The Data Documentation Initiative (DDI): An Introduction for National Statistical Institutes

Arofan Gregory, Open Data Foundation
July 2011

Contents

I. Introduction

This paper introduces DDI to those coming from national statistics institutes (NSIs). While there is a large amount of information regarding DDI available today, sometimes it is difficult to know where to start, and much of it comes from domains which are not familiar to those working with official statistics. Here, we attempt to characterize the flavors and uses of DDI, give some general background on the standards organization (the DDI Alliance), describe available tools, and relate the DDI to other initiatives and standards which are more familiar to this audience.

II. The Different Versions of DDI

DDI has two major development lines: the original DDI (versions 1.0 - 2.5, now called "DDI - Codebook") and the lifecycle-based DDI (versions 3.0 - 3.1, now called "DDI - Lifecycle"). These two lines have different capabilities, which are described below. Both are able to describe microdata sets and their tabulation into aggregate data cubes, with a wealth of related metadata (questions, variables, concepts, categories, codes, etc.).

A. DDI - Codebook: The Original DDI

Originally created by data archives to document the data sets which they archive and disseminate to researchers, DDI – Codebook includes the same information as a traditional codebook, describing variables, question text, and the categories and codes used as response domains and the values of variables. It also captures some other information about the data set ("study" in DDI terms). The data themselves may be held in statistical package formats or in ASCII with enough information in the DDI metadata file to understand which values are associated with which variables. The metadata are structured in an XML format.

With this information, it is possible to generate good documentation after-the-fact from a data file, to provide good search capabilities for data based on the rich metadata, to feed tabulation engines, and to generate set-up files for statistical packages. Tools also exist to read statistical package files and generate DDI documentation and corresponding ASCII data files. There is also a freeware editor for DDI 1/2 which is widely used: the Nesstar Publisher.

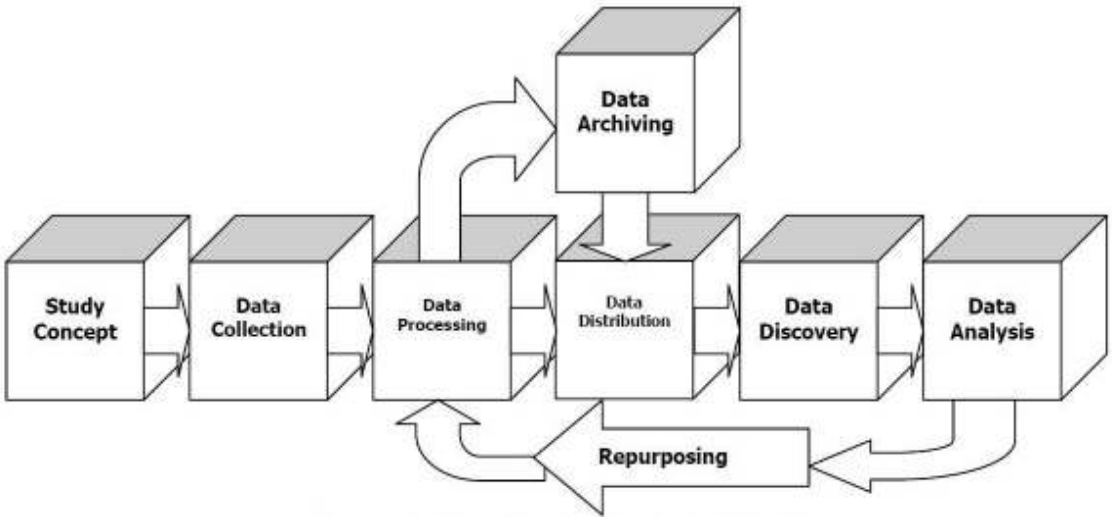
DDI - Codebook does not support the description of data coming from complex longitudinal or repeat-cross-sectional surveys, which have many successive waves. Neither is it capable of describing the surveys or other instruments used in data collection. For each DDI – Codebook document, there is only support for a single language.

Despite these limitations, many statistical agencies have found DDI to be a useful tool for documenting collected data. Mostly, these NSIs are in the developing world, where the "Metadata Management Toolkit" produced by the International Household Survey Network is in widespread use. This toolkit is a set of tools for publishing HTML documentation from the DDI XML, and uses the Nesstar Publisher technology. This version of DDI is also used by almost every European national data archive, being an agreed standard for CESSDA, the community of European data archives. It is also used widely in the US and Canada, and in Australia and New Zealand.

B. DDI - Lifecycle: The New DDI

The newer branch of the DDI family is focused not on after-the-fact documentation of data sets, but on describing metadata as they are created and used throughout the data production lifecycle. This branch of DDI started in version 3.0, and will be maintained in parallel with the DDI - Codebook branch. As might be expected, there is a strong similarity between them in those areas where they contain the same metadata, and mappings between the branches are provided.

The lifecycle model underlying the DDI – Lifecycle specification is shown below.



It is useful to think of DDI as supporting metadata-driven survey design. Over the life course of a survey that results in a data set – from initial conceptualization to data publication and beyond -- a huge amount of metadata is typically produced. These metadata can be recorded in DDI format and re-used as the data collection, processing, tabulation, and reporting/dissemination take place. The DDI metadata are both documentary and also “machine-actionable” – that is, they can be used to drive processes and to support additional steps in the life cycle. This is especially important for describing how microdata are cleaned, edited, tabulated and anonymized in the production of aggregates.

DDI – Lifecycle has capabilities which do not exist in DDI – Codebook. Among these are the abilities to describe survey instruments and other forms of non-survey data collection. Also, more than one data set can be described, so that for data collections which are conducted on a repeat basis the similarities and differences across waves can be clearly identified. There is a strong emphasis on identifying re-use of metadata throughout the lifecycle, and also across different data collections. DDI – Lifecycle is also multi-lingual: for each DDI document ("instance" in XML terminology), all human-readable text can be provided in alternate languages.

DDI – Lifecycle is the subject of much tools development: there are several libraries in Java for working with the standard, and an increasing number of open-source tools for editing, storing, and utilizing the DDI metadata. The main commercial package for working with DDI – Lifecycle metadata is Algenta Technologies' Colectica tool suite. Major open-source projects include the authoring and editing tools from the Danish Data Archive and the tools being developed by the Canadian Research Data Centre Network. Tabulation and visualization tools include those from Space-Time Research (which also supports SDMX). For an overview of tools, consult the DDI Alliance website. Also, every year at the IASSIST conference there is a set of presentations, available online, covering recent DDI – Lifecycle tools developments.

DDI – Lifecycle is a fairly new standard, and is being adopted mostly by those organizations which have more complex needs for metadata: research institutes conducting large-scale longitudinal or repeat cross-sectional surveys, research data centers, and similar organizations. However, there is starting to be more interest from other large data producers, among them such NSIs as the Australian Bureau of Statistics.

III. The DDI Alliance

DDI is developed, maintained, and governed by the DDI Alliance, an international membership-based organization currently housed at ICPSR, the major social science data archive in the U.S. There are approximately 30 members, coming from a variety of backgrounds: data producers and NSIs, data libraries housed in universities, research centers, and secure data facilities. The current list of members can be found on the DDI Alliance website.

Currently, there is an Expert Committee made up of representatives from the member organizations. This group, which itself has various working groups, makes substantive recommendations to improve the specifications and votes on changes to the standards. A Steering Committee, which is made up of an elected chair and vice-chair, and a set of representatives from some of the larger

member institutions, provides oversight. The DDI Director sits on the Steering Committee and is responsible for organizing the day-to-day operations of the organization.

The DDI Alliance publishes a quarterly newsletter, and has lots of information on its website regarding the standard and related events. There is an active user community in Europe, which holds an annual conference – the European DDI Users Group (EDDI). Also, the annual IASSIST conference features many presentations on DDI, and the annual face-to-face meetings of the DDI Expert Committee and Steering Committee are held in the margins of that conference.

IV. DDI, SDMX, and the GSBPM

This section describes how the DDI relates to other standards and reference models, with a strong focus on SDMX and the GSBPM, as those are most familiar to the NSI community.

It is important to note that DDI comes out of the domain of “Social, Behavioral, and Economic” (SBE) research, and thus has terminology which is sometimes unfamiliar (but often closely related to) the terminology used by those working in NSIs. A typical example is the use of the word “study” to refer to a data collection cycle, or survey.

A. The Standards and How They Relate

DDI applications focus on describing microdata and the processing performed on the data as they are integrated, tabulated, etc. Thus, DDI is not generally duplicative of SDMX, even though they have many common features. DDI supports dissemination of both aggregates and microdata, but is most often used in the dissemination of microdata for secondary use by researchers. It is also a standard for use during the data production lifecycle. DDI can support the exchange of metadata and data between organizations, but this is not its primary purpose.

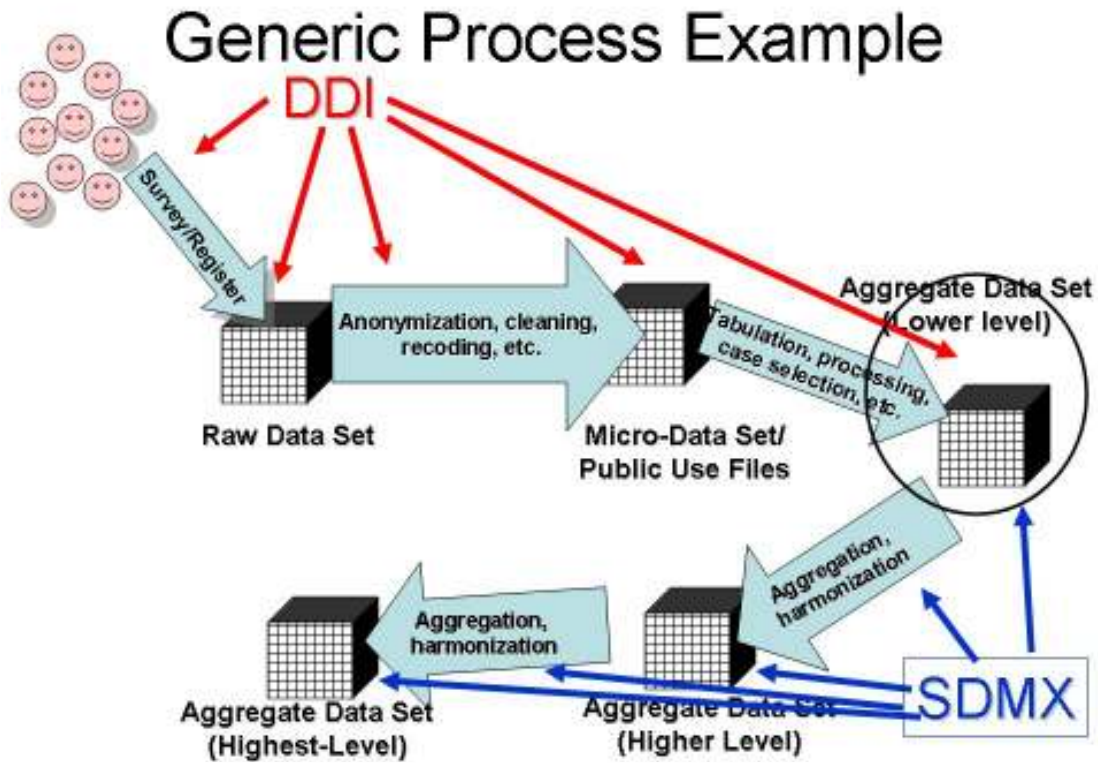
This places it in contrast with SDMX, which has a strong focus on reporting and collection, as well as on dissemination. The use of SDMX in internal production systems – while valuable – is a secondary application of the standard from the perspective of its designers.

Further, DDI is a technical standard, while SDMX has a technical component but also a “content-oriented” component, actively working to promote harmonization of terminology, concepts, and domain structures for aggregate data and related metadata.

Because both standards describe data sets and their structures, there are many common metadata components: both describe concepts, codelists, dimensions and attributes, measures, and the structure of aggregate data cubes. These are

fairly well-aligned between the two standards. Further, DDI and SDMX use similar schemes for maintenance and identification. These alignments at the technical level are intentional.

The focus and purpose of the standards are very different, however. If we look at the generic process of data production as it occurs within an NSI and the organizations to which an NSI reports, we can illustrate the areas where the standards are most appropriate:



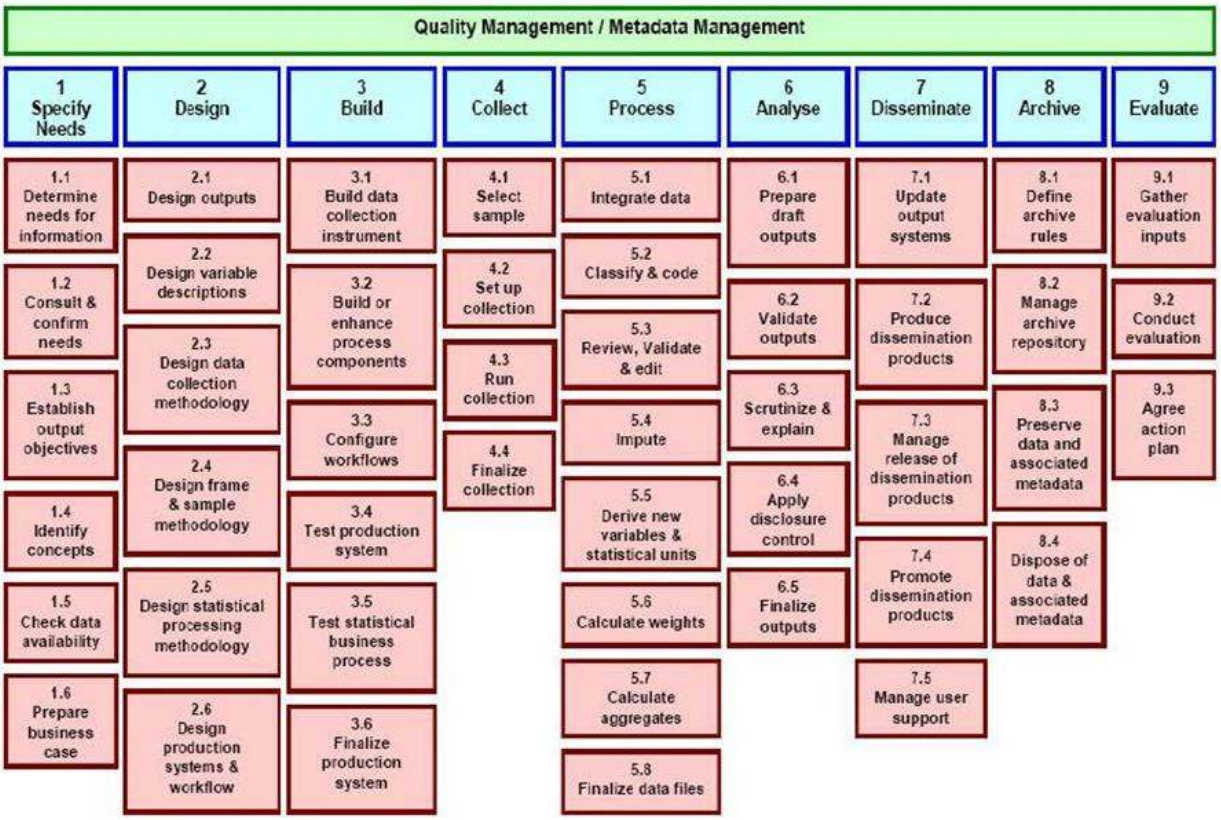
There is an overlap at the point of tabulation, allowing for the possibility of transforming data described with DDI metadata into an SDMX dataset, with related structural metadata.

While DDI – Lifecycle has the capability to represent data in a standard XML format, this is not a feature of the standard which is often used. DDI is primarily a standard for describing metadata, with the data remaining in a format such as CSV, SAS, SPSS, etc. The relevant metadata is attached to specific values in the native data formats using external references such as column-row coordinates.

It is important to understand that DDI has a large number of metadata fields which may or may not be used by any given application. To address the need for

customization of DDI for specific uses and by specific organizations, there is a "DDI Profile" mechanism that enables one to identify which metadata elements are used by a given community or organization, so that a smaller agreed set of information can be identified for use. In initial explorations of ways that DDI and SDMX might be used together (see below), it is assumed that a subset of the larger DDI specification could be identified for use by NSIs, as they may have different requirements from other communities which use the standard.

It is interesting to note that one of the inputs to the Generic Statistical Business Process Model (GSBPM) was the DDI Lifecycle model shown earlier. There are some differences between these models, but they are substantially similar at the top level:



The major differences result from the fact that in social science research, data collection is often grant-funded, and occurs at a single point in time to support some specific research, whereas data collection in NSIs is most often conducted on a repeat basis, every quarter, month, year, etc. Thus, the GSBPM has a much greater focus on evaluation for the next cycle. In both cases, however, there is a great deal of reuse of metadata, which is well-supported by DDI – Lifecycle.

While the GSBPM could describe the collection, processing, and further aggregation of aggregate data, this process is most typical in international statistical organizations. For NSIs, the more usual inputs are not aggregate, but are instead microdata. Thus, NSIs are one type of organization which could benefit from the use of both standards, handing off from DDI to SDMX at the point of tabulation. Tools such as those from Space-Time Research illustrate this clearly: the tabulation engine runs off a database imported as DDI metadata plus ASCII data, but the tabulations are exposed in SDMX formats, as data sets and data structure definitions.

B. Possible Uses of the Standards

There is currently an exploration of the combined use of the two standards to support the GSBPM, as described by a presentation by Steven Vale at the 2010 European DDI User's Group conference and a paper available at http://www.ddialliance.org/resources/publications/working/othertopics/ExploringRelationshipBetweenDDI-SDMX-GSBPM.pdf. Several different scenarios are being considered.

One is general support for the lifecycle, where microdata are described using DDI – Lifecycle, and, once tabulated, the aggregate products are described, reported, and disseminated as SDMX. This model is of obvious applicability for NSIs, because of the nature of the tools in each area: SDMX tools are well suited to reporting and dissemination, while DDI tools tend to focus on data collection with tools such as Blaise, and data processing in SAS, SPSS, Stata, and other statistical packages.

A typical scenario here would employ a DDI tool such as Colectica to describe a survey, which could then automatically be exported as Blaise code to support a CAI data collection. Once collected, the DDI metadata could be used to generate SAS set-up files for processing. A tool such as Space-Time Research's SuperCross could then be used to tabulate the microdata, and render it into SDMX. Once in SDMX format, it could be directly disseminated using a tool such as OECD.stat, or further manipulated in a data warehouse to meet the required data structures for reporting in SDMX format.

Another case of interest in the use of the two standards is more focused on enriching data dissemination. Because DDI metadata are very rich, and describe the process of collection and tabulation, they could potentially be linked to disseminated aggregates, being exposed alongside the aggregate data products on a website, as embedded metadata, or actually presented in native DDI XML format, or mapped into an SDMX metadata report. A related case "mines" DDI metadata for the automatic population of SDMX-based quality reporting (this is a case being implemented by INEGI in Mexico).

One final scenario proposes the use of DDI and SDMX as the publication formats of standard classifications and concepts. Today, such things are typically

exposed in PDF form, or in Excel spreadsheets. Having standard classifications (such as those often produced by NSIs) exposed in parallel SDMX and DDI formats would be of value to other agencies and researchers who use them in their own data collection and processing activities.

C. Related Developments: GSIM, ESSnet CORE, and Others

Two other recent developments in this area should also be discussed, as they relate to SDMX and DDI. The first of these is the Generic Statistical Information Model, or GSIM, which is a project currently being conducted by the "Statistical Network," an informal group of NSIs working together to produce common models, approaches, and technology tools. The idea behind GSIM is to create a companion-piece to the GSBPM, but one which describes statistical data and metadata throughout the process, rather than the process itself. The reference model would presumably be finalized and published by METIS, like the GSBPM itself. Such a model could be the basis of an agreed mapping between DDI and SDMX for use in NSIs, and this approach is being discussed within that community and elsewhere. It is notable that the leader of the GSIM work – the Australian Bureau of Statistics – is also very interested in implementing SDMX and DDI together to support the GSBPM process model.

Eurostat has organized the ESSnet CORE project, which is working to define a common statistical architecture across Europe. This work builds on the earlier CORA project within ESSnet, which was based on the GSBPM. CORE has a focus on describing a technical architecture for services descriptions and workflows, and on idea is that the data model – that is, the inputs and outputs of the services – might be described by the GSIM models, and implemented as DDI and SDMX formats.

It is significant that in a recent document produced by the HLG – BAS committee (a high-level coordination committee for business architecture and strategy formed by the Conference of European Statisticians), there is mention of the use of GSIM and CORE working together in support of the GSBPM. It is easy to see how SDMX and DDI could become the implementation formats to support such a combination.

Additionally, informal discussions have begun between the DDI Alliance and the SDMX Sponsors, to explore what the standards organizations might do to support the use of these standards in tandem.

This is ongoing work, and still very much in the exploratory stages, but the possibility would seem to be a promising one.

Another tangentially related project within Europe is the "Data without Boundaries" (DwB) collaboration, between the European NSIs and the national data archives. This project would allow qualified researchers in Europe to discover and access microdata held in any European archive or NSI through a

single portal. DDI is the metadata standard used by the archives – if it becomes popular in support of the GSBPM, then the same metadata would be available for use within the European research infrastructure being created by the DwB project. It is still too early to know what will come out of this project, but one of the work packages has the goal of looking at metadata standards, particularly DDI and SDMX.

V. Summary

SDMX and DDI could potentially be used together by NSIs, and this possibility is being actively explored today. That these standards might be related to a more general statistical framework is a strong possibility, with the GSBPM being the current centre-piece of such a framework.

Regardless of the results of such exploration today, DDI does offer a set of useful tools for NSIs who are looking for a standard metadata model for microdata, and its processing and tabulation.