Institute of Education Sciences
National Center for Education Statistics

NATIONAL INSTITUTE OF STATISTICAL SCIENCES
EXPERT PANEL REPORT

METADATA AND PARADATA:
INFORMATION COLLECTION AND POTENTIAL INITIATIVES

TABLE OF CONTENTS

METADATA AND PARADATA: INFORMATION COLLECTION AND POTENTIAL INITIATIVES

EXECUTIVE SUMMARY

Metadata and/or paradata accompany federal statistical agency data files to describe or define the data elements and the collection and processing of these data. Practices vary across the Statistical Community of Practice (SCOP). Distinctions between the two terms are not well-defined, and there are no generally accepted standards in use by all the federal statistical agencies. Therefore SCOP commissioned the survey of current definitions and practices in current use that make up this report with the objective of laying the groundwork for development of standardized definitions and practices.

The goals for this report are therefore to provide the necessary background information for developing and implementing standardized definitions and best practices for metadata and paradata within the US federal system. In particular, this report

1. Reviews existing practices and resources both nationally and internationally;
2. Inventories current metadata and paradata practices in federal statistical agencies;
3. Identifies key elements of metadata and paradata that would be useful to federal statistical agencies in development and implementation of standardized definitions.

The following definitions are proposed to SCOP:

- *Metadata*: Formalized data about statistical data needed to search for, display and analyze those data.
- *Paradata*: Formalized data on methodologies, processes and quality associated with the production and assembly of statistical data.

Note: *survey weights should be regarded as data*, even though calculation of them employs paradata about the design and conduct of the survey.

- *Markup Language:* a method for annotating text in a way that is syntactically distinguishable from that text and in consequence is computer processable, (e.g., HTML, HyperText Markup Language).

To date with respect to metadata and paradata, there are essentially two systems on which to build. Each is tied to a description language. DDI is based on XML (Extensible Markup Language); SCMS is based on UML (Unified Modeling Language).

A five-step process is suggested for the development and implementation of ICSP (Interagency Committee on Statistical Policy) agency-wide definitions and practices.

1. Agreement on definitions for data, metadata and paradata with *a concept of metadata that is completely independent of that of file structure.*
2. *Commitment to markup language-based metadata, and to markup language-based data*.

3. “Proof-of-concept” based on one or both of DDI and SDMX, consisting *of full markup-language-based metadata and data files for a modest-scale survey* including development of parsers that would put the data into common formats including development of parsers that would put the data into common formats.
4. Creation of an ICSP-wide *extensible metadata template for surveys*, together with tools for common tasks.
5. Creation of a *repository for metadata for surveys*.

PREFACE

The Federal Statistical Community of Practice and the National Center for Education Statistics (SCOP) requested the National Institute of Statistical Sciences (NISS) to survey existing standards and practices for metadata and for paradata that are currently in use for federal data. The goals for this report are therefore to:

1. Review existing practices and resources both nationally and internationally to help inform efforts to develop and implement standardized metadata and paradata for use in the US statistical system.
2. Inventory current metadata and paradata practices in federal statistical agencies to identify possible best practices and possible gaps.
3. Identify key elements of metadata and paradata that would be useful to federal statistical agencies, moving next to the development of standardized definitions, and ultimately to the implementation of those definitions.

NATIONAL INSTITUTE OF STATISTICAL SCIENCES
REPORT TO FEDERAL STATISTICAL COMMUNITY OF PRACTICE

METADATA AND PARADATA:
INFORMATION, COLLECTION AND POTENTIAL INITIATIVES

BACKGROUND

I. TERMINOLOGY

Metadata and/or paradata accompany federal statistical agency data files to describe or define the data elements and the collection and processing of these data. Practices vary among agencies (and other holders of large data files). Distinctions between the two terms are not well-defined, and there are no generally accepted standards in use by all the federal statistical agencies. Therefore the current report surveys definitions and practices in current use with the objective of laying the groundwork for development of standardized definitions and practices.

The terms metadata and paradata are sometimes used synonymously, but we believe that there is a meaningful, albeit nebulous, distinction. The SDMX Metadata Common Vocabulary (see Section 3.2.2) contains the following definitions:

- *Metadata*: data that defines [sic] and describes [sic] other data and processes
- *Statistical Metadata*: Data about statistical data, comprising data and other documentation that describe objects in a formalized manner. They provide information on data and about processes of producing and using data. Statistical metadata describe statistical data and - to some extent - processes and tools involved in the production and usage of statistical data.

There is no corresponding definition of paradata, but the same source further states that "there is a clear high-level distinction between the metadata needed to search for and display data (Structural metadata) and the metadata that give more information on definitions, methodologies, processes and quality (Reference metadata)."

Some people would define paradata as a subset of what SMDX calls metadata: "information about […] processes and tools involved in the production and usage of statistical data."

It is not clear that a distinction between metadata and paradata will persist into the future, but it is useful to maintain the distinction in the short run, and in particular to understand the current situation. One justification is that although in some settings a distinction between metadata and paradata is not useful, the distinction *is* useful in the context of surveys.

To some extent, similar issues arise in distinguishing data from metadata and paradata, which are discussed further below.

Therefore, we propose the following definitions for SCOP:

- *Metadata*: Formalized data about statistical data needed to search for, display and analyze those data.
- *Paradata*: Formalized data on methodologies, processes and quality associated with the production and assembly of statistical data.

We believe that *survey weights should regarded as data*, even though calculation of them employs paradata about the design and conduct of the survey.

To illustrate (see also Section 8), for a single-wave survey, and ignoring statistical disclosure limitation,

1. The following seem unambiguously to be data:
    a. Values of all variables, including frame variables, stratum identifiers and respondent-provided values, record-level flags indicating that a value is missing, edited or imputed
    b. Unit- and variable-level response rates
    c. Data collection mode, if it varies across records
    d. Weights
2. The following are unambiguously metadata:
    a. Descriptions of the purpose of the survey, population, frame
    b. Design information: sample size, sampling mechanism, PSUs, clustering and stratification
    c. Variable definitions, including measurement units
    d. Global variable classifications: frame, collected, subject to editing, subject to imputation or synthesis, derived from other variables
    e. Survey instruments
    f. Reports of nonresponse bias analyses
    g. Information about data provenance[1]
3. The following are unambiguously paradata, because they are about process:
    a. Operational details of data collection: mode, timing, protocols for handling refusals
    b. Information about interviewers
    c. Cost

There are, however, borderline cases. One example is unit-level information about difficulty of obtaining a response (for instance, the number of follow-ups required). Global, as opposed to record-level information about data quality is probably most usefully seen as metadata.

Information about SDL is problematic in several senses. High-level information about SDL seems properly to be metadata. For example, a variable can be classified in metadata as having been coarsened. By contrast, public metadata might not contain details about data swapping, such as the swap rate and which variables were swapped. To introduce a distinction between confidential and public-use metadata may not be a

[1] Data provenance is a major concern in the digital archiving community, especially insofar as it affects and is informative about data quality. The usage is similar but not identical to that in connection with works of art. A useful introduction is available at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.101.7132&rep=rep1&type=pdf

useful course of action at this time. An intermediate position would be to permit an original dataset and a confidentiality-protected version to have different metadata.

It may also be useful to think of metadata as logically distinct from the data collection process and collected data, in the sense that metadata are definable in advance of any data collection. This would imply that only data and paradata can contain record-level values, and that metadata cannot. . In terms of the discussion above, this would require that reports of nonresponse bias analysis be regarded as either data or paradata.

An essential break with the past is that *metadata does not mean file structure*. Indeed, the markup language-based current thinking discussed in Section 2.2 and the appendices forces metadata to be logically and operationally independent of each other. The same reasoning applies in principle to the terms "codebook" and "data dictionary." In practice, objects identified as one of these seem to closer to being true metadata than an unannotated descriptions of data objects and data products.

1.1 Markup Languages

Nearly all of the remainder of this document is related in some way to *markup* (equivalently, *description*) languages.

The concept of a markup language is straightforward: it is a method for annotating text (or other objects) in a way that is syntactically distinguishable from that text. The most important implication of "syntactic distinguishability" is "computer processability." The most ubiquitous such language is HTML (HyperText Markup Language),[2] which is the method by which information is transmitted in the World Wide Web. The use of a markup language allows the same content to be displayed by multiple web browsers, as well as be imported into other software such as word processors. Another example is the text processing language LaTeX (and underlying TeX engine).

Appendix A of this document contains a metadata-centric example, but it is useful to illustrate with a simpler example. Consider the address

Alan F. Karr
National Institute of Statistical Sciences
19 T. W. Alexander Drive
P.O. Box 14006
Research Triangle Park, NC 27709-4006

A markup language version of this text would be

```
<ADDRESS>
    <NAME>
        <FIRST_NAME>Alan</FIRST_NAME>
        <MIDDLE_NAME>F.</MIDDLE_NAME>
        <SURNAME>Karr</SURNAME>

    </NAME>

    <STREET_ADDRESS>
```

[2] See http://en.wikipedia.org/wiki/HTML

```
            <NUMBER>19</NUMBER>
            <STREET_NAME>T. W. Alexander </STREET_NAME>
            <STREET_SUFFIX>Drive</STREET_SUFFIX>
        </STREET_ADDRESS>
        <BOX_ADDRESS>
            <PREFIX>P.O. Box</PREFIX>
            <NUMBER>14006</NUMBER>
        </BOX_ADDRESS>
        <CITY>Research Triangle Park</CITY>
        <STATE>
            <NAME>North Carolina</NAME>
            <USPS_ABBREVIATION>NC</USPS_ABBREVIATION>
        </STATE>
        <ZIP>
            <5DIGIT>27709</5DIGIT>
            <PLUS4>4006</PLUS4>
        <ZIP>
</ADDRESS>
```

The essential characteristics of this representation are:

Tags of the form <XXX> …</XXX>. The / is the indication of the end of the tag. Syntactic separation of annotation - tag pairs - from the content between them.

The hierarchical structure: tags can be nested within one another, but a tag must end before any parent tag containing it ends.

A parser is software capable of reading a markup language and resolving the annotation and content. Web browsers are one class of examples. A parser capable of reading the markup language version of the address would be able to arbitrarily combine or reorder the content elements.

By comparison, LaTeX uses the basic form

```
\begin{tag}Content\end{tag}
```

For example, text to be italicized appears in the source (markup) document as

```
\begin{italic}This sentence is in italics.\end{italic}
```

and in the parser-processed output document as

This sentence is in italics.

II. REVIEW OF METADATA RESOURCES

2.1 UNECE

The most vigorous official statistics activities appear to those associated with the United Nations Economic Commission for Europe (UNECE). The base URL for this information is

http://www1.unece.org/stat/platform/display/metis/The+Common+Metadata+Framework

Other valuable general resources are:

- Part A: http://www1.unece.org/stat/platform/download/attachments/8683564/CMF_Part_A_final+for+web.pdf?version=1
- Part B: Metadata Concepts, Standards, Models and Registries, which has a wealth of information: http://www1.unece.org/stat/platform/display/metis/Part+B+-+Metadata+Concepts%2C+Standards%2C+Models+and+Registries
- Terminology: http://www.unece.org/stats/publications/53metadaterminology.pdf

In particular, the UNECE identifies what appear to be the (only) four sets of technical standards for metadata:

- The Data Documentation Initiative (DDI) (Section 3.2.1)
- The Statistical Data and Metadata eXchange (SDMX) (Section 3.2.2).
- The Dublin Core (Section 3.2.3)
- ISO Standard 11179 (Section 3.2.4)

The first two of the appear to represent the "state of the art."

2.2 Metadata Standards

Most standard of these are tied to a description language such as XML (Extensible Markup Language)[3] or UML (Unified Modeling Language).[4] While the specific language may change, the fundamental purposes, which are to:

- Facilitate machine readability of data and information;
- Support interchange of information and creation of repositories;
- Separate information content from electronic (or physical) instantiation of the content; seem less likely to change.

The four initiatives listed here seem quite clearly to be the principal ones in the world.

2.2.1 Data Documentation Initiative

Base URL: http://www.ddialliance.org/

This effort is rooted in the social sciences. The principal product is the DDI 3.1 Specification, which is a very large XML schema.

- Full File: E:\NISS\ESSI-Stat\SCOP\Metadata\DDI_3_1_2009-10- 18_Documentation_XMLSchema.zip
- Documentation: E:\NISS\ESSI-Stat\SCOP\Metadata\DDI_3_1_2009-10-18_Documentation_XMLSchema\Documentation\DDI_3.1_Part_I_Overview.pdf

2.2.2 Statistical Data and Metadata Exchange (SDMX)

SMDX has its roots in financial institutions, and is based on UML.

Base URL: http://sdmx.org/

[3] See http://en.wikipedia.org/wiki/XML.
[4] See http://en.wikipedia.org/wiki/Unified_Modeling_Language.

"Focusing on time series and indicators, SDMX is the result of a joined effort from the Bank for International Settlements, the European Central Bank (ECB), EUROSTAT, the International Monetary Fund (IMF), the Organization for Economic Cooperation and Development (OECD), the United Nations (UN), and the World Bank (WB) to create an XML specification to support the exchange of aggregate data and metadata. SDMX provides three types of statistical metadata standards: standards for data formats, standards for metadata and a registry-based architecture to implement these standards and to exchange data between systems.

One of the requirements of SDMX was the awareness of other metadata specifications such as the Data Documentation Initiative (DDI). Any of the DDI metadata - which emphasizes archival metadata and micro-data, rather than aggregate data - is exchangeable in an equivalent SDMX metadata format. This ensures inter-operability of metadata across namespaces."

The most accessible description is the user guide: http://sdmx.org/wp- content/uploads/2009/02/sdmx-userguide-version2009-1-71.pdf.

Standards: SDMX Standards Version 2.0 Complete Package: http://sdmx.org/?page_id=16#package

"SDMX Technical Standards Version 2.0 provide the technical specifications for the exchange of data and metadata based on a common information model. The scope of this effort is to define formats for the exchange of aggregated statistical data and the metadata needed to understand how the data is structured. The *major focus is on data presented as time series, although cross- sectional XML formats are also supported*.

Version 2.0 Technical Standards are backward compatible with the earlier Version 1.0 efforts, which focused on XML- and EDIFACT-syntax data formats. The latest work broadens the technical framework to support wider coverage of metadata exchange as well as a more fully articulated architecture for data and metadata exchange.

These specifications have been developed, reviewed, and adopted by SDMX. Steps will be taken to bring this work forward within the context of the International Standards Organization (ISO), with a view to updating ISO/Technical Specification 17369:2005 SDMX."

Vocabulary: http://sdmx.org/wp content/uploads/2009/01/04_sdmx_cog_annex_4_mcv_2009.pdf

2.2.3 Dublin Core

The Dublin Core Metadata Initiative (DCMI) is an open organization engaged in the development of interoperable metadata standards that support a broad range of purposes and business models. The format is XML/RDF. RDF is the Resource Description Framework.[5]

Base URL: http://dublincore.org/

User Guide: http://dublincore.org/documents/usageguide/

2.2.4 ISO/IEC 11179

This standard focuses on metadata registries.

According the web site, "The 11179 standard is a multipart standard that includes the following parts:

- Part 1: Framework

[5] See http://www.w3.org/RDF/

- Part 2: Classification
- Part 3: Registry metamodel and basic attributes
- Part 4: Formulation of data definition
- Part 5: Naming and identification principles
- Part 6: Registration"

Base URL: http://metadata-stds.org/11179/

Quoting from http://en.wikipedia.org/wiki/ISO/IEC_11179, "The ISO/IEC 11179 model is a result of two principles of semantic theory, combined with basic principles of data modelling. The first principle from semantic theory is the thesaurus type relation between wider and more narrow (or specific) concepts, i.e. the wide concept "income" has a relation to the more narrow concept "net income". The second principle from semantic theory is the relation between a concept and its representation, i.e. "buy" and "purchase" are the same concept even if different terms are used."

Another ISO standard, ISO 23081-1:2006 (http://www.iso.org/iso/catalogue_detail.htm?csnumber=40832), "covers the principles that underpin and govern records management metadata. These principles apply through time to:

- records and their metadata;
- all processes that affect them;
- any system in which they reside;
- any organization that is responsible for their management." This standard seems to be oriented primarily to database management.

2.3 Other Resources

2.3.1 Open Data Foundation

The Open Data Foundation supports tools to:

- Discover the existence of data
- Access the data for research and analysis
- Find detailed information describing the data and its production processes
- Access the data sources and collection instruments from which and with which the data was collected, compiled, and aggregated
- Effectively communicate with the agencies involved in the production, storage, distribution of the data
- Share knowledge with other users

Base URL: http://www.opendatafoundation.org/

Additional documents of interest:

- http://www.opendatafoundation.org/?lvl1=resources&lvl2=papers
- https://docs.google.com/viewer?url=http://www.ratswd.de/download/workingpapers2009/57_09.pdf
- http://www.opendatafoundation.org/papers/66-258-2-PB.pdf

2.3.2 International Household Survey Network

Base URL: http://www.surveynetwork.org/home/

"The mission of the IHSN is to foster the improvement of the availability, accessibility and quality of survey data in developing countries, and to encourage their analysis and use by national and international development decision makers, the research community, and other stakeholders.

Intermediate objectives to achieve these goals are:

- Improved coordination of internationally-sponsored survey programs with emphasis on timing, sequencing, frequency and cost-effectiveness;
- Provision of harmonized technical and methodological guidelines by the international community, in particular related to data collection instruments;
- Availability of a central survey data catalog to better inform users of the availability of survey and census data from all sources;
- Provision of tools and guidelines to data producers, to foster documentation, dissemination and preservation of microdata according to international standards and best practices."

The IHSN also provides tools for "the data documentation, or *metadata*, [that] helps the researcher to:"

- **Find** the data they are interested in. Without names, abstracts, keywords and other important metadata element it might be difficult for a researcher to locate the datasets and variables that meet his or her research requirements. Any cataloguing and resource location system - be it manual or digital - is based on metadata.
- **Understand** what the data are measuring and how the data have been created. Without proper descriptions of the design of the survey and the methods used when collecting and processing the data, the risk is high that the user will misunderstand and even misuse them.
- **Assess** the quality of the data. Information about the data collection standards, as well as any deviations from the planned standards, is important knowledge for any researcher who wants to know whether the data are useful for his or her research project."

The associated URL is http://www.surveynetwork.org/home/index.php?q=tools/documentation/standards

The IHSN web site also cites DDI, SDMX, Dublin Core and ISO 11179.

III. SPECIFIC EFFORTS IN OTHER COUNTRIES

Comparable but more detailed information for the fourteen US agencies that are members of the Interagency Council on Statistical Policy appears in Section 6. The countries discussed here appear to be neither notably ahead of nor notably behind the US>

3.1 United Kingdom

1. e-Government Metadata Standard (E-GMS) is an XML schema that defines the metadata elements for information resources to ensure maximum consistency of metadata across public sector organizations in the UK.

Base URL: http://www.cabinetoffice.gov.uk/govtalk.aspx

Library: http://www.cabinetoffice.gov.uk/govtalk/schemasstandards/xmlschemas/schemalibrary.aspx

2. eCAF XML Schema is an electronic implementation of the Common Assessment Framework (CAF). This schema defines the format for CAF data exchange into and out of National eCAF. It will form part of the solution for transferring Common Assessment information between National and Local eCAf systems.

Base URL: http://www.dcsf.gov.uk/everychildmatters/strategy/deliveringservices1/caf/caframework/

3.2 Australia

Australian Government Recordkeeping Metadata Standard:

https://www.naa.gov.au/information-management/information-management-standards/agls-metadata-standard

Australian Institute of Health and Welfare - Metadata Online Registry (METeOR)

3.3 Canada

Records Management Metadata Standard
http://www.collectionscanada.gc.ca/government/products-services/007002-5001-e.html

This standard is based on the Dublin Core; see
http://www.collectionscanada.gc.ca/government/products-services/007002-5001.1-e.html

In particular, in contains seven Dublin Core descriptive metadata elements: Creator, Description, Identifier, Language, Subject, Title and Type.

3.4 Eurostat

Base URL: http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/metadata

Concepts and Definitions Database (CODED):
http://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_NOM_DTL_GLOSSARY&StrNom=CODED2&StrLanguageCode=EN&CFID=4475384&CFTOKEN=7863e757 b805d521-CF12BB8F-D860-0887-7D24223B18D5EE56&jsessionid=1e51a0ba6ae9f1ab363b2c3517235a22665dTR

IV. SPECIFIC INITIATIVES IN THE US GOVERNMENT

These efforts seem to have succeeded to some and perhaps a significant extent because of narrowness, either with respect to type of data or underlying scientific/policy domain. The list here is not complete, and in particular does not address use of metadata standards in financial or other management-oriented contexts.

4.1 Federal Geographic Data Committee (FGDC)

It appears that the most successful initiative in the US is the FGDC. Nearly all ICSP agencies comply with its standards for geospatial data.

“The Federal Geographic Data Committee (FGDC) is an interagency committee that promotes the coordinated development, use, sharing, and dissemination of geospatial data on a national basis. This

nationwide data publishing effort is known as the National Spatial Data Infrastructure (NSDI). The NSDI is a physical, organizational, and virtual network designed to enable the development and sharing of this nation's digital geographic information resources. FGDC activities are administered through the FGDC Secretariat, hosted by the U.S. Geological Survey."

Base URL: http://www.fgdc.gov/

Metadata: http://www.fgdc.gov/metadata/csdgm/

Content Standards: http://www.fgdc.gov/standards/projects/FGDC-standards- projects/metadata/base-metadata/v2_0698.pdf

4.2 Department of Defense

The Data Services Environment (DSE) of the Department of Defense maintains an XML-based metadata registry. The extent to which it is employed is unclear.

Base URL: https://metadata.ces.mil/dse/homepage.htm

Metadata Registry: https://metadata.ces.mil/mdr/homepage.htm

4.3 National Cancer Institute Cancer Data Standards Repository

The "caDSR is a database and a set of APIs and tools to create, edit, control, deploy, and find common data elements (CDEs) for use by metadata consumers, and information about the UML models and Forms containing CDEs for use in software development." It is based on ISO 11179.

Base URL: US National Cancer Institute - Cancer Data Standards Repository (caDSR)

CDE Browser: https://cdebrowser.nci.nih.gov/CDEBrowser/

4.4 Agency for Healthcare Research and Quality

"The United States Health Information Knowledgebase (USHIK) is a metadata registry of healthcare-related data standards funded and directed by the Agency for Healthcare Research and Quality (AHRQ) with management support in partnership with the Centers for Medicare & Medicaid Services."

Base URL: US Health Information Knowledgebase (USHIK)

4.5 Department of Justice

The Global Justice XML Data Model (JXDM) is model-based, object-oriented, and extensible. The official home of the Global JXDM is the Office of Justice Programs (OJP) Information Technology Website. The Global JXDM namespace is http://www.it.ojp.gov/jxdm. This namespace URL resolves to the Global JXDM release archives and associated documentation."

Base URL: US Department of Justice - Global Justice XML Data Model GJXDM

4.6 Departments of Justice and Homeland Security

"NIEM, the National Information Exchange Model, is a partnership of the U.S. Department of Justice and the Department of Homeland Security. It is designed to develop, disseminate and support enterprise-wide information exchange standards and processes that can enable jurisdictions to effectively share critical

information in emergency situations, as well as support the day-to-day operations of agencies throughout the nation."

Base URL: US National Information Exchange Model NIEM

4.7 Library of Congress

The LoC has created the METS: Metadata Encoding and Transmission Standard.

Base URL: http://www.loc.gov/standards/mets/

Primer: http://www.loc.gov/standards/mets/METSPrimerRevised.pdf

XML-Based Schema: http://www.loc.gov/standards/mets/mets.xsd

V. SURVEY OF ICSP WEBSITES

The web sites of all fourteen members of the Interagency Committee on Statistical Policy were searched for four terms:

- Metadata
- Paradata
- Codebook
- Data Dictionary

The results are summarized here.

5.1 Bureau of Economic Analysis

Metadata: The Special Data Dissemination Standard, described in a 1996 paper, addresses:

- Coverage, Periodicity, and Timeliness of the Data
- Access by the Public
- Integrity of the Data
- Quality of the Data

These items are more reminiscent of OMB's Information Quality Guidelines.

URL: http://www.bea.gov/scb/account_articles/general/1096imf/mclen.htm

Additional Quote: "The [Dissemination Standards Bulletin Board] DSBB will therefore identify publicly countries that have subscribed to the standard and will give wide and easy access to the information describing their data and their dissemination practices (the "metadata")."

The Strategic Plan for 2009-2013 mentions (slide 34 of 34), in the context of National Accounts, "Develop new tools to streamline storage and manipulation of structured metadata."

URL: http://www.bea.gov/about/pdf/strategic_plan_matrix_2009-2013.pdf

Codebook: Search returned no results.

Data Dictionary: Search returned no results.

Paradata: Search returned no results.

5.2 Bureau of Justice Statistics

Metadata: Search returned no results.

Codebook: Available for some datasets.

Data Dictionary: Search returned no results.

Paradata: Search returned no results.

5.3 Bureau of Labor Statistics

Metadata appears to have been viewed as a research issue, and implementation as a future possibility. References include:

- "Metadata Standards and Metadata Registries: an Overview" by Bruce E. Bargmeyer, Environmental Protection Agency, and Daniel W. Gillman, Bureau of Labor Statistics (2000)
- The "Role of Metadata in Statistics" by Cathryn S. Dippo and Bo Sundgren (2000)
- "Data and Metadata from the Terminological Perspective" by Daniel W. Gillman, Frank Farance (2009), which cites
    - Farance, F. and Gillman, D. (2006). The Nature of Data. Working Paper #12 in *Proceedings of the UNECE Workshop on Statistical Metadata*, Geneva, Switzerland

Codebook: Multiple references to NLSY97, but these seem to be mainly industry codes. ftp://ftp.bls.gov/pub/time.series/ce/ce.datatype contains short definitions of data elements

Data Dictionary: Multiple versions available. Example URL, for the Consumer Expenditure Survey: http://www.bls.gov/cex/2008/csxdiarydata.pdf

Paradata: Mentioned only in technical papers.

5.4 Bureau of Transportation Statistics

Metadata: There are many mentions, the majority of which appear to be related to geographical data, especially shapefiles for maps. Both the thinking and the implementation are based on the FGDC.

An example is the National Transportation Atlas Databases 2010, which contains

- Metadata file (.xml): XML encoding of shapefile's metadata
- Metadata file (.htm): HTML encoding of shapefile's metadata.
- Metadata file (.txt) : Text version of shapefile's metadata

URL:

http://www.bts.gov/publications/national_transportation_atlas_database/2010/html/liner_notes.html

BTS' Standards Manual contains the following: "Guideline 6.4.2: File Description. Provide complete documentation for all data files.

- Data producers should determine what metadata standards are current at the time data files are prepared and produce associated metadata for their files that comply with applicable standards.
- Documentation must include a description of the data files including the title, data collection sources, tables that make up the set, inter-relation among tables (e.g., keys), and record layouts for data files.

- Documentation must also include descriptions for each variable in the data set that includes the variable name, description, type (categorical, numerical, date/time, etc.), format, entry restrictions (e.g., categories, range), and missing value codes.
- Indicate changes made to previously released data and the "as of" date of the data file."

URL:

http://www.bts.gov/programs/statistical_policy_and_research/bts_statistical_standards_manual/html/chapter_06.html

Codebook: Some codebooks are available. An example is http://www.bts.gov/programs/omnibus_surveys/targeted_survey/2002_national_transportation_availability_and_use_survey/public_use_data_files/excel/code_book.xls

Data Dictionary: Similar mentions as for metadata, suggesting that the two terms might be seen as interchangeable.

Paradata: Search returned no results.

5.5 Census Bureau

Metadata. All Census geographic data are FDGC compliant.

A research effort in the late 1990's led to a proposal for a metadata standard:

- W. LaPlant, D. Gilman, M. Appel: http://www.census.gov/prod/2/gen/96arc/viiblapl.pdf, (See also http://www.census.gov/prod/2/gen/96arc/viiibgil.pdf)

However, the effort does not appear to have led to any implementations.

Codebook: Large number of results.

Data Dictionary: Very large (461) number of results. Typically, these are text files, as exemplified by http://www.census.gov/sipp/dictionaries/2004prelmw1d.txt

Paradata: Search returned only 6 rather uninformative mentions.

5.6 Economic Research Service (USDA)

Metadata: The search produced only 7 mentions. The web site mentions a 2008 summer internship whose goal was to "improve the metadata on the public website," and specifically to

- Review and report on metadata cataloged in the content management tool and on individual web pages.
- Create a plan for how to improve and leverage metadata across the website.
- Revise metadata across the website to meet established goals.

Discussion of the consequences of this effort could not be located.

Codebook: Information identified as codebooks seems to be included in many Excel data files. An example is http://www.ers.usda.gov/Briefing/CPIFoodAndExpenditures/Data/qfahpd/fruitsandvegetables.xls

Data Dictionary: In some cases, this information is embedded in SAS code to read data files, as exemplified by http://www.ers.usda.gov/Data/foodsecurity/CPS/updatecode1995.htm

In others, it is on the CD containing the data file, an example of which is http://www.ers.usda.gov/Data/foodsecurity/CPS/TCHDOC1202.htm

In still others, it is a standalone file: see http://www.ers.usda.gov/data/foodsecurity/spd/spd99.pdf

Paradata: Search returned no results.

5.7 Energy Information Administration

Metadata appears to be viewed as a research issue. However, "Metadata Products: Descriptions of EIA information products to help customers find what they need. They include directories of all our survey forms, publications, electronic products, models, new releases, energy education resources, and EIA contacts." is discussed at

http://www.eia.doe.gov/smg/asa_meeting_2006/fall/introeia.html

Codebooks seem principally to be descriptions of file structure. An example is: http://www.eia.doe.gov/emeu/recs/recspubuse87/codebooks/All_codebook_files.txt

Data Dictionary: Search returned essentially no results.

Paradata: Search returned no results.

5.8 National Agricultural Statistics Service

Metadata. Geographical are FGDC compliant. An example is the Census of Agriculture FGDC Content Standards for Digital Geospatial Metadata (FGDC-STD-001-1998): http://www.agcensus.usda.gov/Publications/2007/Online_Highlights/Ag_Atlas_Maps/mapfiles/ag_co_metadata_faq_07.html

Codebook: Search returned no results.

Data Dictionary: The term seems to be used synonymously with metadata, and often to mean file structure. An example is http://www.nass.usda.gov/research/Cropland/metadata/metadata_mo03.htm

Paradata: Search returned no results.

5.9 National Center for Education Statistics

Metadata. A metadata task force produced the *Forum Guide to Metadata*, which is available at http://nces.ed.gov/pubs2009/2009805.pdf. It is not clear whether this document lead to any major changes in NCES' practices.

The NCES Statistical Standards discuss metadata, but do not contain strong requirements:

"Standards: **GUIDELINE 7-1-1B**: To facilitate the sharing and use of data elements, national and international standards organizations have produced drafts of several standards for the creation of metadata on data elements. Examples are the International Organization for Standards "Specification and Standardization of Data Elements" standard (ISO/IEC 11179) and the more detailed American National Standards Institute "Metadata for the Management of Shareable Data" Standard (ANSI X3.285) www.ansi.org. These standards continue to be refined. Data producers should determine what metadata

standards are current at the time data files are prepared and produce associated metadata for their files that are in compliance with applicable standards."

Distinctions between metadata and file structure are sometimes vague. "**STANDARD 7-1-2**: A file description and record layout must be provided for each file. The file information/metadata [SIC] header must include the following:

1. Title of the survey (survey name, part, and year as applicable);
2. Name(s) of each file;
3. Reference year for the data;
4. Version number and date of release;
5. Logical record length (in positional files) or number of variables on the file (delimited files);
6. Number of records per case or observation; and
7. Number of cases in the data file. For delimited files also include the delimiters (e.g., comma, space)."

Similarly, "**STANDARD 7-1-3**: For each variable on the file, the file description must include the following:

1. Variable name;
2. Data type (alpha or numeric);
3. Record number (if multiple records per case);
4. Position within the record (beginning-end, or variable number if delimited) within the record, field length, and variable label; and
5. The survey question wording and response categories."

The National Assessment of Educational Progress (NAEP) appears to consider metadata to be the same as file layout: http://nces.ed.gov/nationsreportcard/tdw/database/metadata.asp.

There is some discussion in the context of Department of Education State Longitudinal Stat Systems (SLDS) programs, an example of which is http://nces.ed.gov/programs/slds/state.asp?stateabbr=KS

Codebook: Almost all NCES data products have an electronic codebook and associated documentation. An example is the Early Childhood Longitudinal Study (ECLS-K)
http://nces.ed.gov/ecls/data/ECLSK_K8_Manual_part2.pdf

There are also codebooks for public use files; see for example, http://nces.ed.gov/pubs2010/2010334.pdf.

Data Dictionary: In the Forum Guide, http://nces.ed.gov/pubs2009/metadata/ch2_dictionaries.asp "A data dictionary is an agreed-upon set of clearly and consistently defined elements, definitions, and attributes … Although many items in a data dictionary can be classified as metadata, data dictionaries and metadata systems are not identical. Data dictionaries generally contain only some of the metadata necessary for understanding and navigating data elements and databases and, thus, contain only a subset of the metadata found in a robust metadata system. Metadata systems, on the other hand, generally include the entire range of items used for data system management and analysis, including features for sorting, searching, organizing, and connecting data and metadata (see exhibit 2.3).

The Data Systems Standards: http://nces.ed.gov/dataguidelines/ contains the National Education Data Model (http://nces.ed.gov/forum/datamodel/). There are also references to 2 XML standards

- Postsecondary Electronic Standards Council (PESC)
  PESC supports the maintenance and implementation of electronic data interchange (EDI) and the development, maintenance, and implementation of extensible markup language (XML).
- Schools Interoperability Framework (SIF)
  SIF, an organization that develops open source XML standards for P-12 data, has developed an Implementation Toolkit containing three documents that are designed to serve as an aid for schools.

Paradata: Search returned no results.

5.10 National Center for Health Statistics

Metadata is mentioned many (240+) times, although meaningful implementation seems to be less common. Examples of the latter include:

- PHIN (Public Health Information Network) Metadata Standards: http://www.cdc.gov/phin/library/documents/pdf/PHIN_Vocabulary_Metadata_v1.0.pdf
- EPHT (Environmental Public Health Tracking Network): http://www.cdc.gov/nceh/tracking/netvision/netvision_feat.htm: "Metadata are required for all data sets that will be available through the EPHT Network, and metadata will consist of a standard set of elements. Metadata will be presented in a manner to allow easy searching and discovery of data (e.g., through the use of key words)."

Codebook: This search produced 2150 results. The term appears to be uses in the sense of variable names and definitions. An example, from the National Health and Nutrition Examination Survey (NHANES) is http://www.cdc.gov/nchs/data/nhanes/nhanes_03_04/l28ocp_c.pdf

Data Dictionary, which yielded 338 results appears to mean file layout. An example is http://www.cdc.gov/nchs/data/nhhcs/2007NHHCSPublic-UseFileDataDictionary.pdf

Paradata. NCHS appears to be unique among ICSP agencies in disseminating paradata files, albeit only for a few of its surveys, notably the National Health Interview Survey (NHIS).

Examples of such files are:

- The NHIS Paradata file: http://www.cdc.gov/nchs/nhis/2008paradata.htm and PDF documents cited there.
  - ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHIS/2008/srvydesc_paradata.pdf
  - ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHIS/2008/paradata_summary.pdf

5.11 Office of Environmental Information (US EPA)

Metadata: Most search results point to presentations at OEI-sponsored workshops, such as

- http://www.epa.gov/OEI/symposium/2010/fagan.pdf
- http://www.epa.gov/OEI/symposium/2005/naranjo2.pdf

Information about OEI's GeoData Gateway suggests that geographical data are FDGC compliant.

Codebook: Search returned no results.

Data Dictionary: Search returned no results.

Paradata: Search returned no results.

5.12 Office of Research, Evaluation and Statistics (SSA)

Metadata: Search returned no results.

Codebook. There is some discussion in the context of linkage to other databases at http://www.socialsecurity.gov/policy/docs/ssb/v69n2/v69n2p1.html

Data Dictionary appears to mean file layout. An illustrative example: http://www.socialsecurity.gov/appeals/DataSets/dataDictionary/Data_Dictionary.pdf

Paradata: Search returned no results.

5.13 Division of Science Resources Statistics (NSF)

Metadata: A general discussion appears at http://www.nsf.gov/statistics/seind10/pdf/methodology.pdf

Codebook: (211 mentions) The term sometimes has a seemingly unusual usage, however, to include both survey instruments data summaries.

Data Dictionary: Search returned no results. However, for some public use files, information is available, for instance, for the Graduate Students and Postdoctorates in Science and Engineering Survey (GSS), see http://www.nsf.gov/statistics/srvygradpostdoc/data08/guide2008.pdf.

Paradata: Search returned no results.

5.14 Statistics of Income Division (IRS)

Metadata: which produces 48 results, appear to mean term definitions. See, for instance, http://www.irs.gov/taxstats/article/0,,id=214353,00.html

Codebook: Search returned no results.

Data Dictionary: Search returned no informative results.

Paradata: Search returned no results.

VI. OTHER ITEMS OF INTEREST

These are listed because of possible relevance. In general, US states seem to be rather far behind the Federal government, even though many states make data available on their web sites.

- Minnesota Department of Education Metadata Registry (K-12 Data)
- Minnesota Department of Revenue Property Taxation (Real Estate Transactions)
- NORC Workshop in 2008: http://www.norc.org/news/metadata+workshop+-+the+next+frontier+in+documenting+survey+data.htm

VII. IMPLICATIONS FOR THE SCOP

The following figure, adapted from the DDI documentation, helps establish a context for discussing next steps. It represents the data collection, processing and archiving process. Sound metadata are meant to facilitate each step in this process. The figure also formalizes a somewhat broader view than is now prevalent. In particular it calls attention to the "repurposing" that may now be the dominant manner in which some data are used.

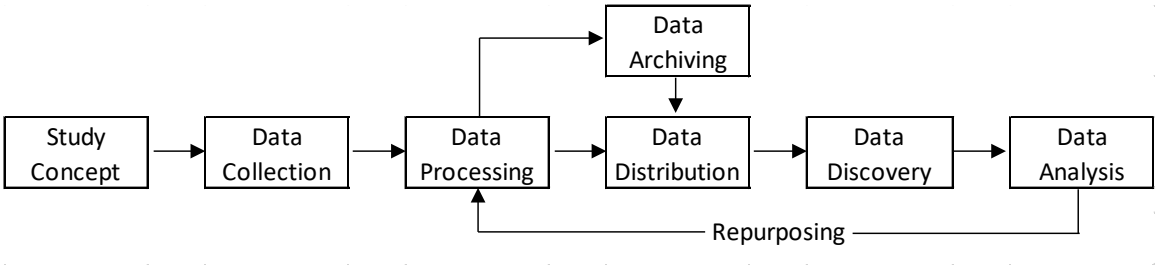

Figure 1: DDI schematic rendition of the "data process."

7.1 The Current Situation

Unlike other problem contexts, with respect to metadata and paradata, it does not appear that the current situation is one of "best practice somewhere but not everywhere." The only solidly successful model is the FGDC. The extent to which this model works because the standards are administered by external body is unclear but suggestive. SCOP itself reflect the absence to date of a collective commitment by the ICSP agencies to move forward. Even intra-agency commitment, other than to the FDGC is uneven. Research-catalyzed efforts in some agencies in the late 1990s to produce metadata schema seem to have foundered.

The "state of the art," in terms of systems on which ICSP agencies can build, does seem to consist of DDI and SDMX. Potentially salient differences between the two are that:

1. DDI is based on XML, while SDMX is based on UML.
2. DDI has its origins in social sciences, which may make existing schema more compatible with the kinds of survey data collected and disseminated by the ICSP agencies. SDMX, by contrast, arose in the context of financial data, which are clearly relevant to some ICSP agencies, but less so to others.
3. The participation of EUROSTAT , on the other hand, represents the commitment of a major statistical agency to SDMX.

The International Household Survey Network (Section 3.3.2), while a step in the right direction, does not constitute the basis for a metadata system.

7.2 Next Steps

There seems to be one essential first step: agreement on definitions for data, metadata and paradata. Section 2 outlines initial steps on this path. Above all else, *there must be a concept of metadata that is completely independent of that of file structure*. There is likely to be disagreement about the value of, in computer science language, "deprecating" the terms "codebook" and "data dictionary" and replacing them with "metadata," but doing this does seem necessary in the long run.

The second essential step, which amounts to recognition of reality, is *commitment to markup language-based metadata, and ultimately to markup language-based data*. The two appendices to this report make concrete on a small scale what this entails; below we propose what we believe to be a first "step in the right direction." Ultimately, this path leads to a major change in the way ICSP agencies disseminate their data, in which system-specific (for instance, SAS or Excel) files or formats (for instance, CSV) would be replaced by markup language-based files and parsers.[6] Nothing else will work in the long run.[7]

Operationally, this step seems to entail a choice between DDI and SDMX (and by implication, between XML and UML) as the basis for data and metadata. An important prerequisite to this decision would be for the ICSP agencies to undertake, or ask an external body to undertake a detailed comparison between DDI and SDMX.

A sensible third step would be to prepare a demonstration ("proof-of-concept") case, based on one of both of DDI and SDMX, consisting *of full markup-language-based metadata and data files for a modest-scale survey* conducted by one of the ICSP agencies, including development of parsers that would put the data into common formats. The survey should be complex enough to permit full understanding of the issues, but not so complex (and in particular, not longitudinal) that the effort is either too lengthy or too expensive. In even of NCES' role in SCOP, a cross- sectional survey such as the Schools and Staffing Survey (SASS) might be a suitable choice.

Based on the outcome of the third step, more difficult next step would follow: creation of an ICSP-wide *extensible metadata template for surveys*, together with tools for common tasks. The administrative framework might be analogous to that of the FGDC, with an (external?) Oversight Committee that controls the core of the template and operates a mechanism for making additions. The goal would be to capture a set of major components common to essentially all ICSP-agency surveys, including (This list is meant to be illustrative, not prescriptive!)

- "Facts:" Agency, legislation, purpose, dates, contractor, contractor number, cost
- Key Words: For identification of datasets
- Design: Population, frame, sampling (PSAs, …), sample size, design weights,
- Data collection Instrument and Mode(s)
- Data Collection: Time period(s), unit response rate (unit history),
- Variables
    - Definition
    - Format, Units, Flags (missing, edit, imputation)
    - Item response rate
    - Classification: frame, collected, obtained from administrative data, calculated from other variables
    - Access restrictions
    - Documentation, in the form of URLs or other pointers, including information no adjustments for nonresponse bias and methodologies for edit, imputation and SDL.

[6] Agencies could, of course, apply the parsers to create and disseminate files in currently "popular" formats.

[7] Even today, many researchers cannot handle legacy file formats such as fixed-width fields in which padding with space characters prevents numbers from being recognized as such. Even recognizing the problems can require hexadecimal dumps of files.

Undertaking this step involves, or at leads to, collective agency commitment to prepare, or have contractors prepare, core metadata for all future (and some past?) surveys.

The remaining step, which is more complex than the others, would be creation of a *repository for metadata for surveys*.

Other issues need to be addressed. The most pressing of these was discussed in Section 2: creation of a conceptual and operational basis for dealing with confidentiality and SDL in the context of markup language-based metadata (and, for that matter, markup language-based data). A second issue is cross-survey compatibility, which is embedded in the "extensible metadata template" step described above, and is very important because it instantiates standardization across surveys. A third issue, which seems crucial, is to ensure that the metadata template facilitates linking datasets. A final issue, which may less pressing, is provenance. Currently, agencies tightly monitor and control their datasets, even when collection is done by a contractor. However, as more and more sharing and repurposing of data occurs, failure to document provenance may have more serious consequences.

APPENDICES

Appendix A: An Illustrative Example

This example is designed to show in an informative but not overwhelming way how markup language-based metadata works.

Consider the following data table.

Name	Gender	Age	Height (cm)	Weight (kg)
Joe Smith	M	33	170	73
Bob Jones	M	26	195	65 [Imputed]
Mary White	F	57	145	[Missing]

Then, although more information could be included, a DDI-style[8] description of the data alone is given by the following metadata object.

```
<DATA DESCRIPTION>

    <ATTRIBUTE>
        <NAME>Name</NAME>
        <TYPE>Text</TYPE>
        <MISSING INDICATOR>Empty </MISSING INDICATOR>
        <IMPUTATION >
            <IMPUTATION PRESENT>No</IMPUTATION PRESENT>
        </IMPUTATION >
    </ATTRIBUTE>

    <ATTRIBUTE>
        <NAME>Gender</NAME>
        <TYPE>Text</TYPE>
        <ALLOWABLE VALUE>M</ALLOWABLE VALUE>
        <ALLOWABLE VALUE>F</ALLOWABLE VALUE>
        <MISSING INDICATOR>Empty </MISSING INDICATOR>
        <IMPUTATION>
            <IMPUTATION PRESENT>No</IMPUTATION PRESENT>
        </IMPUTATION>
    </ATTRIBUTE>

    <ATTRIBUTE>
        <NAME>Age</NAME>
        <TYPE>Numerical</TYPE>
        <UNITS>Years</UNITS>
        <PRECISION>5-Year Ranges</PRECISION>
        <ALLOWABLE VALUE>0-5</ALLOWABLE VALUE>
```

[8] IMPORTANT: This specification is not DDI compliant, and is not, for clarity, put into the DDI schema.

```
        …
        <ALLOWABLE VALUE>196-200</ALLOWABLE VALUE>
        <MISSING INDICATOR>Empty </MISSING INDICATOR>
        <IMPUTATION>
                <IMPUTATION PRESENT>No</IMPUTATION PRESENT>
        </IMPUTATION>
</ATTRIBUTE>

<ATTRIBUTE>
        <NAME>Height</NAME>
        <TYPE>Numerical</TYPE>
        <UNITS>Centimeters</UNITS>
        <PRECISION>Integer</PRECISION>
        <ALLOWABLE VALUE>150</ALLOWABLE VALUE>
        …
        <ALLOWABLE VALUE>400</ALLOWABLE VALUE>
        <MISSING INDICATOR>Empty </MISSING INDICATOR>
        <IMPUTATION>
                <IMPUTATION PRESENT>No</IMPUTATION PRESENT>
                <IMPUTATION FLAG>N/A</IMPUTATION FLAG>
                <IMPUTATION METHOD>N/A</IMPUTATION METHOD>
        </IMPUTATION >
</ATTRIBUTE>

<ATTRIBUTE>
        <NAME>Weight</NAME>
        <TYPE>Numerical</TYPE>
        <UNITS>Kilograms</UNITS>
        <PRECISION>Integer</PRECISION>
        <ALLOWABLE VALUE>25</ALLOWABLE VALUE>
        …
        <ALLOWABLE VALUE>150</ALLOWABLE VALUE>
        <MISSING INDICATOR>Empty </MISSING INDICATOR>
        <IMPUTATION>
                <IMPUTATION PRESENT>Yes</IMPUTATION PRESENT>
                <IMPUTATION FLAG>WeightImputeFlag=1</IMPUTATION FLAG>
                <IMPUTATION METHOD>Weight = Height/3</IMPUTATION METHOD>
        </IMPUTATION>
</ATTRIBUTE>
```

```
<ATTRIBUTE>
        <NAME>WeightImputeFlag</NAME>
        <TYPE>Binary</TYPE>
        <ALLOWABLE VALUE>0</ALLOWABLE VALUE>
        <ALLOWABLE VALUE>1</ALLOWABLE VALUE>
        <MISSING INDICATOR>Empty </MISSING INDICATOR>
</ATTRIBUTE>
```

</DATA DESCRIPTION>

For each attribute, this description contains some of: its name, its type (text or numerical), the units in which it is reported, the numerical precision, allowable values, the form in which missing values are reported, and whether imputation is possible, as well as if so, the manner which in which it is indicated, the method (imputing weight as height divided by 3 is purely illustrative).

An example of an associated markup language version of the actual data file is then. With a suitable parser, this information can be put into any format, including (tab or comma) delimited text, a SAS data object, an R data object, and an Excel file. Using the metadata, it can check whether the values in the file are valid.

<DATA FILE>

```
<RECORD>
        <Name> Joe Smith</Name>
        <Gender>M</Gender>
        <Age>31-35</Age>
        <Height>180</Height>
        <Weight>73</Weight>
        <WeightImputeFlag>0</WeightImputeFlag>
</RECORD>

<RECORD>
        <Name> Bob Jones</Name>
        <Gender>M</Gender>
        <Age>26-30</Age>
        <Height>195</Height>
        <Weight>65</Weight>
        <WeightImputeFlag>1</WeightImputeFlag>
</RECORD>

<RECORD>
        <Name> Mary White</Name>
        <Gender>F</Gender>
        <Age>56-60</Age>
        <Height>180</Height>
```

```
        <Weight>73</Weight>
        <WeightImputeFlag>0</WeightImputeFlag>
    </RECORD>
</DATA FILE>
```

Note that there is neither logical nor implementation-driven need that the attributes of a record be any prescribed order. The following file is completely equivalent to the one above, and proper parsers would have no problem dealing with it.

```
<DATA FILE>

    <RECORD>
        <Age>31-35</Age>
        <Name> Joe Smith</Name>
        <WeightImputeFlag>0</WeightImputeFlag>
        <Gender>M</Gender>
        <Weight>73</Weight>
        <Height>180</Height>
    </RECORD>

    <RECORD>
        <Height>195</Height>
        <Gender>M</Gender>
        <Age>26-30</Age>
        <WeightImputeFlag>1</WeightImputeFlag>
        <Weight>65</Weight>
        <Name> Bob Jones</Name>
    </RECORD>

    <RECORD>
        <Height>180</Height>
        <Weight>73</Weight>
        <Gender>F</Gender>
        <Age>56-60</Age>
        <WeightImputeFlag>0</WeightImputeFlag>
        <Name> Mary White</Name>
    </RECORD>

</DATA FILE>
```

Building, for example, a CSV data file would require the following markup language description of a physical data product.

```
<PHYSICAL DATA PRODUCT>

    <FORM>Delimited</FORM>
    <RECORD>
        <FIELD>
            <NAME>ID</NAME>
            <ATTRIBUTE>Name</ATTRIBUTE>
            <ORDER>1</ORDER>
            <SEPARATOR>,</SEPARATOR>
        </FIELD>
        <FIELD>
            <NAME>Sex</NAME>
            <ATTRIBUTE>Gender</ATTRIBUTE>
            <ORDER>2</ORDER>
            <SEPARATOR>,</SEPARATOR>
        </FIELD>
        <FIELD>
            <NAME>Age in Years</NAME>
            <ATTRIBUTE>Age</ATTRIBUTE>
            <ORDER>3</ORDER>
            <SEPARATOR>,</SEPARATOR>
        </FIELD>
        <FIELD>
            <NAME>Height (cm)</NAME>
            <ATTRIBUTE>Height</ATTRIBUTE>
            <ORDER>4</ORDER>
            <SEPARATOR>,</SEPARATOR>
        </FIELD>
        <FIELD>
            <NAME>Weight (kg)</NAME>
            <ATTRIBUTE>Weight</ATTRIBUTE>
            <ORDER>5</ORDER>
            <SEPARATOR>,</SEPARATOR>
        </FIELD>
        <FIELD>
            <NAME>Impute Flag for Weight</NAME>
            <ATTRIBUTE>WeightImputeFlag</ATTRIBUTE>
            <ORDER>6</ORDER>
            <SEPARATOR>,</SEPARATOR>
        </FIELD>
        <EOR>Carriage return/Line feed</EOR>
    </RECORD>
```

</PHYSICAL DATA PRODUCT>

The contents of the CSV file itself, would then be:

“ID”, “Sex”, “Age in Years”, “Height (cm)”,”Weight (kg)”, “Impute Flag for Weight" “Joe Smith”,M,33,170,73,0

“Bob Jones”,M,26,195,65,1

“Mary White”,F,57,145,,0

Each line of this file would end with a carriage return character followed by a line feed character. The parser would omit the comma separator following last entry in each line. The quotation marks are required for CSV compliance, and are removed by applications that read CSV files.

Appendix B: WDSL for Data Swapping

This (real) example contains the XML-based WSDL (Web Services Description Language) file for a Web services implementation of data swapping created by NISS. Briefly, this service allows users to perform data swapping using remote software; see "NISSWebSwap: A Web Service for data swapping," by A. Sanil, S. Gomatam, A. F. Karr and S. Liu, *Journal of Statistical Software* **8(7)** (2003) for a complete description. The markup structure should be apparent. It allows communication of the file name and parameters for the swapping.

```
<?xml version="1.0" encoding="UTF-8"?>

<definitions name="Swap_dataService" targetNamespace=http://WebSwap_swap.org/wsdl.

    xmlns:tns="http://WebSwap_swap.org/wsdl" xmlns=http://schemas.xmlsoap.org/wsdl/
    xmlns:soap="http://schemas.xmlsoap.org/wsdl/soap/" xmlns:ns2="http://WebSwap_swap.org/types"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema">


<types>

<schema targetNamespace=http://WebSwap_swap.org/types
    xmlns:xsi=http://www.w3.org/2001/XMLSchema-instance
    xmlns:tns="http://WebSwap_swap.org/types" xmlns:soap-
    enc="http://schemas.xmlsoap.org/soap/encoding/" xmlns:wsdl="http://schemas.xmlsoap.org/wsdl/"
    xmlns="http://www.w3.org/2001/XMLSchema">


<complexType name="SwapData">
        <sequence>
                <element name="numFields" type="int"/>
                <element name="outputFile" type="string"/>
                <element name="numRecords" type="int"/>
                <element name="riskCutoff" type="double"/>
                <element name="data" type="tns:ArrayOfArrayOfstring"/>
                <element name="dataFile" type="string"/>
                <element name="constraints" type="base64Binary"/>
                <element name="log" type="tns:ArrayOfstring"/>
                <element name="swapRate" type="double"/>
                <element name="riskFraction" type="double"/>
                <element name="logFile" type="string"/>
                <element name="csvType" type="string"/></sequence>
</complexType>


<complexType name="ArrayOfArrayOfstring">
        <complexContent>
                <restriction base="soap-enc:Array">
                        <attribute ref="soap-enc:arrayType" wsdl:arrayType="tns:ArrayOfstring[]"/>
```

```
                </restriction>
        </complexContent>
</complexType>

<complexType name="ArrayOfstring">
        <complexContent>
                <restriction base="soap-enc:Array">
                <attribute ref="soap-enc:arrayType" wsdl:arrayType="string[]"/>
                </restriction>
        </complexContent>
</complexType>

</schema>
</types>

<message name="doSwap">
        <part name="SwapData_1" type="ns2:SwapData"/>
</message>

<message name="doSwapResponse">
        <part name="result" type="ns2:SwapData"/>
</message>

<portType name="SwapIF">
        <operation name="doSwap">
                <input message="tns:doSwap"/>
                <output message="tns:doSwapResponse"/>
        </operation>
</portType>

<binding name="SwapIFBinding" type="tns:SwapIF">
        <operation name="doSwap">
                <input>
                        <soap:body encodingStyle="http://schmas.xmlsoap.org/soap/encoding/"
                        use="encoed" namespace="http://WebSwap_swap.org/wsdl"/>
                </input>
                <output>
                        <soap:body encodingStyle=http://schemas.xmlsoap.org/soap/encoding/
                        use="encoded" namespace="http://WebSwap_swap.org/wsdl"/>
                </output>
```

```
            <soap:operation soapAction=""/>
        </operation>

<soap:binding transport="http://schemas.xmlsoap.org/soap/http" style="rpc"/>
</binding>

<service name="Swap_data">
        <port name="SwapIFPort" binding="tns:SwapIFBinding">
            <soap:address location="http://www.niss.web-services:8080/WebSwap/SwapIF"/>
        </port>
</service>

</definitions>
```

Appendix C: Author

Alan F. Karr, PhD

Director, National Institute of Statistical Sciences