DDI and SDMX: Complementary, Not Competing, Standards

Arofan Gregory (agregory@opendatafoundation.org)
Pascal Heus (pheus@opendatafoundation.org)

Version 1.0, July 2007

Contents

Overview 2
I. Introducing the Standards 2
II. Data and Metadata Formats 3
  A. SDMX 4
  B. DDI 5
  C. DDI/SDMX Overlap 6
III. Process Management and the Data Lifecycle 7
IV. Registries and Repositories 9
V. Implementation of the Standards 10
  A. DDI 10
    1. Upstream Capture of Metadata/Supporting Data Collection 11
    2. Group Surveys in Catalog or Series 11
    3. Managing Question Banks 11
    4. Capture of Metadata Regarding Researcher Activity 12
    5. Comparison-by-Design for Series of Studies 12
    6. Mining the Archive for After-the-Fact Comparability 13
  B. SDMX 13
  C. SDMX Plus DDI 15
VI. Summary 16

Overview

Recently, two technical standards for statistical and research data and metadata have been receiving much attention. Particularly for those working with both micro-data and time-series aggregates, there can be some confusion as to the relationship between these standards, and questions about which may be more appropriate for use in a particular application or institution. This paper describes the basic scope of each standard, and provides some information which may help in making a decision about which of them is most suitable.

I. Introducing the Standards

The Statistical Data and Metadata Exchange[1] (SDMX) technical specifications come out of the world of official statistics and aim to foster standards for the exchange of statistical information. They have been created by the Statistical Data and Metadata Exchange Initiative. The initiative is a cooperative effort between seven international organizations: the Bank for International Settlement[2] (BIS), the International Monetary Fund[3] (IMF), the European Central Bank[4] (ECB), Eurostat[5], the World Bank[6] (WB), the Organization for Economic Co-operation and Development[7] (OECD), and the United Nations Statistical Division[8] (UNSD). The output of this initiative is not just the technical standards, but also addresses the harmonization of terms, classifications, and concepts which are broadly used in the realm of aggregate statistics. The technical standards are now in their second version. The first version is an ISO Technical Specification, ISO-17369. The second version has been put forward for ISO status.

The Data Documentation Initiative[9] (DDI) is a specification for capturing metadata about social science data. It is maintained by the Data Documentation Initiative Alliance, a membership-driven consortium including universities, data archives, and national and international organizations. The specification was originally created to capture the information found in survey codebooks, which remains the focus of the first two versions. The new 3.0 version - now in Candidate Draft status for implementation testing - covers the whole data lifecycle, from the survey instrument design to archiving, dissemination and repurposing, allowing

[1] http://www.sdmx.org
[2] http://www.bis.org
[3] http://www.imf.org
[4] http://www.ecb.int
[5] http://ec.europa.eu/eurostat
[6] http://www.worldbank.org
[7] http://www.oecd.org
[8] http://unstats.un.org
[9] http://www.ddialliance.org

for a description of re-codes, processing, and comparison of studies by design or after-the-fact.

SDMX and the latest version of the DDI have been intentionally designed to align themselves with each other as well as with other metadata standards. Because much of the micro-data described by DDI instances is aggregated into the higher-level data sets found at the time-series level, it is not surprising that the two have been designed to work well together. Although there is some overlap in their descriptive capacity, they can best be characterized as complementary, rather than competing.

One point of similarity between SDMX and the 3.0 version of DDI is the existence in each case of a conceptual model, which forms the basis of the XML implementation. In SDMX, there is the SDMX Information Model, which is a meta-model of the exchange and dissemination processes around aggregate statistics. In DDI, the model is a more specific metadata model based on a particular view of the data lifecycle. These models have many similarities, in part because both are aligned with ISO/IEC 11179[10] (a widely accepted standard concerning semantics and metadata registries), but also because they were authored by a team which had a significant degree of cross-membership in both initiatives. This use of conceptual models to inform the creation of standard XML schemas represents the current state of practice in the design of international XML standards, and can be seen in many different standards, of which DDI and SDMX are only two.

This paper will characterize the basic features of each version of the standards, describing both how they support applications within their intended scope, and also how they can be used together within a single application to provide complementary functionalities. This presentation is organized according to how each standard addresses data formats and metadata formats, how they view process management and the data lifecycle, and how each addresses the topic of registries and repositories. A discussion of how each can fit into implementations is provided as a summary.

II. Data and Metadata Formats

One of the key differences between DDI and SDMX is the intended use of the standards. DDI, in its early versions, was primarily intended as an archival standard, providing an electronic format for descriptive, human-readable metadata for researchers in place of the paper codebooks which had previously served as documentation for survey data. This orientation is important - DDI has

[10] http://metadata-stds.org/11179/

a requirement to describe studies after-the-fact, and as such cannot make assumptions about how the data set in question is structured.

A. SDMX

SDMX has a different focus - it is designed to facilitate the automated exchange and processing of data and metadata between organizations. There is thus no requirement for SDMX to describe a wide range of different types of data structures - it imposes a typical data structure, which can be mapped into and out of by the different counterparties involved in the exchange. This difference in intention is important in understanding how the two standards function.

SDMX has several different data and metadata formats: for time-series data, for cross-sectional data, for describing the structures of data sets ("structural metadata"), for independent metadata sets (termed "reference metadata"), and for describing the structures of independent metadata sets (another form of "structural metadata"). In the 1.0 version of the SDMX Technical Specifications, there was no provision for independent exchange of non-structural metadata - this was added in the 2.0 version of the specifications. Examples of this type of metadata include footnote metadata, metadata about data quality, statistical metadata, and methodological metadata. Typically, independent metadata is produced and disseminated in exchanges which are separate from - but may be in reference to - the exchange and dissemination of statistical data.

Because of the very broad range of concepts found in the many domains within official statistics, SDMX was designed as a meta-model. This means that the structural metadata formats are used to configure the data formats and metadata formats, indicating which concepts will be used in the reporting of data and metadata. SDMX places no restrictions on the concepts used: each exchange has a structure which indicates the concepts to be used, and how they are represented, based on the preferences of the counterparties. These structural formats are known as "data structure definitions (DSDs)" (or "key families") and "metadata structure definitions (MSDs)". The structure of a data or metadata format is clearly separate from the accompanying data or metadata payload format. An interesting approach is used toward the XML implementation of these various structures: the XML schemas for data and metadata sets are generated from the structural metadata. Thus, if I have a concept such as "TOPIC" in my structural definition, I will have a corresponding XML construct such as <TOPIC> in my XML format for the data or metadata payload. Further, SDMX provides a range of equivalent data and metadata formats which correspond to different technical use cases for the same set of data or independent metadata.

This approach is best illustrated using examples of each type. Here is a partial summary of the contents of a data structure definition. Note that this is typically encoded in an XML instance, but here has been spelled out in plain text for readability.

Concept: FREQUENCY (functions as a dimension - uses Frequency code-list: Annual, Quarterly, Monthly, etc.)
Concept: REFERENCE_AREA (functions as a dimension - uses ISO Country code-list)
Concept: OBSERVATION_VALUE (functions as a measure - numeric datatype)
Concept: TOPIC (functions as a dimension - uses a code-list which enumerates the topics of the statistical dataset)
Concept: TIME (functions as a dimension - uses ISO date representation)

In the derived SDMX data format, the resulting XML (in the "Compact" format) for a time series would look like this:

```
<org:DataSet>
      <org:Series FREQUENCY="A" REFERENCE_AREA="CH" TOPIC="03">
            <org:Obs TIME="2004" OBSERVATION_VALUE="3.145"/>
            <org:Obs TIME="2005" OBSERVATION_VALUE="2.96"/>
            <org:Obs TIME="2006" OBSERVATION_VALUE="3.457"/>
            <org:Obs TIME="2007" OBSERVATION_VALUE="4.206"/>
      </org:Series>
</org:DataSet>
```

Note: The prefix "org:" on each tag simply indicates the namespace for the particular schema which corresponds to the data structure definition. For each concept, there is a tag, which is placed in a specific position within the schema based on whether it functions as a measure, as time (which is a special type of dimension), or as a regular dimension. Each tag corresponds to a concept in the data structure definition. Note that whoever writes the data structure definition - by choosing the concepts, and how they function - dictates what the resulting XML data format will look like.

A very similar approach is used in the SDMX Metadata Report format, where a specified concept becomes a "metadata attribute". Concepts are arranged in a presentational hierarchy, and this is reflected in a hierarchical set of XML elements, each one of which represents the corresponding concept (with which it shares a name).

B. DDI

DDI in the 1.* and 2.* versions does not have a format for data - the data is typically held in text-based or proprietary file formats. While the DDI metadata can describe these files to make them easily processible, there is no XML data format. The ability to store data in the XML has been introduced in DDI version 3.0.

In DDI, there is a hybrid approach to the encoding of metadata. On the one hand, many fields in the XML schemas reflect specific concepts which are hard-coded

into the schemas (for example, <Citation> fields and <Universe>.) On the other hand, concepts can be specified and bound to Variables, such that these become user-configurable descriptors of the data. This latter approach is similar to the one found in SDMX regarding concepts and dimensions in our example above.

In DDI, it is possible to describe micro-data, and it is also possible to describe tabular layouts and multi-dimensional cubes. Further, these structures may include the multi-dimensional data. In SDMX, data is typically organized in multi-dimensional cubes (although a tabular format can be seen - and described - as a simple example of this). As such, it is possible to describe a tabular micro-data structure in SDMX. The XML is however not optimized for this.

C. DDI/SDMX Overlap

This is the major point of overlap between these two standards: SDMX describes the structure of multi-dimensional data cubes, and then provides the data formats for these. DDI can be used to do exactly the same thing. Despite this overlap, there is a significant difference between the standards in this area.

SDMX allows for only very regular, "clean" cube structures, and assumes that any other type of cube structure can be mapped into the "clean" SDMX structure before exchange. DDI - because it has a requirement to describe data cubes after-the-fact for documentation purposes - must allow for the description of any type of multi-dimensional cube whatsoever. This means that SDMX cubes tend to be simpler and easier to process, because they have been more completely regularized before being put into the standard XML. DDI cubes are exactly as their original creator made them, which can be anything from completely clean to very messy indeed.

To summarize the scope of each standard in this area:

SDMX provides XML formats for describing data and independent metadata structures, which can be user-configured to hold any concepts desired. They also provide XML formats based on these configurations. The concept of exchanging a data set or a metadata set is the primary focus in SDMX, which is optimized for the exchange of aggregate data. The typical case is the exchange of time series data.

DDI also provides the ability to describe a rich set of metadata in an XML format, with an emphasis on micro-data, but also allowing for tabular formats and multi-dimensional cubes. In the 3.0 version, DDI supports all phases of the lifecycle from a description of concepts and the survey instrument used to collect data to the end product held in a data archive and used for analysis. DDI 3.0 also provides an XML format for micro-data and tabular/multi-dimensional data, but

very often the data is held in text or statistical software specific binary files. The user-configurable aspects of DDI ("variables") are mixed with specific metadata fields.

III. Process Management and the Data Lifecycle

SDMX and DDI take very different views of the data exchange process and data lifecycle. This stems from the differences in focus and scope of these standards. DDI 1.*/2.* was concerned with the data archivists' perspective: the description, after-the-fact, of all the details of a data set used for research. Historically, this was found in a print publication known as a codebook, and the XML in these versions is still a recognizable an electronic codebook. In the 3.0 version, DDI has taken a more ambitious approach in which the entire lifecycle of data collection, processing, and archiving is addressed.

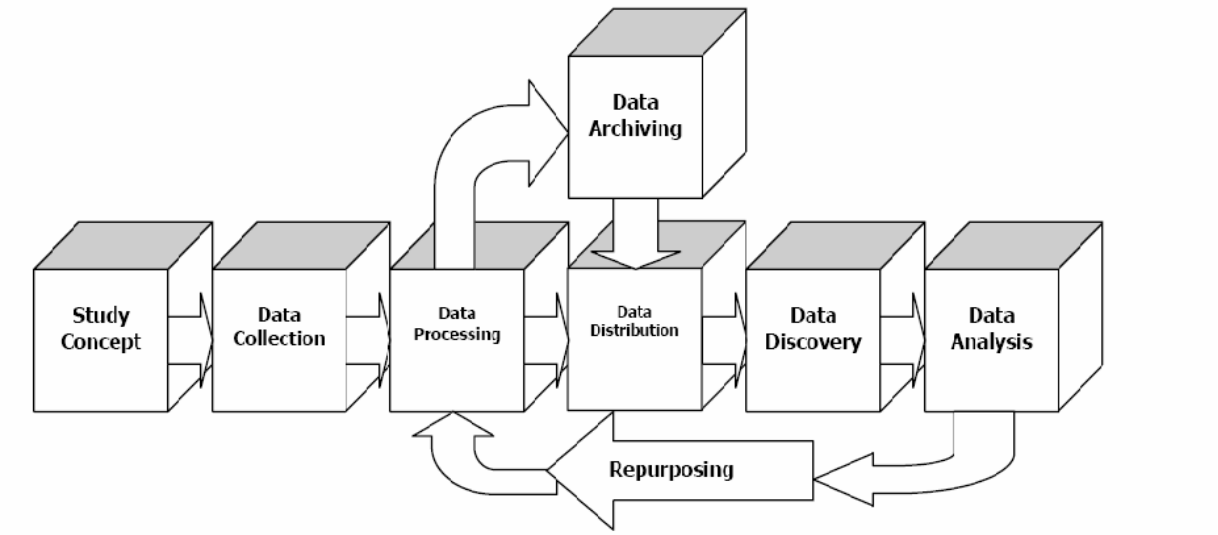


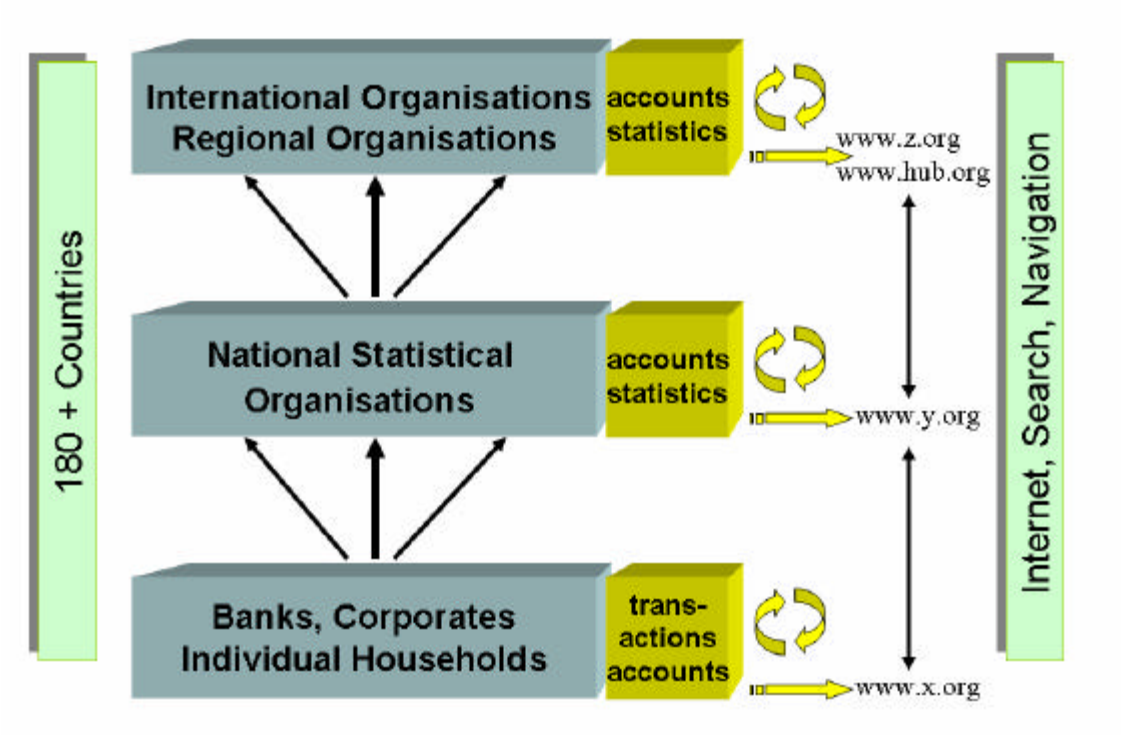
Figure: Combined Life Cycle Model

Source: DDI Alliance, Overview to DDI 3.0, 2007.

This is an abstract process model, resembling at a high level the typical stages of study design, data collection, processing/re-coding, and archiving. Support is provided for series such as panel and longitudinal studies, and after-the-fact comparison of studies is also supported. Appropriate pieces of metadata are attached to each stage of the process, and some metadata may be updated as it travels through the lifecycle.

SDMX has a different design. It does not assume any single stage-by-stage lifecycle, even at a high level. Because the flows of aggregate data do not involve the use of survey instruments and subsequent processing and re-coding, the

lifecycle approach is of limited utility within the SDMX scope. Instead, the management of regular data collection and dissemination is the focus. Thus, while the SDMX information model allows for the generic description of statistical processing, its major focus is on which organizations regularly contribute specific parts to a resulting complete data or metadata set. This includes information about advance release calendar and specific coverage of particular topics. SDMX covers the management of data reporting and dissemination as it flows through the statistical chain.



Source: SDMX Initiative, various presentations, 2002.

It is easy to see how these differences have arisen: SDMX, coming from a world of aggregate, official statistics, must deal with the regular provision of data from a large number of providers and its compilation and subsequent dissemination. SDMX supports many efficiency gains within this process. DDI, on the other hand, aims to solve a different problem: the collection of micro-data and its subsequent processing and re-processing. The primary issue for this lifecycle view is management throughout the lifecycle of a single study or a series of studies, and the provision of rich documentation to researchers who are using the resulting studies within the context of a data archive.

In many cases, both the official aggregate statistics which are the focus of SDMX and the survey micro-data which are the focus of DDI may be consumed by a single end-user, who might be a researcher, student, journalist, economist, policy maker or statistician. The differences in the standards result not from having a

different set of end users, but from having different challenges in terms of providing the end data and metadata to these users.

IV. Registries and Repositories

Both SDMX in version 2.0 and DDI in version 3.0 are designed to work with registries. Before discussing this topic, it is necessary to define what is meant by a registry. Strictly speaking, a registry is a piece of software - often deployed on a network or the Internet - where the existence of resource is located and classified, and an address provided where it can be found, external to the registry itself. This type of mechanism has been implemented in many different ways, and is a standard part of the web services technology which represents the state of play on the Internet. The standard Universal Description, Discovery and Integration registry is an example of this type.

Often, however, there is more meant by the term "registry" in the statistical or research data context that a simple look-up table of resources. The ISO 15000 ebXML[11] registry is termed a "registry/repository", which more correctly indicates its function. It serves for some resources as a registry, and for others as a repository, or centralized database. This latter model is the one which is most often meant in the realm of statistics and research data. Another very common model is the administration lifecycle model found in ISO/IEC 11179, which provides no services interfaces, but simply provides a model for the maintenance of "administered items" such as data (or metadata) elements. This model is arguably that of a repository, except that it specifies no standard implementation.

The standard registries listed above are not specific to statistics or research data. There are many examples of such registries, however. SDMX version 2.0 provides a set of standard registry interfaces, with behaviors for registry services specified. This registry has a base layer which contains structural metadata (a repository layer), a middle layer which contains provisioning metadata (another repository layer), and a top layer which contains pointers to data sets and independent metadata sets (a registry layer). At the bottom two layers, the metadata is stored in a centralized repository which is used to handle data-set and metadata-set registration in a true registry layer (the data and metadata sets are distributed around the network).

Other registry examples in the domain of statistics and research data include concept banks and question banks. These can be implemented as either registries or repositories, but are probably typically the latter - centralized databases containing the questions or concepts directly. DDI 3.0 is designed to support both of these types of "registry" applications, but there is another type of

[11] http://www.ebxml.org/

registry which is probably of greater utility in the DDI 3.0 world - a lifecycle registry. As explained above, DDI 3.0 metadata sets can be collected throughout the data lifecycle, creating a metadata set which grows and changes as it moves through the DDI lifecycle model. A registry mechanism could be used to track the distributed metadata set as it evolves across the lifecycle. This could be very useful in those cases where more than one organization or department is involved in survey design and administration, data collection, data processing, and data archiving, as is often the case.

As a final note, it should be stated that the SDMX statistical registry is one which is essentially neutral towards the standard used to mark up the data and metadata to which it points. Prototype examples exist of SDMX registries which can be used to connect SDMX aggregates with their source data, as described with DDI instances.

V. Implementation of the Standards

From the discussion above, it is clear that DDI in its 1.*/2.* and 3.0 versions is designed for very different purposes than are the two versions of the SDMX Technical Specifications. This section provides some description of the tools which implement these standards, and the ways in which the standards can be used either individually or together in some real-world functional scenarios.

A. DDI

The classic case for using DDI - especially for versions 1.*/2.*, but no less for version 3.0 - is the documentation of studies resulting from the administration of surveys. Population and agricultural censuses and household, enterprises and other sample surveys, all lend themselves to the use of DDI as an after-the-fact way that archives can document the metadata needed by researchers to make best use of the data. Such tools as the International Household Survey Network's (IHSN) Microdata Management Toolkit or the Nesstar software demonstrate how the metadata collected around a study can enormously improve navigation and understanding of the data collected.

The IHSN's Microdata Management Toolkit is an especially clear example of how this works: not only are the metadata for a study made available in a standard XML format, but they are also used to automatically generate the survey documentation in a PDF report, CD-ROM and/or website which in turn greatly facilitates the discovery and access to both the metadata and data. If public use files exist, the data themselves can also be made part of the distribution package. Further, the metadata can be used to generate the dataset for popular statistical packages, so that researchers and users can get work using their preferred software. There are many similar implementations throughout the DDI

user community, based on commercial tools and also written specifically for individual archives.

DDI version 3.0 promises to support other classes of applications as well. Some examples are: (A) upstream capture of production metadata/supporting data collection; (B) group surveys in catalogs or series, (C) manage question banks, (D) capture of metadata regarding researcher activity; (E) comparison-by-design for series of studies; and (F) mining the archive for after-the-fact comparability. While there are certainly many other cases which could be discussed, these serve as illustrative of how DDI 3.0 will likely be implemented in the short term.

1. Upstream Capture of Metadata/Supporting Data Collection

Because DDI has the ability to describe a survey instrument, designed and administered using production tools such as Blaise, CASES, or CSPro, it provides an ability to solve a problem which has long plagued data archivists: how to get the metadata captured upstream, so that the archivist need not try to reproduce the metadata after-the-fact when documenting a data set. This case can be understood from two perspectives: for the organization which is archiving the data at the end of the process, DDI facilitates upstream metadata capture, saving them a good deal of effort in getting the metadata. For the organization doing the production, this is less of a consideration. What the DDI buys them in this scenario is a way of passing the metadata between the survey instrument design stage, the survey administration stage, and the data review and editing stage. DDI 3.0 provides a standard XML format which could be useful for automating many processes in this sequence, depending on the choice of tools used for these functions. It is also the case that upstream metadata capture simplifies the problems associated with data cleaning.

2. Group Surveys in Catalog or Series

Grouping studies is one of the most fundamental needs. Data producers, archives and researchers constantly organize surveys in catalogs, sets of longitudinal studies, collections by concepts, survey families, etc. This functionality is unfortunately not available in the DDI 1/2.x specification and has been addressed in the 3.0 version through the integration of flexible grouping mechanisms.

3. Managing Question Banks

In DDI 3.0, the “question” metadata elements have been separated from the “variable” elements and can now be documented on their own. This not only maximizes reusability and provides for a rich mechanism to document questions, but also allows users to manage questions and their related codes, categories

and concepts as standalone products. This means that questions can now be documented and maintained independently of surveys which support the design and maintenance of question banks. This can also be complemented with the new survey instrument module to capture the flow of questionnaires and integrate in survey design tools as mentioned above.

4. Capture of Metadata Regarding Researcher Activity

A researcher will often take several variables from various data sets and combine them into a single "virtual" data set to support the research in hand. This process involves case selection, harmonization and derivation of new variables, re-coding, etc. DDI 3.0 provides the ability to describe this activity, so that the work performed by a researcher can be more fully documented. If such metadata is persisted, it becomes more easily possible to replicate the data from which a researcher has drawn their conclusions. Further, researchers generate a lot of knowledge about the variables they are working with as they do their research, and this knowledge - if captured - can be of use to other researchers later, when working with the same variables. A simple example of this is the process of re-coding: in DDI 3.0, a researcher performing a recode can capture not only the mapping between the code lists being used, but also the command-line processes used to actually perform the re-code. Another researcher might (especially in the case of standard code-lists) be able to re-use this metadata to automate another, similar process. This type of metadata is not always so directly useful to later researchers, but it is of value to those who want to understand the earlier analysis which was performed when working with the same data.

Note that this functionality could also be useful for data producers who often maintain multiple versions of the survey dataset (archive, licensed, public use, user specific) or may want to capture the design changes across the production process.

5. Comparison-by-Design for Series of Studies

Many longitudinal and panel studies have standard or slowly-evolving structures which are re-used as each wave of the study is conducted. In DDI 3.0, the metadata describing each wave of the study can be captured, and the evolution of the study over time can be easily viewed. This metadata is very useful in understanding the differences between the studies in a series, and also in their processing: wherever things are re-used, then it is likely that the code for processing of those same things can also be reused.

6. Mining the Archive for After-the-Fact Comparability

Especially when the archive is capturing metadata regarding the researcher's activities, it becomes possible to create applications which can look through the metadata about the archive's contents and automatically identify potentially comparable variables. One of the features of DDI 3.0 is that it requires at a minimum the identification of the concepts underlying all the questions and variables used in a study. This is the starting-point for identifying variables from different studies which are potentially comparable. Add to this the metadata generated by researchers who are performing harmonization and re-coding, and the set of information for identifying comparability becomes much richer.

These are just a few examples of how DDI 3.0 may be implemented. It should be noted that a common theme runs throughout these cases: by leveraging the machine-processible aspects of standard metadata, not only do all downstream users benefit, but those who generate the metadata can use it to make their own tasks easier. Metadata is always produced in any data collection or processing - this can either be captured or lost. DDI 3.0 takes the approach that it should be captured, shared, and leveraged for the benefit of all participants throughout the lifecycle.

B. SDMX

SDMX takes a different approach, focusing on increasing efficiencies and usability around the exchange of data and metadata, rather than on the capture and leveraging of metadata throughout the lifecycle. As mentioned above, this result in part from the fact that aggregate statistics do not pose the same challenges in terms of the lifecycle as survey data and micro-data do. It should be mentioned that there has been a strong focus on quality initiatives at the national and international level, and that while some aspects of quality in aggregate statistics are not within the scope of a technical standard, other aspects are, such as timeliness, accessibility, and usability.

One of the major cases for SDMX is the reduction of the reporting burden, both from the national level to the regional and international level, and among regional and international organizations. Often, the same data is reported many times by organizations to other organizations, and in each case a slightly different format for the data is required. SDMX addresses this issue by standardizing the formats, and by providing that the needed metadata accompanies the data. These bilateral data exchanges are often conducted using comma separated values (CSV) to format the data. Because CSV can be formatted in a large number of ways, it is not always possible to understand a data transmission without a specific knowledge of the format. Even if the formatting is evident in a similar

type of file such as a spreadsheet, the hooks for automating the intake of such files are not highly predictable, and are prone to error.

By using a standard XML format for both data and the accompanying structural metadata, SDMX makes it easier to understand the files received from counterparties. The same can be said for independent metadata, which is often made available as Word documents or in other formats which have little or no structure. This fact alone eases the burden of reporting data: it can be formatted once, in a standard fashion, and then reported to all recipients. For both sender and receiver, the difficulties of managing multiple transmissions are simplified.

SDMX also leverages the Internet to increase these efficiency gains: by providing for a centralized registry mechanism, reporters can simply put their data (or independent metadata) onto an accessible site, and then register the fact that it is there. Any counterparties can receive a notification from the registry that the data is available, and retrieve it from the location provided by the registry. This changes what has historically been a "push" technology into a "pull" technology, which is more efficient when dealing with networks such as the Internet.

There are several effects of this approach: one is to lower the reporting burden for statistical organizations at the national level: they expend fewer resources in preparing their data for multiple counterparties, and they have a simpler mechanism for delivering it to them. This results, ultimately, in more timely data reporting. Several examples of this exist today: the Joint External Debt Hub[12] (JEDH) is one example of how this mechanism can work; another is the prototype prepared by the Food and Agriculture Organization (FAO) for the reporting of data from several West African countries to the regional organization, based on the CountrySTAT and RegionSTAT tools. Other examples can be found on the SDMX web site[13].

It should be noted that SDMX places an emphasis not only on the use of the technical standards, but also on the harmonization of data structures within statistical domains. Drafts of the SDMX Content-Oriented Guidelines can be found at the SDMX site given above.

Another major case for the use of SDMX is the dissemination of data and metadata on websites. Today, it is very common to find CSV downloads available on statistical websites. While this is useful, it has the same problems as for the exchange of statistical data between organizations: the lack of metadata makes it hard to load the data into the user's systems. When looking at dissemination of data to end users, this problem is particularly acute, since it is far less likely that an individual user will be able to speak directly to the individual responsible for formatting the CSV file. SDMX provides an excellent format for

[12] http://www.jedh.org/
[13] http://www.sdmx.org

download directly from websites: the XML itself is often used for web presentation, and as an importable format it is metadata rich and easy to process. Most database platforms have native tools for working with XML.

Additionally, the existence of a centralized registry provides a wealth of metadata which can assist in the navigation of large numbers of data flows. SDMX-aware browsers are now being created which will leverage the existence of SDMX registries to provide for update notifications whenever new data on a particular topic, from a particular organization, becomes available. This reduces the need for users to continually search for the latest data in which they are interested.

The net effect of SDMX on dissemination is an increase in usability and accessibility for the end user, whether they are individuals or institutions. It should be noted that there are now an increasing number of tools being made available which provide support for the SDMX technical standards.

C. SDMX Plus DDI

Perhaps the most interesting case is one where the complementary SDMX and DDI standards are used in combination. This is often seen at the international and national levels, where users of aggregate data wish to know more about the sources of the aggregates. The SDMX registry provides a centralized place where the existence of data - whether aggregate or micro-data - can be discovered. Further, the linkages between the aggregates and their source data - or at least the metadata about the source data - can be made. Thus, I could easily provide links from the tables published off the aggregates to the documentation of the surveys which provided the micro-data inputs.

Location of Vaccination and Consultation

Location	1st Quintile	2nd Quintile	3rd Quintile	4th Quintile	5th Quintile	All Quintiles	Male	Female	All Genders
Health Centre	56.15	43.19	47.96	51.79	45.53	49.01	49.01	47.57	56.15
Hospital	16.13	23.57	27.55	25.64	30.68	27.42	27.42	25.29	16.13
Private Clinic	4.85	11.23	5.06	3.78	4.69	4.92	4.92	5.99	4.85
Mobile Unit	8.23	7.92	7.67	6.61	6.37	7.72	7.72	6.44	8.23
School	2.3	2.66	0.82	1.83	0.66	1.16	1.16	1.67	2.3
Home	12.34	11.43	10.94	10.34	12.07	9.76	9.76	13.04	12.34

Source: Nigeria Living Standards Survey (NLSS) 2004 - variable vaccgiven in Preventive, health, vaccination file.
SourceURL: http://www.nigerianstat.gov.ng/nlss/2006/index.html
DDISourceURL: hh-nga-nlss-2004-individual.xml
DataSourceURL: INDIVIDUAL HOUSEHOLD MEMBERS.Nesstar
SurveyURL: PARTA.pdf

SDMX-ML
LOC_VAC_DATA.html
Loc_vac_consult_kf.xml
LOC_VAC_GENERIC_DATA.xml
table7.xslt

The screen above shows a simple representation of what this might look like. It is taken from a demonstration of this idea which was based on material from the website of the Nigerian Statistical Agency, which today uses DDI. An aggregate table was marked up in SDMX, and links to the source data established. In this presentation, the user can easily navigate from the HTML presentation of the aggregate table to the SDMX XML file, or can go to the DDI documentation of the source data. There are also links which have been mined out of the DDI data, to the PDF representation of the survey used to collect the data, and to the data files themselves. Nesstar and the IHSN Metadata Management Toolkit both support a file format which combines the DDI metadata with the data files, and allows for easy navigation of the file with a free browser tool. In this case, the survey data are publicly accessible, so the user would have access not only to the aggregates, but also to full documentation of the survey and even the micro-data itself.

This type of functionality is based on the existence of a centralized registry where links to the various resources in this picture can be combined. Because the SDMX registry specifications allow for the linking of aggregate data files to the data and metadata files associated with the source, the existence of a standard XML format for the source metadata becomes very powerful. In essence, the SDMX registry acts as a catalogue where all of the aggregates and their source files can be discovered, and the links traversed. Note that the SDMX registry does not provide a user interface per se, but acts as a metadata resource for those who wish to build user applications which benefit from the metadata resources it contains.

It should also be noted that SDMX multi-dimensional data and structural metadata can easily be translated into DDI multi-dimensional data and metadata. The ability to perform a clean crosswalk at this point, where the standards overlap in functionality, actually provides a high degree of flexibility in their use. Users of DDI can interoperate with users of SDMX, given the appropriate software transformations. Having these two types of standard data and metadata linked inside a registry allows us to leverage this fact, and to connect aggregate data with its source data in a standard way which has never before been easily possible.

VI. Summary

It should be clear from the discussion here that DDI and SDMX are standards which are related, but which are not in competition. They are very different in scope: where DDI is aimed at solving problems with the documentation of research, and across the micro-data lifecycle, SDMX is concerned with creating efficiencies around the exchange of aggregate data. DDI comes from the world of

data archives and social sciences researchers; SDMX springs from the world of official statistics.

The fact that these two standards are well aligned means that they can be combined in powerful ways, and that users of the two standards can move data from one standard format to the other fairly easily. The choice of which standard to use depends on the focus of the organization which is doing a standards-based implementation. Hopefully, the presentation of the various implementation cases above serves as a guide to how each standard can be applied. It may well be that both standards are useful within an organization which deals both with micro-data and aggregates.