IHSN technical note on metadata standards - DRAFT

Introduction

Discovering, accessing, and using statistical microdata, and publishing resulting analytical outputs, are fundamental elements of socioeconomic or health research. Provisioning services and tools that support and facilitates such activities requires surrounding the data with relevant infrastructure and comprehensive information. These aspects are key to enabling the delivery of data documentation to the users, the automation of processes or information flows, and knowledge sharing.

The purpose of this document is to provide a high level overview of domain specific metadata standards, underlying information technologies, and related best practices, to enable establishment of data management and processing infrastructures in support of effective research. We particularly focus on the Data Documentation Initiative (DDI), a widely used specification for documenting and managing socioeconomic and health micro-datasets.

One of the main objective is to highlight the benefits of adopting and leveraging metadata standards and related information technologies for (1) facilitating broad dissemination and sharing of data by producers and custodians and (2) providing tools to researchers for documenting and publishing secondary data and publications, when applicable in accordance to data management plans.

Metadata

Overview

Metadata, often defined as “data about the data”, is a term broadly used to refer to information that adds significance, context, or knowledge to existing information. Whenever we need to make any kind of informed decision or to

take an action, we want to have access to as much information as possible to ensure that our choice is the right one. We constantly use metadata in our everyday life: buying goods (brand, price, nutrition fact), ordering food in a restaurant (the menu), purchasing a car, dressing-up based on weather forecast (metadata about cities or regions), investing in stock market, and many other situations. Without metadata, we would be ill informed and unable to make sound and effective decisions.

In the case of socioeconomic or health statistics, metadata concerns complementing statistical datasets with comprehensive documentation, imparting meaning to the quantitative or qualitative information, to ensure access, effective and responsible use, preservation, and sharing.

For microdata, this goes beyond the simple data dictionary. Metadata are intended to help researchers understand what the data are measuring and how they have been created. Without a proper description of a survey's design and the methods used when collecting and processing the data, there is a significant risk that the user will misunderstand and even misuse them.

Good documentation also reduces the amount of user support statistical staff must offer external users of their microdata. Metadata are also intended to help users assess the quality of data. Knowledge of data collection standards – as well as of any deviations from the planned standards – is important to any researchers who wish to know whether particular data are useful to them.

Lastly, metadata are needed to develop data discovery tools, such as survey catalogues that help researchers locate datasets of interest. Note that data not intended for public dissemination must also be fully documented. Producing good metadata helps build the institutional memory of data collection, and can assist in training.

Comprehensive metadata content includes:

- Explanatory Materials: data collection methods, dataset structures, technical information, data dictionary, classifications, derived variables, weighting/grossing, data sources, confidentiality/anonymization
- Contextual Information: project description, topical / geographical / temporal coverages, provenance, series/time-series
- Cataloguing Materials: bibliographic records, citation, keyword/concepts, other search discovery attributes

Kinds of Metadata

Metadata can be broken down in various flavours:

- Structural metadata describe the structure of data sets, whether these are tabular in nature or simply files of raw data or microdata. Which variable's value appears in which column? Which row represents which case? Are there hierarchical relationships? Etc.
- Reference metadata (also known as "descriptive" metadata) consist of what is often thought of as "footnote" metadata, whether this is about methodology, sampling, quality measurements, production notes, etc. This is a very broad term, which can cover a range of information, regarding everything from single data values to entire collections of data.
- Administrative metadata are the data created by the process of administering data, in their collection, production, publication, or archiving.
- Behavioral metadata (also known as "paradata") is information about the reaction and behavior of users when they are working with data, and respondents while data are being collected (in this case, it is paradata about a collection instrument). This can be of interest to those who act as data librarians – to help them better manage their data collections – but can also be of direct interest to researchers – what did other researchers do with the data? How did respondents react when asked a question?

Sound metadata standards and management systems must provide mechanisms to broadly capture all these kinds of metadata.

Metadata & Information Technology

Metadata is a very natural part of most modern IT infrastructures, given the strong focus modern technology places on information. If technology depends on the exchange and use of information – or data – then the metadata describing that information can be very critical in the creation of systems which perform tasks in an automated way.

The rise of the Internet in the late 90's has lead to the need for organizations to effectively exchange information with each other. This necessitated the

availability of a common language and harmonized architecture to isolate proprietary information systems from each other. It also called for mechanisms to describe standardized information models and ensure the description of data in a broadly understandable way.

This has resulted in the establishment of the Extensible Markup Language (XML) and Service Oriented Architecture (SOA). The combination or XML and SOA can be used to wrap existing systems or infrastructures with an implementation neutral layer or directly implement solutions based on the standards, enabling the publication and exchange of information in a standard way. This technology suite has been around for over a decade, is very mature, and widely recognized as an IT industry standard practice. It has played an essential role in the success story of the Internet.

Importantly, XML is open and free for all to use. The XML standards themselves are maintained by the World Wide Web Consortium and publicly available. This implies that XML not only provides a common language and facilitates metadata management but also is also easy and cost effective to adopt as a technology. XML documents are stored in a standard text format, ensuring that anyone can read them and long term preservation.

About XML

As its name implies, XML is first a language, which means it, comes with syntactical and grammatical rules that describes its syntax. These are pretty simple in nature but ensure that we can ensure that an XML document is properly written or “well-formed”.

XML is however not only a language but also a collection of technologies available to perform various operations on the underlying data or metadata: XML schema, for describing document structure; XPath and XQuery for querying and searching XML; SOAP or REST to facilitate the exchange of information; and many others. Together, they provide a comprehensive suite for managing metadata.

The ability to validating an XML document against the rule of the language and structure is one as the many benefits of XML, as it helps ensuring that information exchanged between agencies meets the agreed upon formats..

Human Readable and Machine Actionable

It is important to note that metadata are equally useful to individual beings and computer systems.

In the first case, the metadata are typically descriptive in nature and need to be delivered in user-friendly formats (human readable), such as documents, web pages, or printed publications. Their purpose is to share and deliver knowledge.

In the second, metadata is formatted in a computer friendly form with well-defined structure and organized content to enable processing by software (machine actionable). The purpose here is to enable efficient task automation and business process management.

As it turns out that converting data or metadata from XML into human readable formats are operations that can be software automated, using XML as a storage format caters to both needs (which is not true the other way around).

Metadata are commonly perceived as a mechanism for providing documentation to users. Combining high quality human readable and machine actionable metadata however is key to unlocking the core features. Machine actionable metadata are essential for automating processes, which is one of the most important benefit of adopting metadata standards and technologies. One of the challenges is therefore to ensure quality and comprehensiveness of the metadata in such regards.

Metadata Standards

In order to exchange information between systems, agencies, or individuals, we need to agree upon common descriptive structures. For example, describing a book commonly involves capturing a title and author but may or may not involve an ISBN number, number of pages, or format. Some elements of information may be required while others could be optional or repeatable. Such attributes also have a “type” that can be as simple as a string, number, range, date, or be itself a complex structures. As noted above, XML provides all the necessary mechanisms to achieve this using for example

XML schemas. The resulting model is then often referred to as a "specification" or "standard".

It is important to note that not all standards are created equal and the term "standard" is used to describe specifications at different scope level. We need to distinguish between:

- International/Industry Standards: widely recognized and endorsed by agencies around the globe, preferably approved by organizations such as ISO, OASIS, or W3C.
- Regional Standards: similar to international but not particularly globally used or endorsed
- National Standards: country specific standards commonly used by local organizations or endorsed by national agencies (i.e NIST)
- Institutional Standards: used within the confinement of a specific agency or consortium (i.e internal classifications, information systems, etc.)
- Individual Standards: commonly out of the mind on a single or small groups of individuals, to be used within a project (and should not really be called a standard)

When selecting a standard, it is critical to ensure that it is fit for purpose. For the management of statistical data, globally or widely accepted ones are clearly the preference. Sustainability and openness of a specification are also important criteria.

Developing and agreeing on standard model can be a challenging process, widely depending on the complexity of the objects being described, and the scope of the standard, the organizational processes, and often the number of agencies involved. Different domains and entities of course call for different models and there are therefore many different XML specifications. For example, there are standards for news feeds (NewsML, NITF, RSS), documents (OpenOffice or Microsoft OpenDocument), US Patents, individual's resume/CVs, books, stock quotes, geographies, and many others. Most of the XML components are actually themselves XML standards (XSLT, XSchema, etc.) - just like the English dictionary, XML is written and defined using XML.

When it comes to socioeconomic data and official statistics, a few standards have emerged in the last decade to describe the data and its related processes. The two core specifications are the Data Documentation Initiative (DDI), mainly concerned with microdata or low-level administrative data, and the Statistical Data and Metadata Exchange Standard (SDMX), focusing on

the publication and exchange of aggregated data, indicators and time series. These have been designed to complement each other (aggregated data often finding its source in microdata) and work hand in hand with other standards such as ISO 11179 (data elements, value domains, classifications, metadata registries), ISO19115 (geography), Dublin Core (generic resources), and others. Together they form a comprehensive collection of metadata specifications for the management and use of data.

Detailed information on the DDI is provided in the following section. Its relationship to SDMX, other standards and technologies (such as RDF), and related activities (harmonization, open data, linked data, etc.,) and then briefly discussed.

Mappings and Transformations

Note that it is not uncommon for multiple standards to exist within the same domain, sometime competing but most often complementing each other. This is not particularly an issue as each can cater for different needs within the domain. DDI and SDMX are good illustration of complementing standards. DDI-Codebook and DDI-Lifecycle, the two flavours of DDI, are an example of variations for different purposes.

As the data lifecycle can be long and complex or involve many different entities, it is also possible for metadata to cross-domain boundaries. In which case, it may need to be converted between specifications or reshaped to provide a different view or perspective.

XML is also one of several formats for serializing or storing metadata. Other formats, such as RDF, JSON, SQL databases, CSV, and others can be used.

XML natively provide mechanisms to support such requirements and convert or “map” metadata across standards or formats (i.e. using transformations).

Data Documentation Initiative (DDI)

Overview / History

The Data Documentation Initiative (DDI) is an XML specification for describing and facilitating the management and processing of microdata. DDI emerged in the late 90's from the socioeconomic data archiving community, as an instrument to document surveys and censuses datasets. The Interuniversity Consortium for Political and Social Research (ICPSR) in the USA, one of the world's largest data archive, was instrumental in its inception. The DDI was rapidly adopted by other archives across North America and Europe, in particular by members of the Council of European Social Science Data Archives (CESSDA) community. The Nesstar software, developed by the UK Data Archive and the Norwegian Social Science Data Services (NSD) with the financial support of the European Union, played an essential role in its initial success and adoption.

In 2004, the DDI Alliance, a membership based organization, was established as the governing body for the specification. It counts today over 35 members agencies. The Alliance at that time recognized that the existing specification, whose mandate has primarily focused on the data archivists, was not meeting the needs of early data production stages, post-dissemination activities, or complex survey programs such as longitudinal studies. It also suffered from some technical weaknesses and flexibility in terms of global identification and reuse of metadata elements. This had slowed the adoption of the DDI by large statistical agencies whose complex requirements called for a more technical standard. The call was therefore made for the specification to cover a broader range of features, including managing early survey design stages, questionnaires, classification banks, and others activities. DDI at that time was at version 2.1 and the Alliance technical experts group initiated work on a next generation DDI version 3, whose design revolved around a data "lifecycle", ranging from early study conceptualization by producers into the use and repurposing of the resulting data by researchers. This took nearly four years of efforts and DDI 3.0 was released in April of 2008. Version 3.1 was since published, with 3.2 expected later this year. To minimize some confusion around the solely numbered versions and clearly differentiate between the two streams of DDI, the Alliance has named the DDI 1.x-2.x versions DDI-

Codebook (DDI-C) and DDI 3.x versions DDI Lifecycle (DDI-L). Version 2.5 of DDI-Codebook was recently released, adding some features at the request of key stakeholders and facilitating the back and forth conversation into DDI-Lifecycle.

DDI has become a global standard for documenting and managing socioeconomic, health, and other microdata. While this is being considered and actively discussed by the DDI Alliance, the DDI-XML specification is currently not an ISO standard.

IASSIST

The International Association for Social Sciences Information Services and Technology (IASSIST), an international organization of professionals working in and with IT and data services to support research and teaching in the social sciences, provides a space for archivists and technologists meet has naturally become the home of the DDI community. The IASSIST annual conference is traditionally the annual conference around DDI, with the DDI Alliance meeting taking place around the event.

DDI-Codebook (v1.x-v2.x)

Versions 1.0 through 2.1 and the recently published version 2.5 of the DDI are referred to as “DDI-Codebook” or “DDI-C”. These XML specifications focus on the documentation of variables contained in microdata files and surround them with contextual metadata such as descriptive information on the study that lead to their creation and reference documentation such as reports, questionnaires, technical manuals and the likes.

A DDI-C document can describe complex datasets composed of multiple files, including their primary keys and relationships. It is fundamentally a very rich “codebook”. At the variable level, in addition of capturing information commonly available in a traditional data dictionary (variable name, label, format, value labels), DDI-C can describe various characteristics such as a definition, question texts, interviewer instructions, universe, weights, derivation/imputation, or summary statistics (min/max/stddev/...), and others. DDI-C captures extensive information on the survey itself - such as title, abstract, data collection, sampling and other methodologies, stakeholders, access policies, and contact information - and for further details can reference external resources such as reports, questionnaires, and technical documents.

While not often used, DDI-C also has the ability to describe data tables, or “cubes”, when constructed directly from the survey variables (as table dimensions). Altogether, the DDI-C metadata captures in a single document all necessary information to access and make effective use of the data for various analytical purposes. In addition to provide invaluable information to the users, it’s XML nature enables processing by application for integration in information systems and process automation.

The various versions of DDI-C that have been published since 2000 reflects various enhancements made to the specification based on user demand. The most commonly found version of DDI-Codebook is 1.2.2 as it is the one generated by the popular Nesstar and IHSN Microdata Management Toolkit software. Version 2.5 of DDI-C was recently released to include new metadata elements and facilitate bi-directional conversion between DDI-Codebook and DDI-Lifecycle.

Today, DDI-Codebook is the most commonly found format of DDI-XML metadata. The specification has been around for over a decade, is very mature, and software such as the Nesstar, the IHSN Toolkit, IHSN NADA, and others, greatly facilitate its adoption and use.

DDI-C Users

As previously mentioned, the first institutional group of DDI-C adopters were data archives across North America and Europe, lead by ICPSR and the Minnesota Population Center in the US, and members of the CESSDA across Europe. These agencies continue today to leverage the specification and play a leading role in its direction.

At the international level, the World Bank became an early adopter in 2000, initially to support internal and country level projects such as the Africa Household Survey Databank and the Data Development Platform. Following the inception of the International Household Survey Network in September 2004 as a recommendation of the Marrakech Action Plan for Statistics, the World Bank collaborated with Nesstar Ltd. to enhance their Nesstar Publisher tool, rebrand it as the IHSN Metadata Editor, and complemented it with open source utilities to provide a comprehensive data documentation and packaging software suite, which resulted in the Microdata Management Toolkit (MMT). Through the World Bank / PARIS 21 Accelerated Data Program (APD) and other initiatives, the Toolkit was rolled-out into national statistical offices and other statistical agencies in developing nations. The

software and related best practices became very popular and has since been deployed in ?? countries around the globe, leading to thousands of surveys being documented using the DDI. It has also been adopted by other international organization or institutions.. As the Nesstar Publisher became freeware, the suite is now completely free of charge. In 2009, to complement the MMT and provide a web based data and metadata dissemination solution, the IHSN developed and released the National Data Archive (NADA) package, a pHp based open source application leveraging DDI-Codebook XML to populate and manage the catalog. NADA has been deployed in numerous countries and organizations, making hundreds of surveys available to researchers. The World Bank itself using it for the dissemination of its microdata.

DDI-C Tools

For DDI-Codebook, the most commonly used packages for managing the data and metadata are the Nesstar software suite and the IHSN Microdata Management Toolkit. Together, these tools are used in hundreds of organizations around the globe, and have been instrumental in the success of DDI. The IHSN NADA package complements these packages by providing a lightweight web based catalog for the discovery, search and retrieval of data and documentation. DDI-Codebook is also supported by packages such as Berkeley SDA tabulation engine and several utilities have been developed to facilitate using DDI with common statistical packages.

The free availability of essential tools to produce, maintain, and publish DDI-C makes this version of the standard fairly easy to adopt by anyone.

DDI-Lifecycle (3.x)

DDI-Lifecyle is the advanced version of DDI specification designed to meet the needs of large agencies, complex datasets, and enterprise grade metadata management systems, and use case beyond DDI-C. It significantly expands the sets of features available in DDI-C but in doing so introduce some level of complexity in terms of specification, management techniques, and tools.

Some of the initial requirements of DDI-L included:

- Improve and expand the machine-actionable aspects of the DDI to support programming and software systems
- Support CAI instruments through expanded description of the questionnaire (content and question flow)
- Support the description of data series (longitudinal surveys, panel studies, recurring waves, etc.)
- Support comparison, in particular comparison by design but also comparison-after-the fact (harmonization)
- Improve support for describing complex data files (record and file linkages)
- Enable the maintenance of concept, universe, question, classification or variable banks
- Provide improved support for geographic content to facilitate linking to geographic files (shape files, boundary files, etc.)

DDI-L has a strong emphasis on reuse and aligns on the traditional IT principle that an element of information should only be documented once (also known as third normalized form in the database world). Identification mechanisms also ensure that essential metadata elements are assigned a globally unique identifier (a URN), facilitating global publication on the Internet. As DDI-L metadata is stored in well-defined isolated containers (called “schemes”), it enables distributing the burden of maintaining the metadata across multiple agencies/departments/users throughout the various stages of data production, archiving, dissemination, and analysis.

DDI-L Adoption

DDI-Lifecycle is currently seeing rapid adoption, particularly in agencies managing large or complex data collections or needing to integrate solutions in enterprise IT infrastructures.

Several initiatives are ongoing, some involving leading agencies such as the Australian Bureau of Statistics, Statistics New Zealand, the Canadian Research Data Centre Network in collaboration with Statistics Canada, GESIS and several other institutions in Germany. The Data without Boundaries project, involving numerous agencies across Europe, is planning to leverage DDI-L to support a cross-country research infrastructure. As a result, both custom and generic tools have begun to emerge.

DDI-L Tools

From the tools perspective, DDI-Lifecycle is in an emerging phase. While applications are becoming available, most are still at the prototype/testing level or are institutional implementations whose usage outside the agency is not particularly possible or easy to achieve.

In terms of commercially supported tools, Algenta Colectica# was the first product available around DDI-Lifecycle. Since version 11, Stat/Transfer#, a popular data conversion package, support DDI-Lifecycle XML as a metadata export format. Metadata Technology North America will be releasing later this year the OpenMetadata Framework#, an open source platform for implementing tools around metadata (DDI and others). A free light DDI Editor will be made available as an example of a desktop tool implementation. Other tools, utilities, and services around DDI are expected to become available under the OpenMetadata.org portal.

As DDI-Lifecycle was designed to meet numerous use cases, a great variety of tools meeting different needs will emerge in the very near future, with DDI-L as a common language... For the time being, adopting DDI-L often requires developing new tools or extending existing systems, and is therefore a choice that requires due diligence.

The DDI Alliance maintains a list of available packages and utilities availble on the web at http://www.ddialliance.org/resources/tools

DDI-C and DDI-L

Both DDI-C and DDI-L provide powerful features enabling the effective management, access, and use of statistical microdata. Beyond the traditional archiving/dissemination/documentation use case, DDI-L is by design the right choice. In the area where both specifications overlap, the decision may be determined by several factors, including:

- Tools availability: DDI-C comes with easy to use software, DDI-L on the other hand will likely require some level of development and investment by the user (at least for the time being)
- IT capacity: though not required, DDI-L commonly operates in an enterprise IT environment involving client/server infrastructure. DDI-C users can operate from a simple desktop.

- Learning curve: DDI-L specification is complex and requires the user to understand its underlying principles; DDI-C is structured much more like the traditional codebook or data dictionary, and users find it easier to understand.
- Number of Maintainers: DDI-C is typically maintained by a single agency, DDI-L facilitates distributing the load or reusing existing metadata
- Alignment on SDMX: as further discussed below, DDI-L has been designed to work hand-in-hand with the SDMX standard and overlaps in technical design.
- Enterprise Integration: DDI-L is a better fit for integration in large or complex information systems, especially because it emphasizes reuse of information.

It is also important to recognize that both specifications can work hand in hand and are not at all mutually exclusive. Many DDI-L users today are actually transitioning from a DDI-C environment and likely to continue to use DDI-C tools for some time. Transforming metadata back and forth (when relevant) is a fairly trivial task, which is now made easier thanks to the availability of DDI-C 2.5, which has features designed to enable transformation into DDI-L. It is also fairly common for new user to start with DDI-C (using today's tools) in anticipation of migrating to DDI-L in the near future. It is therefore important for to keep the option open as such approach can deliver the best of both worlds.

DDI and Health Data

[this section needs work / ideas]

While DDI was for long primarily used around socioeconomic data, increasing demand has emerged in the recent years from the health sector for documenting surveys and other data collections. In particular, he UCL Institute for Child Health# (ICH) in the United Kingdom has piloted DDI-Codebook with selected survey under a Secure Epidemiology Research Platform# (SERPent) pilot project (as a proof of concept). ICH also expects to used DDI-Lifecycle to support the recently initiated UK Birth Cohort Study#. In the USA, NORC at the University of Chicago is using DDI for supporting work around big data from the Center for Medicare and Medicaid Service in its virtual data enclave#. At a recent EUCONNET meeting (a collaboration of child cohort studies in the health research area, funded by the European Union) in

Edinburgh, several of the world's leading child cohort studies met and specifically discussed metadata and data management issues. Of these studies, several were starting implementations of DDI-L, notably at the University of Essex in the UK and the University of Bamberg in Germany. In the UK, the Medical Research Council (MRC) recently decided to use DDI-L as its standard for collecting metadata for its discovery portal, the MRC Gateway, gathering information from a range of health-related longitudinal studies in the UK. (Veerle Van den Eynden, MRC Data Support Service, "Supporting the sharing of longitudinal health data", presentation at IASSIST 2012).

The upsurge of interest in DDI-L within the health research community, particularly for longitudinal studies, is an interesting phenomenon. While there are standards specific to such health-related subjects as clinical trials (CDISC is the most popular) these standards are not designed specifically to support data management across the lifecycle. For this, DDI-L seems to be a better tool. In addition, health research and sociological research are starting to incorporate data which in the past would have been considered out of scope: for studies such as the MIDAS study on aging in the US - which is managed using DDI-L - the data sets now incorporate bio-markers and MRI scans of respondents, data which traditionally belonged in the health-research domain. Health research has started to incorporate more and more data about lifestyle, which traditionally has been considered in the sociological domain. It is thus perhaps not surprising that DDI-L is of increasing interest among the health research community.

Key Benefits

This section highlights some of the key benefits of adopting the DDI, metadata standards, and XML technologies for statistical microdata management.

Comprehensive Documentation

Microdata are often very complex. Effective and responsible use for research or other purposes requires not only a deep understanding of the data themselves but also extensive familiarity about the processes and methodologies used for their creation. DDI provides a rich set of elements to capture such information, which can be used for both delivering documentation in user-friendly format or for automating analytical processes. The conversion of the information from the XML format into web pages or

documents can easily be achieved by leveraging XML native technology such as XSL transformations.

Quality

Using a standard such as DDI has significant impact on many of the quality dimension so the data such as:

- Accuracy/Consistency: Compiling metadata around microdata often require performing various quality assurance procedures, which often leads to the discovery of inaccuracies, subsequently resulting in an overall improvement in terms of data consistency and accuracy. Ideally, metadata should be leveraged to drive or support data production, further minimizing the risk of error and facilitating overall quality control.
- Accessibility: DDI metadata greatly facilitate discovery and access to the data. The availability of comprehensive metadata enables search engines to deliver a rich set of effective functionalities.
- Timeliness: leveraging information technologies and machine actionable metadata enables process automation. Metadata driven data management system can rapidly executed tasks that would otherwise be time consuming and error prone if performed manually by individuals

The combination of the above addresses numerous issues contributing the overall usefulness of the data.

Reusable tools / architectures

Standards based tools and IT architecture has the major advantage to be reusable. Just like formats like PDF or Microsoft are de facto standard for documents, the DDI is emerging as the best practice around microdata. As a significant portion of the software being developed around the DDI specification has tendency to be open source, it greatly facilitates reuse and fosters collaborative efforts for reuse, enhancement, or extensions. It also reduces the overall implementation costs and promotes sustainability and transparency.

Packaging for preservation, publication, and exchange

The vast majority of statistical packages and database software use proprietary file formats that are not particularly compatible with each other and

require the purchase of commercial licenses. Users often struggle and spend a significant amount of resources in simply opening, importing, or converting data. These formats are also not particularly a good fit for long term preservation as there is no guarantee that the software will continue to be around for decades to come or be backward compatible. The actual binary encoding of the file is also often kept behind closed doors and protected by IP or copyrights.

To address such issues, ASCII text is typically used as generic format for dissemination or long term preservation. The major drawback of ASCII data however is that they carry no or very little metadata, which result in significant loss of information and usefulness. While the data is now openly accessible, users need to invest resources in recreating knowledge such as the data dictionary, offsetting some of the benefits.

Combining ASCII with DDI-XML alleviates such issues and provides the best of both worlds: open data with rich metadata that goes beyond what proprietary formats can offer. Importing the text data into statistical or database packages for processing or analysis can largely be automated by leveraging the XML to generate the necessary ingestion scripts or programs. Such script generator commonly only needs to be implemented once per package, possibly as an XSL transformation or program. ASCII+DDI therefore can be used as a canonical format for long term preservation and the packaging of datasets for publication or exchange. It enable their use with a wide range of software, further resulting in significant resource saving for the users.

Automation

This aspect is inherent to the adoption of XML technologies in general but is very significant in the case of statistical data. The ability to have tasks performed by software applications can have a tremendous impact on data quality. As most of today’s tools and statistical packages are not particularly metadata rich or aware, numerous operations need to be manually performed, requiring costly resources investments, being prone to human error, and taking time to complete. The availability of well-structured machine actionable metadata enables replacing these often simple or repetitive tasks with efficient tools, resulting in higher quality and timely data. Metadata such as DDI also enable the integration of these tasks in business process management systems, resulting in effective data workflows.

Harmonization, Comparability, and Linked Data

Bringing data together across waves or from multiple sources is a very common requirement, particularly for researchers. This cannot sensibly be achieved without a deep understanding of the data and how they compare across sources. DDI metadata provides comprehensive information to support and enable such processes. It first does so by delivering extensive documentation to the user. But more importantly, the metadata can also be used to partially, if not fully, automate the necessary data transformation processes. Without metadata such as DDI - as it if often the case today - users often need to painfully search for and extract this information from data dictionaries and technical documentation, perform significant validation, and develop necessary scripts and programs, resulting in huge time and effort investments. Such process is often repeated over and over again as the harmonization or transformation processes are not attached to the data.

Enhanced Publication, Replication, Citation

An important aspect of research is concerned with the validation of the results as well as potential reuse of the new data to further the ensuing knowledge. Too often though, the published output of a research project is limited to a paper, with the underlying data being often inaccessible if at all archived and preserved.

Packaging data alongside the publication has long been advocated as a recommended practice. Gary King replication standard# for example calls for any research output to hold that “sufficient information exists with which to understand, evaluate, and build upon a prior work if a third party can replicate the results without any additional information from the author." This combines with the idea of “enhanced publication”**#** advocated by SURF whereby publications are linked with additional material, like research data, models, algorithms, illustrative images, metadata sets or post-publication data such as comments or rankings. Such obligations are also becoming critical for researchers as an increasing number of funding agencies rightfully require to accompany research proposal with a data management plan. This is nowadays further facilitated by initiatives such as DataCite# and the underlying Digital Object Identifiers# (DOI), which became an ISO standard last May#

DDI metadata can enable all of the above by providing mechanisms for packaging data for publication and surround them with identification and

extensive knowledge, in a standard format, facilitating preservation, dissemination, and sharing.

This is essential to support a world wide web environment that is becoming more and more data aware and centric, and initiatives such as data.gov

Statistical Data and Metadata Exchange (SDMX)

The Statistical Data and Metadata Exchange (SDMX) technical specifications come out of the world of official statistics and aim to foster standards for the exchange of statistical information. Because of the nature of these data, the focus is on aggregated statistics, indicators, and time series like macrodata. Unlike for microdata, the structure of these datasets is fairly consistent and predictable. 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.

The SDMX initiative is a cooperative effort between seven international organizations: the Bank for International Settlement (BIS), the International Monetary Fund (IMF), the European Central Bank (ECB), Eurostat , the World Bank (WB), the Organization for Economic Co-operation and Development (OECD), and the United Nations Statistical Division (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 are registered with ISO as Technical Specification, ISO-17369.

SDMX has several 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.

SDMX is currently primarily in use by international organizations, central banks, and selected national statistical agencies.

SDMX and DDI

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 such as ISO11179. Because much of the microdata described by DDI instances is typically aggregated into the higher-level data sets found at the time-series level, this is not surprising. Although there is some overlap in their descriptive capacity, they can best be characterized as complementary, rather than competing. One focuses on microdata, the other on macrodata.

The most obvious overlap between the standards is in the description of data tables, commonly referred as “cubes” due to their multidimensional nature. There are however 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. In addition, DDI describes tabular that are directly derived from the underlying microdata. SDMX on the other hand can capture data cubes from any data source, microdata, processed microdata, administrative data, and other.

Additional overlap exists in the way DDI and SDMX describes concepts and classifications. The technical design of the DDI-Lifecycle also significantly draws from SDMX, in particular in regards to identifiers (URNs) and schema design.

A significant difference between the two specifications is that SDMX actually carries the data while DDI describes the variables whose data is stored in external files (ASCII, statistical packages, databases).

An important benefit of using specifications both together is the ability to maintain linkages between published tables holding aggregate data derived from microdata. A cell in a time series table can in that case directly or indirectly be related to one of more variables in an underlying survey. These relationships can be captured in SDMX and DDI and provide critical information to researcher, enable the rapid navigation from the aggregated to the microdata source, enable on the fly table computation, or facilitate data mining.

There is currently a project to coordinate development between the two standards bodies. Facilitated by UN/ECE, the “SDMX-DDI Dialogue” is an effort to establish how best the two standards can work together. This project is creating a joint SDMX-DDI vocabulary, to make it easier for users to work with both standards, and is researching topics such as how DDI and SDMX can work together in a microdata access scenario, in teh reporting and collecting of register data, and in a number of other use cases. The DDI Alliance and the SDMX Sponsors have endorsed the use of the two standard for data management across the data production lifecycle. (http://www1.unece.org/stat/platform/display/metis/SDMX+DDI+Dialogue+-+Overview+Page)

There is also much interest in how the two standards can be used to support the Generic Statistical Business Process Model (GSBPM), which is a popular reference model developed under the UN/ECE’s METIS workshop on statistical metadata (http://www1.unece.org/stat/platform/display/metis/The+Generic+Statistical+Business+Process+Model). This model was based in part on the DDI Lifecycle model published as part of the DDI-L specifications. Currently, there is additional work going on under the auspices of the High-Level Group on Business Architecture in Statistics (HLG-BAS), a committee of the Conference of European Statisticians (CES), to develop the Generic Statistical Information Model (GSIM). This model - expected to be published early in 2013 - incorporates both DDI and SDMX, and has started work to show how SDMX and DDI could be used in implementations of the model (http://www1.unece.org/stat/platform/display/metis/Generic+Statistical+Information+Model+%28GSIM%29).

Challenges

Both DDI and SMX represent powerful tools for data management, sharing, dissemination, and use. However, their implementation does not come without

its own challenges and costs. As with many types of IT, there is typically an up-front investment required of organizations which implement them. It should be understood that these challenges are not only technical in nature.

With both SDMX and DDI, we are beginning to see both commercial and open-source tools which can help implementers to solve the technical challenges. However, the standards require a culture of metadata and data management which does not always exist in organizations today.

First, many organizations manage metadata in an ad-hoc fashion, letting each department capture and store metadata in a different fashion. This makes it difficult for all the departments of an organization to agree on a single standard way of managing this content. Typically, much effort will be required to standardize the metadata management practices across an organization.

Having selected a standard approach to metadata and data management, all the staff of an organization must be educated in the new approaches. Whether this is done with SDMX, DDI, or a combination of the two, there is a skill-set required to work with new tools and to learn what may be unfamiliar terminology. This must be incorporated into the organization's training programs for its staff.

Finally, there is the issue of change management. Re-working of data management systems can cause many types of change to an organization's operations: typically, this involves the elimination of "silos" within the organization. If all data and metadata are managed in a standard way, then they become visible across an organization. While there are many benefits to this from the perspective of the organization's management, it represents a change in the way resources are allocated and data management functions are performed. This can drive changes to the implementor's internal organization. These changes can be problematic, depending on how an implementing organization operates. Outlining an non-intrusive transition strategy minimizing impact on ongoing processes is essential to ensure successful institutionalization of metadata.

Conclusions

[to do]