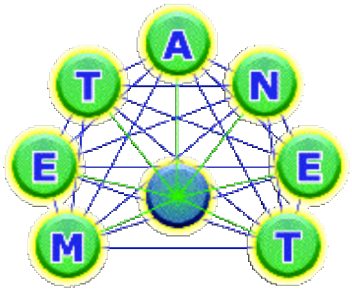

The Data Documentation Initiative (DDI) metadata specification

Jostein Ryssevik,
Nesstar Ltd

Introduction

The Data Documentation Initiative (DDI) is an international program to produce a metadata specification for the description of social science data resources. The program was initiated in 1994 by the Inter-University Consortium for Political and Social Research (ICPSR). Contributors to the efforts of the DDI come from social science data archives and libraries in USA, Canada and Europe and from major producers of statistical data (like the US Bureau of the Census, the US Bureau of Labour statistics, Statistics Canada and Health Canada). Until recently the work has been carried out by a committee based on voluntary participation. A reorganization to a more strongly knit consortium model will be put in place by the end of 2002.

The original aim of the DDI committee was to replace the existing and widely used OSIRIS Codebook/data dictionary standard with a more modern and Web-aware specification that could be used to structure the description of the content of social science data archives. The first preliminary version of the DDI specification came in the form of an SGML DTD, which in 1997 it was converted into an XML DTD. [i] The migration to XML happened just a few months after the W3C released the very first working draft of the XML specification. The DDI was consequently one of the very first major metadata initiatives using the new framework.

The first official version of the DDI specification (version 1.0) was published in March 2000. Publication followed extensive beta-testing in a variety of environments as well as software implementation by major projects like Nesstar.[ii] Since the first public release,

several extensions have been made to the specification, most noteworthy the inclusion of a sub-model for description of aggregated data structured as multidimensional tables (cubes). Several other extensions, like more precise descriptions of data coming out of CATI/CAPI-like environments are on the drawing table. Plans do also exists for a DDI II process that will revisit the entire DDI-specification from a modeling perspective and possibly produce alternative syntactical representations in frameworks like RDF Scheme and/or XML-Schema.

The uptake of the specification in the data archive community has been quite remarkable. Several major European data archives have already migrated their entire holdings to the DDI and an effort to create a harmonized and integrated interface to all the national social science data archives in Europe based on the new standard has been initiated. In North America, the worlds biggest data archive, ICPSR, as well as several minor data centers and libraries, are in the process of moving to the DDI. The acceptance of the DDI is also growing outside the data archive community. Health Canada is, as an example, making their new data dissemination tool webDAIS compatible with the DDI specification.

The environment of the DDI

To understand the DDI, it is important to understand the environment from which it has grown. The social science data archives are rarely engaged in the collection of primary data, but serve as brokers between various data providers and the academic community. Their holdings contain data from the public sector (statistical agencies, central government etc), the commercial sector (opinion and market research companies) and academic research. The archives do not only preserve data for future use but also add their own value to the collections:

- Data received by the archives goes through a variety of checks and cleaning procedures to ensure their integrity.
- Any system or software dependencies are stripped away to make sure that data can be read at any time in the future.
- Comprehensive computer-readable metadata are developed.
- Data from various sources are often integrated and harmonised in order to produce easy-to-use information products (on-line databases, CD-roms etc.).
- Data are catalogued and made accessible through electronic search and retrieval systems.
- In order to encourage the use of statistical data among students, teaching packages and interactive statistical laboratories, are developed.

Due to the extensive refinements of the data sources, data and related services from the archives are frequently requested by non-academic users. This includes users from the public sector, as well as from the mass media and private companies. To the extent that services to non-academic users do not run counter to the agreements with the data depositors, access is usually granted.

The characteristics of the user communities go a long way to explain the high priority that the archives have given to the development of metadata:

- Users of archived data have rarely been engaged in the creation of a dataset.
- Archived data will frequently be used for other research purposes than intended by the creators (secondary analysis).
- Archived data will frequently be used many years after they were created.
- Academic users are often comparing and combining data from a broad range of sources (across time and space).

The common denominator of the four characteristics is an emphasis on the relative distance between the end users of a statistical material and the production process. Whereas the creators and primary users of statistics might possess "undocumented" and informal knowledge, which will guide them in the analysis process, secondary users must rely on the amount of formal meta-data that travels along with the data in order to exploit their full potential. For this reason it might be said that social science data are only made accessible through their metadata. The metadata provides the bridges between the producers of data and their users and convey information that is essential for secondary analysts.

The products of the DDI

So far, the DDI committee has delivered two products: the **DDI 1.0 specification** in the form of an XML DTD and an extensive **Tag Library** describing the role and use of the various elements in the specification. Both products can be found at the DDI website along with further information about the work of the committee.[iii] Several draft versions including resent changes and additions can also be found at this site.

The structure of DDI 1.0 specification

An XML DTD (Document Type Definition) provides the rules for applying XML to a document of a specific type. The DTD defines the elements that the document is composed of, the attributes of these elements, and their logical relationships to other elements. The elements will usually be arranged in a hierarchical or tree-like structure.

The DDI-tree contains five main branches, or sections:

1. **The Document Description,** which consists of bibliographic information describing the metadata document and the sources that have been used to create it (this section can thus be looked upon as metadata for the metadata, or mete-metadata if you like).

   Description of this section in the TagLibrary: The Document Description consists of bibliographic information describing the DDI-compliant document itself as a whole. This Document Description can be considered the wrapper or header whose elements uniquely describe the full contents of the compliant DDI file. Since the Document Description section is used to identify the DDI-compliant file within an electronic resource discovery environment, this section should be as complete as possible. The author in the Document Description should be the individual(s) or organization(s) directly responsible for the intellectual content of the DDI version, as

distinct from the person(s) or organization(s) responsible for the intellectual content of the earlier paper or electronic edition from which the DDI edition may have been derived. The producer in the Document Description should be the agency or person that prepared the marked-up document. Note that the Document Description section contains a Documentation Source subsection (1.4) consisting of information about the source of the DDI-compliant file-- that is, the hardcopy or electronic codebook that served as the source for the marked-up codebook. These sections allow the creator of the DDI file to produce version, responsibility, and other descriptions relating to *both* the creation of that DDI file as a separate and reformatted version of source materials (either print or electronic) *and* the original source materials themselves.

The Document Description has the following sub-sections:

- **Citation:** The bibliographic information describing the marked-up codebook
- **Guide to the documentation**: List of terms and definitions used in the document.
- **Documentation status**: Used to indicate at what stage the documentation is in the production process.
- **Documentation source:** Citation for the source document. This element encodes the bibliographic information describing the source codebook

2. **The study description,** which contains information about the data collection.

Description of this section in the Tag Library: The Study Description consists of information about the data collection, study, or compilation that the DDI-compliant documentation file describes. This section includes information about how the study should be cited, who collected or compiled the data, who distributes the data, keywords about the content of the data, summary (abstract) of the content of the data, data collection methods and processing, etc. Note that some content of the Study Description's Citation -- e.g., Responsibility Statement -- may be identical to that of the Documentation Citation. This is usually the case when the producer of a data collection also produced the print or electronic codebook for that data collection.

The Study Description has the following sub-sections:

- **Citation:** The bibliographic information describing the data collection
- **Study scope**: This section contains information about the data collection's scope across several dimensions, including substantive content, geography, and time. The section is also containing a placeholder for an abstract as well as keywords.
- **Methodology and processing**: This section describes the methodology and processing involved in a data collection (including data collection methodology, sampling procedures, deviations from the sampling plan, data collection modes and instruments, control operations, data cleaning, weighting, response rates etc.)
- **Data access**: This section describes access conditions and terms of use for the data collection.
- **Other study description materials**: This section describes other materials that are related to the study description. This is primarily descriptions of the content

and use of the study, such as appendices, sampling information, weighting details, methodological and technical details, publications based upon the study content, related studies or collections of studies, etc.

3. **The Data Files Description,** which describes each single file of a data collection (formats, dimensions, processing information, missing data information etc.)

   Description of this section in the Tag Library: The File Description consists of information about the particular data file(s) containing numeric and/or numeric + textual information that the DDI-compliant file describes. This section consists of items describing the characteristics and contents of file(s) that comprise the study as described in the Study Description. There may be multiple file descriptions if there are multiple files in the collection.

   The File Description has the following sub-sections:

   - **File description**: This section contains elements to describe the physical format and layout/structure of each single datafile.
   - **Notes**: Additional textual information about each file.

4. **The variable description,** which describe each single variable in a datafile (format, variable and value labels, definitions, question texts, imputations etc.). While section 3 provides the physical description of a dataset, section 4 provides the logical.

   Description of this section in the Tag Library: (missing for this section)

   The Variable Description has the following sub-sections:

   - **Variable group**: This set of elements provides a mechanism to organise variables into groups and sub-groups. The mechanism can be used to make a dataset more easy to navigate (by providing a hierarchical Table of contents) or to group variables that are to be treated in a specific way (like a multiple response group in survey data).
   - **Variable**: The variable sub-section contains an extensive set of elements to describe the characteristics of a single variable. This includes names and labels, question texts and other relevant information from the data capturing process, definitions, information about imputation, security and access conditions, weighting and missing data information, summary statistics, description of the value domain etc. Note that for enumerated value domains (discrete variables), this set of elements can also be used to describe each single permissible value, including their frequencies.
   - **NCube (added in version 1.02)**: This section is used to describe data cubes or multidimensional tables. Note that the Variable element above is used to describe the variables that constitute the dimensions of a cube.
   - **Notes**: Additional textual information about the variables.

5. **Other Study-Related Materials,** which can include references to reports and publications, other machine-readable documentation that are relevant to the users of the study (referenced by URI's) etc.

   Description of this section in the Tag Library: This section allows for the inclusion of other materials that are related to the study as identified and labeled by the DTD users (encoders). The materials may be entered as PCDATA (ASCII text) directly into the document (through use of the "txt" element). This section may also serve as a "container" for other machine-readable materials such as data definition statements by providing a brief description of the study-related materials accompanied by the attributes "type" and "level" defining the material further. The "URI" attribute may be used to indicate the location of the other study-related materials. Other Study-Related Materials may include: questionnaires, coding notes, SPSS/SAS/STATA setups (and others), user manuals, continuity guides, sample computer software programs, glossaries of terms, interviewer/project instructions, maps, database schema, data dictionaries, show cards, coding information, interview schedules, missing values information, frequency files, variable maps, etc.

   Other Study-Related materials have no sub-sections.

Elements and attributes

In XML content can either be entered as an **element text** (shorter or longer blocks of text appearing between a start-tag and an end-tag), or as **attribute values** (a name or term defining the state or value of an attribute). Attributes can either be **enumerated** (requiring that values must be taken from a list provided in the DTD - a controlled vocabulary) or **non-enumerated** (where the values are chosen more freely).

Given the ability to control the content of attributes it is a good design principle to use attribute values to drive software processes and element text to deliver messages to human readers. The DDI is, to a large extent, following this principle.

The timeMeth element, which is supposed to contain a description of the time method or time dimension of the data collection, might serve as an example. The element text can be used to give a human language description of the method, whereas the method-attribute can include a controlled vocabulary, which more easily can be understood by a software system. An example of how this structure can be used in concrete mark-up is shown below:

```
<method><dataColl><timeMeth method='panel'>The study is a panel
survey where 50% of the sample are replaced at each
subsequent.....etc. </timeMeth></dataColl></method>
```

Controlled vocabularies

The use of controlled vocabularies has been heavily discussed within the DDI committee. From a general point of view controlled vocabularies add structure and predictability to the specification – thereby making the metadata instances more easily understood by a software process. The downside is, of course, related to the difficulties of defining a complete set of terms that might fit all purposes and users. Controlled vocabularies might easily become too restrictive, thereby providing reasons to break the standard.

The DDI provides controlled vocabularies for a number of attributes. Examples are:

> …the type-attribute of the "file structure" element which might take the values : (rectangular|hierarchical|relational)
>
> …the type-attribute of the "summary statistics" element which might take the values: (mean,median|mode|vald|invd|min|max|stdev)

A more extensible approach to controlled vocabularies is in use for a more limited set of attributes. In this approach there is no list of values declared in the DTD, but a mechanism is provided whereby metadata authors can include a reference (name and URI) to an externally defined vocabulary (a simple list of terms, a more extensive thesaurus or an ontology). The specification of this mechanism for the keyword-element is shown below.

```
<!ELEMENT keyword     (#PCDATA | %a.phrase;)*                    >
<!ATTLIST keyword     %a.global;
                      vocab   CDATA   #IMPLIED
                      vocabURI   CDATA   #IMPLIED                >
```

Trees and relations

The structure of XML is fundamentally tree oriented and one of the important functions of a DTD is to declare how the various elements are to be nested. The tree structure provides a simple data model that can be used to represent the relationships between the described objects. The DDI, to a large extent, builds on this tree structure. Category values are, as an example, nested below categories, which again are nested below variablesetc.

However, the DDI also uses the internal referencing mechanism of XML (IDREF attributes pointing from one element to another) to break out of this tree structure. An example of this is the relationship between variables and variables groups. To allow a variable to appear in more than one group, variables are not nested below the groups but are instead referenced from the group elements by means of IDREFs. An example of how

this might be used is shown below (where the group labelled “Introduction” includes several variables, amongst them the variable “Population density” with the ID V5):

```
<varGrp ID="G1" type="subject" var="V1 V2 V5 V6 V7 V8 V14
    V16 V17 V18 V19">
    <labl>Introduction</labl>
</varGrp>

<var ID="V5" name="POPDEN" files="F1" dcml="2"
    intrvl="contin">
    <labl level="variable">Population density Q10</labl>
</var>
```

Another example is references from datafiles, variable groups and variables to identified parts of the methodology and processing section of the study description (using a reserved attribute called methrefs).

Multilingualism

To facilitate the production of multilingual metadata instances, every DDI-element contains an xml-lang attribute. The xml:lang attribute might as an example be used to include abstracts in more than one language or to repeat question texts in all the source languages of a cross-national survey.

The problem with the solution is that it requires every element of the DDI to be repeatable. This corrupts the cardinality of the metadata structure creating severe difficulties for any processing software. Consequently a refinement of this solution can be expected.

Obligatory and recommended elements

A stated goal of the DDI programme has been to define a superset of all possible elements and attributes used to describe social science data resources. The result is consequently a very rich specification with defined placeholders for almost any piece of information that a data producer or distributor might find appropriate to associate with a dataset.

The committee has on the other hand had more difficulties coming to an agreement as to which elements should be defined as obligatory. The argument against obligatory elements has been to keep the entry-level for using the DDI as low as possible to encourage widespread acceptance in various user communities. In version 1.0 this argument has been taken to it’s extreme in that only one element – the title – is declared as strictly obligatory. This is, of course, not a very satisfactory solution from an interoperability point of view. It is also creating problems for application providers that

need more predictability as to the type of information they can expect to find in a DDI instance.

An effect of this unwillingness to enforce the use of a broader set of metadata elements is that application providers as well as communities have started the process of defining their own list of obligatory DDI elements. An example of the former is Nesstar that is in the process of developing an "application profile"[iv] that formally defines a set of additional requirements to the DDI specification that needs to be fulfilled in order to deliver DDI-data through the system. An example of the latter is the Cessda DDI-Group that is defining rules (obligatory elements, further controlled vocabularies etc.) for using the DDI among the European social science data archives.[v] None of these initiatives are breaking or extending the standard, only refining its use in a particular contexts or environments.

As a compromise between making elements obligatory and allowing metadata authors to pick and choose from the rich inventory of available DDI elements, the committee has designated a list of elements as "strongly recommended". The list includes 14 elements with a direct mapping to Dublin Core as well as 30 additional elements.[vi]

A discussion of limitations and challenges

An external evaluation of the DDI program was conducted in February 2001. The evaluators was generally very positive to the work and products of the committee and was particularly pleased with its rapid adoption and uptake by the target community – the social science data archives and libraries on both sides of the Atlantic. The DDI specification was seen as "a strategic component of the infrastructure necessary to support the exchange of structured social research survey data". What the evaluators did not mention was the DDIs ability to move beyond its original domain and to bridge the gap between the data producing and data archiving communities.

Regarding the technical quality and direction of the current DDI specification, two issues were emphasised: One was the survey-data bias of the DDI and the need to extend the specification to support more complex data types – most notably time-series and aggregated data. The other was the lack of modularity and extensibility that to a large extent must be seen as a consequence of the inherent limitations of the XML DTD framework. [vii]

In the following section these and others limitations of the DDI approach will be discussed briefly.

The survey-data bias

The DDI was originally developed to describe the most typical information object of the social science data archive, the independent survey data file. Although references are

made to other types of data, all important concepts and most of the logic are derived from this starting point.

The integration of elements to describe cubes or multidimensional tables represents the first major step beyond survey data. Work is also in progress to create a refined specification for data of a hierarchical and relational nature. Both of these exercises have however demonstrated that proper support for more complex data structures might require substantial changes to the original DDI architecture – amongst other a more consequent separation of the description of logic from the description of physical storage. The move from the relatively standardised and predictable world of the rectangular micro-level data matrix cannot easily be achieved by "conceptual stretching".

A pure "bottom-up" approach

The DDI specification has been developed to describe concrete files or products coming out of the statistical production process. Given it roots in social science data archiving this is quite natural. The information objects of the data archives are finalised products that have cut the lifeline to the various production processes and put into the hands of the users. It might even be said that the main purpose of the DDI is to create a structure that can carry as much information as possible from the production stage to the users.

A consequence of this approach is, however, that the DDI do not have a level of abstraction above a concrete dataset or statistical product. There is a one-to one relationship between a DDI instance and the physical data it is meant to describe. The DDI is tied to the dataset, or put differently; the DDI abstraction ladder stops with the dataset (it is "metadata after data" or pure "bottom up"). As a consequence there is no methods in the DDI to describe abstract statistical concepts that might be represented in more than one concrete study.

This is totally different from a standard like ISO/ICE 11179 that draws a distinction between abstract concepts and conceptual domains on the one hand, and variables and value representations on the other. Such a distinction is important because it allows an application to identify variables that are measuring the same concept, and consequently to identify comparable variables across datasets. In the DDI there is no way of referencing identical variables represented in more than one study, and even series of survey instances where the majority of variables are identical from wave to wave has to described instance by instance.

Modularity

The DDI specification has its roots in a "book" metaphor. It was originally seen as a digital equivalent of a paper document – the well established codebook or data dictionary. Even though the DDI codebooks are divided into 5 chapters or sections (see above), they are not built according to a modular architecture that allow information and application providers to select bits an pieces and "snap" them together on a more freely basis.

The lack of modularity is closely related to the chosen framework – the XML DTD. Compared to XML Schema that allows a specification to be broken into a set of reusable components an XML DTD will always force you into a more monolithic approach. The DTDs also lack support for the most fundamental prerequisite for modular metadata – XML namespaces.

Extensibility

Another consequence of the DTD framework is the lack of a proper extensibility mechanism. Within the confines of a DTD there are no ways to add local extensions without compromising the interoperability of the core specification. You either accept the specification as it is without any additions or you break it.

For a big and complex specification like the DDI this is a major problem that can easily hurt the adoption process. There will always be situation where the specific needs of a given application or resource type will call for local extensions. Without a mechanism that allows extensions to be made without breaking the standard the chances are high that application providers might sacrifice interoperability for local efficiency and relevance.

Metadata modelling

All the limitations discussed above are well known to the DDI committee and high on the agenda for the DDI 2 process. In order to solve the problems it has been discussed to make a clearer distinction between syntax and semantics – that is to separate the semantic model underlying the DDI specification from its current syntactical representation (the XML DTD). Although the DDI obviously is based upon a model of social science data resources, this model has never been clearly formulated or made explicit in a general modelling language like UML.

It is expected that an explicit formulation of the "DDI model" will reveal the weak points and prepare the ground for a more flexible, modular and extensible specification. It will also provide a more solid starting point for a migration to alternative syntactical representations like XML Schema or RDF Schema. The explicit formulation of the "DDI model" will, in fact, make the choice of framework for syntactical representation less salient. The DDI might become a multi-part standard with a semantic model in the core and "bindings" to one or more syntactical representations. This is an approach taken by several metadata standardisation activities, amongst other the IEEE Learning Object Metadata standard[viii] and the new version of ISO/IEC 11179 Metadata registries.

[i] The translation from SGML to XML was done by Jan Nielsen from the Danish Data Archive as part of his disertation: From OSIRIS to XML. Markup and Internet Presentation of Structured Data Documentation. Odense, Denmark, 1977.

[ii] A description of the relationship between Nesstar and the DDI can be found in: Ryssevik, Jostein: The parallell stories of NESSTAR and the DDI. Paper given at the UN/ECE Work Session on Statistical Metadata, Geneva, Switzerland, 22-24 September 1999.

[iii] The DDI web-site can be found at: http://www.icpsr.umich.edu/DDI/

[iv] For a discussion of the term application profiles, see Heery, Rachel and Manjula Patel, Application Profiles: Mixing and Matching Metadata Schemas, *Ariadne,* Issue 25 (September 2000) <http://www.ariadne.ac.uk/issue25/app-profiles/intro.html>

[v] Information about the work of the Cessda DDI-Group can be found at: http://www.sidos.ch/CDG/

[vi] The list can found at: http://www.icpsr.umich.edu/DDI/CODEBOOK/recommended.html

[vii] The complete evaluation report can be found at: http://www.icpsr.umich.edu/DDI/PAPERS/evalsummary.pdf

[viii] Fore more information, see: http://ltsc.ieee.org/wg12/