stage of a personal visit that units can be identified as vacant. That has a direct impact on the quality (sampling and nonsampling) of vacancy estimates and results in an underestimate of vacancy. ACS sample units that change status from vacant to occupied during the collection period have a greater chance of being incorrectly categorized as “occupied” by the survey than do units that change from occupied to vacant. Under the ACS design, sample addresses that are mailed to in March, for example, may not be interviewed until May, and over this three- month period, the occupancy status of a sample unit can change. Because the first two data collection stages are successful only in collecting data for occupied units, nearly all vacant units are not identified until the third month of collection, giving units that are vacant the opportunity over two months to change status and become occupied. This inequality produces a downward bias in the vacancy rates produced by the ACS methods, increasing the mean square error on the survey’s vacancy data.

Filename: 20070423_Metadata.doc
Directory: C:\Documents and Settings\sjolda\Local
Settings\Temporary Internet Files\OLK89
Template: C:\program files\microsoft office\templates\Normal.dot
Title: Reconnaissance of Data Confidentiality Tools
Subject:
Author: Forrest McIlwain
Keywords:
Comments:
Creation Date: 4/23/2007 5:12:00 PM
Change Number: 2
Last Saved On: 4/23/2007 5:12:00 PM
Last Saved By: Laura Smith
Total Editing Time: 34 Minutes
Last Printed On: 4/26/2007 4:18:00 PM
As of Last Complete Printing
Number of Pages: 53
Number of Words: 17,138 (approx.)
Number of Characters: 97,690 (approx.)