AN ARCHIVE'S PERSPECTIVE ON DDI 3

By Mari Kleemola with Michelle Edwards, Sanda Ionescu, Uwe Jensen, Olof Olsson, Ørnulf Risnes, and Wendy Thomas

10/01/2010

DDI Working Paper Series – Document 7

This paper is part of a series that focuses on DDI usage and how the metadata specification should be applied in a variety of settings by a variety of organizations and individuals. Support for this working paper series was provided by the authors' home institutions; by GESIS - Leibniz Institute for the Social Sciences; by Schloss Dagstuhl - Leibniz Center for Informatics; and by the DDI Alliance.

An Archive's Perspective on DDI 3

BY MARI KLEEMOLA, MICHELLE EDWARDS, SANDA IONESCU, UWE JENSEN, OLOF OLSSON, ØRNULF RISNES, AND WENDY THOMAS

ABSTRACT

Ongoing data series tend to be complex in terms of management, processing, and analysis. A team in Europe made an exploration and evaluation of DDI 3 to determine its suitability for marking up and managing cross-national comparative surveys like the International Social Survey Programme data. Their findings may help others interested in organizing survey series using the DDI's Group and Resource Package mechanisms.

This paper describes the DDI 3 evaluation process at the Finnish Social Science Data Archive (FSD), and the issues that came up during this work. Several reasons led FSD to investigate moving from DDI 2.1 towards DDI 3: for example, multilinguality issues and the need to avoid repeating information when documenting series of studies, as well as the need to better manage the growing amount of metadata. The ISSP 2006 Finnish data was chosen as the use case. A subset of these data was documented in DDI 3.1.

This project is also a component of the CESSDA PPP work. CESSDA PPP stands for the Preparatory Phase Project for a Major Upgrade of the Council of European Social Science Data Archives (CESSDA) Research Infrastructure. The aim of the project is to plan the future development of the CESSDA RI and focus on tackling and resolving a number of strategic, financial, and legal issues in order to ensure that European social science and humanities researchers have access to, and gain support for, the data resources they require to conduct research of the highest quality, irrespective of the location of either researcher or data within the European Research Area. For further information please see http://www.cessda.org/project/

BACKGROUND (ORGANIZATION and CONTEXT)

The Finnish Social Science Data Archive (FSD) is a national resource center for social science research and teaching. FSD archives, promotes, and disseminates digital data for research, teaching, and learning purposes. The Archive is funded by the Finnish Ministry of Education and is a separate unit of the University of Tampere.

Currently FSD's holdings include over 700 Finnish survey datasets. All studies are documented using DDI 2.1 at the study and variable level. Documentation is available in both Finnish and English. Many of the surveys have been conducted in Finnish and Swedish (the two official languages in Finland), adding another layer to FSD's documentation.

Although DDI 2.1 has served the FSD very well in the past, changes to the DDI standard that deal specifically with multilinguality and which help reduce the duplication of metadata entry have prompted the FSD to look more closely at moving towards DDI 3. This investigation was also prompted by the increasing number of studies and metadata that FSD holds, challenging its current metadata system. In addition, plans to move away from DDI 2.1 XML files towards a database system have prompted the consideration of moving towards DDI 3.1. It is also anticipated that in the future several different types of documentation will be

produced from a centralized FSD system, for example FSD's www catalogue(s), metadata for preservation purposes, entries for the CESSDA catalogue, Question Banks/Indices, and records for the Finnish National Digital Library. The structure of DDI 3 encouraging the reuse of existing metadata resources makes it an attractive choice. Furthermore, FSD participates in the CESSDA PPP workpackage 8, where one of the tasks was to evaluate DDI 3.

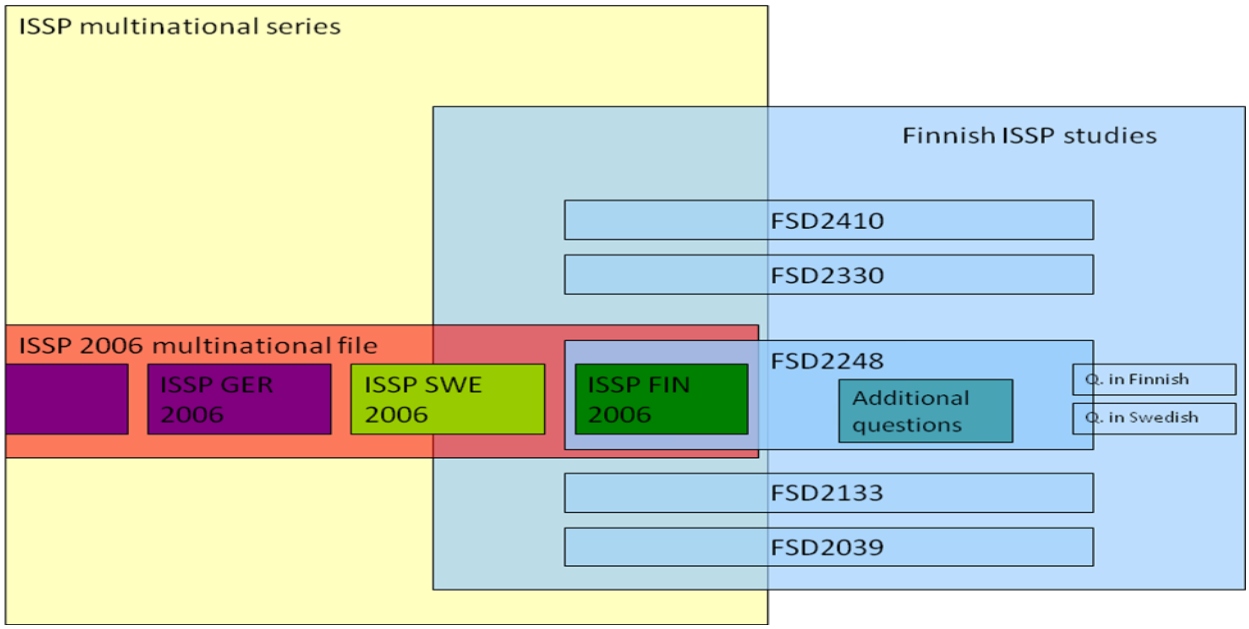
Enhancing the experience of FSD data users is another important goal in looking at DDI 3. To provide researchers with a better service, FSD would like to incorporate and include more metadata than is currently available in DDI 2.1 and its current operational database. FSD is also interested in including information required to build a question bank, locate comparable variables, and allow geographic interfaces. Maintaining documentation in different languages also needs a more sophisticated solution than what is currently provided by DDI 2.1.

To determine if and how DDI 3 would satisfy these needs, a subset of variables from the International Social Survey Programme (ISSP) 2006 Finnish data was chosen as a use case. The ISSP is a continuing annual program of cross-national collaboration on surveys covering topics that are important for social science research. See www.issp.org for further information.

In collaboration with Statistics Finland and the Department of Social Research at the University of Tampere, FSD is responsible for collecting the Finnish ISSP data. FSD processes the data and sends it to GESIS for merging into the multinational file. ISSP 2006 Finnish data is a single study that belongs to several groups/series, thus making it an ideal use case. The main questionnaire is in English and is translated into both Finnish and Swedish. By creating DDI 3 documentation, FSD may be able to collaborate with its ISSP partners in the sharing and reuse of metadata.

USE CASE / REQUIREMENTS

Context of the Use Case



The above diagram showcases how the FSD_ISSP survey fits into the ISSP multinational series. The potential for reuse of metadata and the interaction between the different languages are apparent in the diagram.

Studies that are designed to be compared, like longitudinal studies or cross-national collaborations and multi-national studies, may be documented in DDI 3 using the grouping structure and taking advantage of the inheritance feature. Another option to document similarities between variables or other items is to use the comparison module. The comparison module is most useful when studies were not designed to be comparable (see "Using DDI 3 for Comparison", at http://www.ddialliance.org/sites/default/files/UsingDDI 3ForComparison_0.pdf).

There are several ways to group studies, depending on the point of view. Each individual Finnish ISSP study comprises a core module (green in the picture) and additional questions (dark blue). The theme of the core module changes every year, and themes re-occur. Thus it would be possible to group by theme. On the other hand, each Finnish ISSP core will be merged with other countries' data to form the multi-national file. Thus the grouping could be made by wave/data collection year.

In this use case, from FSD's perspective, most of the common information was at the study level, so it made sense to group by country, creating a group called "Finnish ISSP studies".

For the purposes of this Use Case, a subset of eighteen variables were chosen from the Finnish ISSP file. These include:

Archive and ID variables - Technical variables

- FSD specific

Substantial variables - types

- Simple question: Variable V4 : Q1: Obey laws without exception
- Question with items (small battery):
  - V5 Q2a: Public protest meetings
  - V6 Q2b: Protest demonstrations
  - V7 Q2c: National anti-government strike

ISSP demographics and interview char. - types

- SEX: R: Sex

Country specific Questions / Variables

- Ctry specific: degree
- Country specific education: Finland

In addition the following (for FSD's test purposes):

- (K2) Year of birth (integer) - the ISSP variable AGE is calculated from this variable
- (K20) WRKSUP : R: Supervises others at work (a filter variable)
- (K21) Is R responsible for… (a battery with three subitems and a different universe due to the filter K20, not in the basic ISSP 2006 questionnaire)
- (K22) WRKTYPE : R: Workg f priv., pub sector, selfempl.

The variables were selected to represent different types of questions in the study instrument and different kinds of variables (coded, open-ended, "batteries" with subquestions). The instrument was both in Finnish and Swedish, and the SPSS data file in Finnish. At a later stage the documentation will be translated into English to merge with the international file, but that part of the data lifecycle is beyond the scope of this Use Case.

The final markup was created using Oxygen XML Editor V. 10.0, with guidance provided by the DDI 3 Help Center[1] . Useful hints were found in the Best Practices Documents[2] and the DDI 3.0 Proof of Concept materials[3]. Source material included also FSD's existing DDI 2.1 markup in Finnish and in English.

In the first stage of the markup, the CESSDA Metadata Core[4] was used as the base for the study level documentation. We also explored DDI 3 Editor Lite[5], an authoring tool created at ICPSR, ODaF DeXtris[6], an open source utility that facilitates the use and understanding of XML files, DDI DExT[7], the product of a collaborative project between UKDA and the Open Data Foundation ODaF, and the StatsProgs2DDI[8] program created at GESIS. For a summary of our experiences see the table below.

Tool	**Description**	**Comment**
CESSDA Metadata Markup Core	List of CESSDA recommended elements in DDI 3 presented in the form of an XML template ready to use for creating markup.	Gave a good starting point, is 2.0 compatible. No group level metadata information was typed in using an XML editor.
DDI 3 Editor Lite V.2.2	An authoring tool produced by ICPSR that supports the production of DDI 3.0 Instances. It generates DDI 3.0-XML markup providing basic study and variable-level descriptions of simple, survey-type datasets.	Makes DDI 3 invisible to the user, so creating a DDI 3.0 file was easy. Allowed the documentation of the data lifecycle. Our test file included basic study information and a couple of variables, concepts, organisations, and

[1] http://gandalf.opendatafoundation.org/infocenter/index.jsp

[2] http://www.ddialliance.org/resources/publications/working/bestpractices

[3] http://www.ddialliance.org/specification/proof-of-concept

[4] http://www.ddialliance.org/sites/default/files/cessda-rec_0.pdf

[5] http://www.ddialliance.org/DDI 3/editor-lite-3/DDI_v2dot2.html

[6] http://www.opendatafoundation.org/tools/dextris/

[7] http://tools.ddialliance.org/?lvl1=product&lvl2=dext

[8] http://db.zuma-mannheim.de/DDI/StatsProgs2DDI/StatsProgs2DDI.html

		individuals. The file was later used as a template and a guide. Not possible to group studies.
ODaF DeXtris V. 2007.03	DeXtris is an open source utility whose objective is to facilitate the use and understanding of XML files. The conversion utility included in DeXtris is a prototype that provides direct mappings of DDI 1/2.x elements to DDI 3.	Was used to convert existing DDI 2.1 study description into 3.0. It is possible to browse question, variable, and code schemes, making it easier to understand DDI 3 structure and how things are linked. Does not take full advantage of the DDI 3 specification, for example categories are duplicated, so post-processing is needed.
DExT Tools V. 1.0 M1	The application is based on the DDI specification and supports metadata export to DDI 2.0 and DDI 3.0. This initial product is a proof of concept and focuses on SPSS as the input data format with data export capabilities to ASCII, SAS, Stata, and SPSS.	Exported a DDI 3.0 file from our SPSS sav file. The user interface showed the record layout nicely. We used the VariableScheme as the starting point for our markup, although variable statistics were not included.
StatsProgs2DDI 3.0 V. 2007-05-09, revision 130	StatsProgs2DDI 3.0 is a tool to generate variable-level documentation in DDI 3.0 format by converting statistical package system files.	We used StatsProgs2DDI 3.0 to generate variable statistics. No user interface; an XSLT processor and SPSS knowledge required, as well as some manual post-processing.

All these tools provided very helpful insights into DDI 3 but it soon became evident that no single tool would suffice and in order to take full advantage of DDI 3 functionalities (like re-using categories and question texts) manual post-processing would be needed. Although we were able to utilize the tools for separate sections of markup, eventually the main work was carried out manually, with information being either typed in, or copied and pasted. This approach helped us to gain more in-depth knowledge of DDI 3, although it did cause some frustration on the way.

For the use case, we consulted widely with others to understand the use of DDI 3 in an archive and in this paper we present some of the advice gathered and lessons learned. The following questions/design choices will be addressed by this use case:

1. Grouping Decisions
2. Inheritance
3. Defining a Group vs. defining a Series
4. Using URNs and IDs
5. Working with different languages
6. Identifying Terms from Vocabularies and Code Lists
7. Concepts, Subjects, Keywords
8. Use of OtherMaterial

POSSIBLE SOLUTIONS and DESIGN CHOICES

Grouping Decisions

Groups are used to document related data. They may contain subgroups or individual study items, and may also contain metadata applicable to all members of the group, inherited down the grouping structure, and subject to overrides at the lower levels.

When determining how to use the group module, we used the following guidelines:

- Start by defining your basic study units
- Look at the similarities and differences between your files
- Are there enough similarities to create a group?

When creating a group, the basic principle is that metadata changes are documented at lower levels in the specification. The same set of studies can be grouped in various ways, for example: according to topics, or by country, or time period. In all these cases, the end structure may be very similar. The decision as to which way to go will depend on the usefulness of the groups. The grouping decisions in this use case come from a "process approach", but if we look at the data from an "information delivery perspective" or from the researchers' perspective, the resulting group organization could be different. In our use case, the most stable information is the country level metadata, so the group was defined as "Finnish ISSP studies" with subgroups by module.

When planning groupings, building, and using a decision tree may be helpful. Remember to consider the data lifecycle and how it fits into the big picture and try to avoid making decisions before determining the usefulness of the group and reviewing possible tools. Organizational best practices will play a major role here. They should outline what can and cannot be supported between institutions and partners.

DDI 3 is much broader than DDI 2, which may escalate problems from the technical side. You need to look at constraints when deciding what to use in your DDI instance and decide what will work in most cases. Note that a tool is often developed for one perspective – grouping by Geography, for example, but it is possible to re-group the information by some other factor. One could create a profile to match the needs for a tool, fill in all the appropriate attributes, and make changes according to the level of the group breakout.

One approach to deciding whether grouping will work in your situation is to build from what you have:

1. Look at your base file and how you might create a group suitable to its content and relationship to other files.

2. Look at your next unit – what is common with the first one you worked with? Can you use a group module? Document similarities as well as any differences that exist between your first dataset and the second.

3. Think about grouping via schemes (e.g., question and variable schemes – these schemes are maintainable, versionable, and identifiable in DDI 3)

Grouping at a high level is not always the answer. In our use case, the survey consists of two kinds of questions: the ISSP core questions and the Finnish specific questions. (For a visual representation, please refer to the diagram above.) One might think of creating two groups based on these two kinds of questions.

However, in this particular case, this is not an issue of creating groups but rather a question of creating different question schemes and the appropriate references.

In this case it makes the most sense to start by creating a question scheme with the ISSP core set of questions and a second scheme for the Finnish specific questions. You can have separate schemes for each set of questions as identified above, or incorporate the questions within the same scheme. When deciding which route to take, try to anticipate your needs in an attempt to make your life easier in the future. For instance, if you know that in the future your core ISSP question set may be on a different schedule than the Finnish specific questions, then it might be easier to have the questions in two separate schemes. This way it will be easier to move things out in a subset, in other words it is easier to move blocks around rather than restructure groups and break up schemes. By breaking up a scheme you will encounter issues with IDs and versioning; moving a block is much easier to deal with. When deciding how to proceed, look at what could change and what will probably not change. Group and/or package items in a way that makes it easier to manage especially with respect to versioning and reconfiguration.

Note also that a maintainable ID and item ID are tough to break up – avoid at all costs.

Inheritance

Metadata is inherited down the grouping structure. For instance, if there is an abstract for the study unit belonging to a group, the subgroup will inherit it. If however, the documentation changes for one part of the group, then you can override the element by adding the new information at the subgroup level. If you are working with an identifiable item, then the "action" attribute can be used.

Using the "action attribute"[9]

The "action" attribute is used to indicate that the element (referred to as the 'object') being described is being added, updated, or deleted at the local level. This applies to all objects that are either Identifiable, Versionable, or Maintainable.

ADD: Object is added to the inherited structure.

UPDATE: The listed properties (sub-elements) of the object are to be used in place of the properties of the same type in the inherited object with this ID for local processing.

DELETE: The listed properties of the object have been removed from the inherited group for local processing. There will be a new ID only in the case of ADD.

Note that the purpose of @action is to provide a non-inheritable change to the metadata. This lets you make a "one-off" change without having to create a new sub-group. For instance:

- A study added a single variable in one study: use ADD
- A study used all except one variable: use DELETE
- A study used a variable but gave it a different name: use UPDATE

[9] The information in this section reflects the content of the Corrigendum issued on October 1, 2010, which clarifies the use of the action attribute in DDI 3.1.

In the case of ADD you are creating a new variable with a new ID and a full set of properties that will be used in the single study and not inherited in later studies (If later inherited, you would create a subgroup). For DELETE you provide the action and the ID of the variable not used and its full property set. If you want to document a change to an existing variable, use UPDATE and then provide a variable with the ID of the variable that is being edited. Then change whatever you need to, listing all elements of the specified property type that are available at the local level; properties not listed from the original would remain the same. For example, an update of the English language content of the property VariableName that is inherited in multiple languages must include all available language versions of the VariableName that are available for use at the local level, even if the other languages have not changed. Other properties such as Label or Description that are not listed will be inherited from the group.

If, however, you are introducing a change in a variable you want to inherit from this point on, create a new subgroup that uses the previous scheme and create a new variable, same ID, but new version:

```
<VariableScheme id="VS_1" version="1.1.0">
 <VariableSchemeReference>
  <ID>VS_1</ID>
  <Version>1.0.0</Version>
  <Exclude>
   <ID>Var_1</ID>
   <Version>1.0.0</Version>
  </Exclude>
 </VariableSchemeReference>
 <Variable id="Var_1" version="1.1.0">
 <!-- changed variable here -->
 </Variable>
</VariableScheme>
```

Now everything in this new subgroup will use all of the version 1.0.0 variables except Var_1 which is version 1.1.0.

It is worth remembering that inheritance is about information that someone intended to be the same and that you can also take advantage of the comparison module to capture differences.

Defining a Group vs. Defining a Series

The series statement contains information about the series to which a study unit or group of study units belongs. One may point to the URL of a series repository and then use the Name field to indicate the series itself as identified in that repository. Fields also exist for describing the series and providing abbreviations.

A series can be expressed as a group; the ISSP could be viewed as a series. Try to think in terms of a citation. Other forms of series include Sage Publications series or a series of studies conducted by a funding agency. If the studies in the series have nothing more in common than the funding agency, there is no need for inheritance.

A series is not necessarily comparable even though the data files can be expressed as a group. Individual studies can be documented as part of a series whereas grouping is a means of describing them in terms of inheritability.

Referenceable / Reusable Instances / Modules

As the use case progressed, the need to create FSD's own local resource packages for referenceable and reusable items became evident. Much of the metadata in the Group and Archive modules, as well as Categories and Questions, would be common to many FSD studies. It is easier to maintain the common information in one place and include it by reference in the study descriptions.

Useful Resource Packages for FSD would include, for example, r:OrganizationScheme that contains descriptions of organizations and individuals referenced in other sections of the DDI, and r:Coverage, which describes the temporal, geographic, and topical coverage of the study.

Packages of reusable metadata should be maintainable objects, i.e., documentation that is maintained by a specified agency, is versionable, and can be referenced (is identified). DDI is a tiered system and since there are a lot of Identifiable items these will provide flexibility in how things are maintained. All maintainable, versionable and identifiable objects implicitly have an agency, a version, and an ID.

Using URNs and IDs

The modular structure of DDI 3 relies on referencing objects by using their unique identifiers. All identifiable objects are reusable. There are two ways to provide identification for a DDI 3.0 object: using a set of XML fields or using a specially-structured URN. The structured URN approach is preferred. The URN is made up of several parts including, at a minimum, the agency ID and version once the document is published. The best way to deal with IDs and URNs is to use a tool to create them (See also: DDI Best Practices: Management of DDI 3.0 Unique Identifiers[10]). In our markup we created IDs manually; URNs were neither created nor explored.

Working with Different Languages

Working in a multilingual environment will lead to several questions, some of which may require an organizational policy and decision. In our use case, the original ISSP questionnaire was in English. The questionnaire was translated first into Finnish and then into Swedish (the version spoken in Finland). In the archival process, FSD adds metadata in Finnish and in English, but not in Swedish.

Adding question text for different languages should be done separately. Within the same identified QuestionItem element the question text may be repeated in different languages that are identified by the xml:lang attribute.

```
<d:QuestionItem id="K1">
 <d:QuestionText xml:lang="fi">
   <d:LiteralText><d:Text>Sukupuolenne?</d:Text></d:LiteralText>
 </d:QuestionText>
 <d:QuestionText xml:lang="sv-FI">
   <d:LiteralText><d:Text>Ert kön?</d:Text></d:LiteralText>
 </d:QuestionText>
</d:QuestionItem>
```

[10] http://ddi.icpsr.umich.edu/sites/default/files/bp/DDIBestPractices_ManagementOfDDIIdentifiers.doc.pdf

Using the language attribute creates a tight bonding and makes it easier to pull out information in an automated way.

It is possible to use the attributes "translatable" and "translated" in QuestionText. By default the attribute "translatable" is set to true, and "translated" to false. If an item is translated it should link to the source of the translation. This can also be accomplished with grouping. Recognize that most people are probably not interested in whether the question was translated or not; they'll simply pick the language they need. Keep in mind that there are also items that you should not translate. For example: SAS missing codes - human readable vs. machine actionable – should not be translated.

Identifying Terms from Vocabularies and Code Lists

This question came up when documenting subject and keywords. The elements to use are r:Keyword and r:Subject. Both may contain a string, so our first solution was to add the chosen term into the element. When documenting in Finnish and in English, the result was:

<r:Subject xml:lang="fi">sosiologia</r:Subject>

<r:Subject xml:lang="en">sociology</r:Subject>

This markup is problematic. The terms "sosiologia" and "sociology" have the same meaning, but there is no way a machine can deduce it from the above information. They are just words.

The solution is to use Controlled Vocabularies (CVs). Both r:Keyword and r:Subject are of the type InternationalCodeValueType, i.e., they provide a code value and not the term itself, and a reference to the code list from which the value is taken. If we had a controlled vocabulary where the terms would have codes, labels, and descriptions in multiple languages, we could refer to the term by using its code:

<r:Subject xml:lang="fi" codeListID="FSD_topic_classification" codeListAgency="FSD">term40</r:Subject>

When using the language attribute "fi", we are indicating use of the Finnish label for "term40". The attribute "lang" is optional. If you want to retrieve all the labels in all languages, do not designate an xml:lang (you can always reference a specific language if you are referencing this item from somewhere else).

An organization can either use existing CVs or create its own vocabularies. A CV can be published as an independent resource using Genericode, a standard format for defining code lists. DDI 3 allows the use of Controlled Vocabularies in several places. Some CVs are embedded in the DDI already, and the Controlled Vocabularies Working Group of the DDI Alliance is developing further vocabularies, which will be published in Genericode on the DDI Web site.

CVs provide the possibility to check the validity of the entry, thus adding to the quality of the metadata. CVs also facilitate finding comparable data. In addition, multilingual CVs can help automate the translation of metadata and facilitate queries in different languages (see also: Best Practice Document about Controlled Vocabularies[11]).

Concepts, Subjects, Keywords

All variables can point to a concept. DDI 3 concepts are derived from the ISO11179[12], which is a model for managing concepts and data elements. A concept may be a person, place, or thing and any aspect about

[11] http://www.ddialliance.org/bp/DDIBestPractices_ControlledVocabularies.doc.pdf

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

them. Ideally concepts should be defined by the researcher, and not the data archive. For this use case we created a short ConceptScheme, listing a few concepts that relate to the data we are describing. Each QuestionItem contains a ConceptReference. The Variable may inherit the concept by pointing to a question (with the QuestionReference element).

When describing multiple studies, it is advisable to create an external Resource Package containing the ConceptScheme, and include the concepts by reference.

It is worth noting that even if two variables (or questions) relate to the same concept, they are not necessarily identical. For example, income is a concept but not all income questions or variables are identical.

In DDI 3, Subjects indicate the topical coverage of the data described in a particular module/section, and keywords are meant to support searches on topical coverage. Subject classification schemes and keywords exist primarily for external search engines, provide insight to the contents, and create additional means of accessing data. Subject terms are usually a tightly structured set of terms, whereas keywords may be less structured. Subjects and keywords may be used to describe topical coverage at several levels in the specification: Group, Study Unit, Data Collection, Logical Product, etc.

Use of OtherMaterial

OtherMaterial is used to reference external resources related to the content of the relevant module. It includes a citation, an external reference using a URL (or other URI), and a reference to the item within the module to which the external resource is related. Note that OtherMaterial may be used in several places. Anticipating future needs is the key. If the list of other materials is expected to be dynamic, you may wish to physically separate it.

In our use case, for instance, we used OtherMaterial to point to the Finnish ISSP project Web site from the Group module and to add information about publications related to the study. Because publications will be added from time to time, we chose to add publication information in a:Archive/r:OtherMaterial to provide ease of updating without having to version the s:StudyUnit maintainable object. Note that Dublin Core tags are available but not required here since the citation elements include DC core elements.

There are also a number of elements in DDI 3 that are of "OtherMaterial" type. In addition to the generic listing of OtherMaterial in most modules, some elements use the structure of OtherMaterial within a specific domain. For example, the following elements are of the "Other Material" type:

- d:ExternalInterviewerInstructionReference uses r:OtherMaterial as an extension base allowing an external reference to an interviewer instruction that is held externally in a non-DDI format
- d:ExternalInformation (in d:GenerationInstruction)
- d:ExternalAid (in d:ControlConstruct)
- l:Generation (in l:Category)

OtherMaterial allows language differentiations, so you can also have country-specific extensions. Whenever working with related materials, use Relationship to indicate that the material is related and RelationshipDescription to describe the nature of the relationship.

ISSUES

Here we give examples of some detailed markup issues that arose while working with this use case; these kind of questions may easily come up when a new user starts to examine DDI 3. To help users solve these kinds of problems, the DDI Alliance maintains a DDI Users' Listserv, which allows people who are interested in the development and implementation of the DDI specification to communicate with one another and with the committee that is guiding the development of the specification.

r:Creator: no way of telling the language of affiliation

We added information about authors to the element r:Creator. This in turn created a problem, since there is an xml:lang attribute, but no way of specifying the language of the affiliation, so there is a need to repeat:

```
<r:Creator xml:lang="fi" affiliation="Tampereen yliopisto">Blom, Raimo</r:Creator>
<r:Creator xml:lang="en" affiliation="University of Tampere">Blom, Raimo</r:Creator>
```

The code is technically correct because if someone was pulling information and wanted the "English version," they would get the second entry; if they were looking for Finnish, they would only get the first. However, a better tie-in to the Organization scheme content would be beneficial.

The problem was caused by structural constraints - r:Creator and r:Contributor map to Dublin Core. The best solution is to continue to use multiple entries for documenting affiliation in different languages, although it becomes repetitive. The name of the organization can then be used to search an Organization scheme. Think of Organization Schemes as Resource Packages for creators. Affiliation would then be interpreted as a reference to the Organization scheme. Ideally you would want to describe the individuals involved in the study and the roles they played, and not restrict the content of the element to the agency they are affiliated with. The same principles apply to r:Publisher and r:Contributor.

g:Abstract/r:Content (and other r:Content elements): Only one Content permitted, how to mark different language versions

It is the element r:Content – not Abstract – that contains the language attribute. However this attribute should perhaps be in Abstract, not in Content. This and a number of other language-related issues that came up during our use case have been reported to the Technical Implementation Committee (TIC).

r:Copyright: Not possible to use xhtml or make a reference

This is not possible since it is a statement. It is recognized that this could be a long piece of text with logos, but this is something that translates to Dublin Core. It needs to translate to simple DC that supports no formatting. Another problem is that Copyright is not repeatable, thus it can only be specified in one language.

a:Telephone and a:URL have attribute "privacy" but a:Email does not

Email should also have a privacy attribute. This is a bug to be corrected.

d:ControlConstructScheme

The Control Construct consists of a series of elements used to describe the sequence and flow of questions and supporting information within a data collection instrument. One of the Control Construct elements is QuestionConstruct, which in turns holds information about each survey question, like ResponseUnit, AnalysisUnit,

and UniverseReference. These could easily be the same for each question, but they cannot be inherited, thus the need to heavily repeat them for each QuestionConstruct.

In this particular case, inheriting information is a dangerous way to go – once you've removed something from context there is loss of information. The way to deal with this kind of situation is to find a tool that will handle default values. Example: in surveys the unit from which the responses are obtained is usually a person, so you would want the tool to say that the default value for ResponseUnit is person.

OUTLOOK / CONCLUSION

The intended uses for metadata influence the way an organization should approach and use DDI 3. Planning before markup is essential. Think, for example, about how and what to group; what information is common and can be inherited; and what information is reusable and could be gathered into Resource Packages. Organizational practices, needs, and constraints will affect decisions. There are different structures and several possible ways to arrange your metadata, to optimize the gains.

In our use case, we were able to import all of the information from FSD's DDI 2.1 markup to DDI 3 – and what is more important, we were able to add information that we have previously had only in our operational database, not in DDI 2.1 files. The added metadata is mainly information that is common for most/many FSD studies, so it could be incorporated into Resource Packages (Question Schemes, Organisation Schemes, vocabularies). We also welcome the support for controlled vocabularies and multilinguality. DDI 3 will allow much better data management than our current DDI 2.1-based system and eventually better services for the researcher community.

On the other hand, DDI 3 is much broader and more flexible than DDI 2, which in turn may escalate problems on the technical side, especially for smaller data archives. From a programming perspective it is easier to parse an XML specification when it is tightly bound. The flexibility available in the DDI can be tough to work with from a computing perspective.

Our markup consisted of 4000+ lines. With all the referencing, and the need to create unique IDs, the markup is not easy to create, read, or process manually. Tools are needed, or rather a toolbox containing tools that work together: different tools for different phases of the life cycle and for different purposes (for example, for creating URNs and adding default values when information needs to be repeated).

From an archiving perspective, DDI 3 seems a bit of overkill. For long-term preservation, the complexity could be reduced, even if it means losing functionality. On the other hand, DDI 3 opens up possibilities for curation and preservation, because it allows information to be captured throughout the life cycle.

The number of studies and metadata that FSD holds is increasing, our database system needs to be updated, and in the future we need to produce interoperable metadata for various purposes. At the moment we think that in the long term, there would be payoff in moving to DDI 3. However, further planning and resources are needed.

Finally, this use case has shown that when exploring DDI 3, you should not let the complexities frighten you - the DDI Community is ready and willing to help.

APPENDIX A

The paper is one of several papers which are the outcome of a workshop held at Schloss Dagstuhl - Leibniz Center for Informatics in Wadern, Germany, November 2-6, 2009.

Workshop title:
Workshop on Implementation of DDI 3 - Advanced Topics

Organizers:
Arofan Gregory (Open Data Foundation, Tucson, Arizona, USA)
Wendy Thomas (Minnesota Population Center, University of Minnesota, USA)
Mary Vardigan (Inter-university Consortium for Political and Social Research [ICPSR], University of Michigan, USA)
Joachim Wackerow (GESIS, Leibniz Institute for the Social Sciences, Germany)
Link: http://www.dagstuhl.de/09452

This series was edited by Michelle Edwards, Larry Hoyle and Mary Vardigan.

The authors of the paper would like to acknowledge others who participated in this workshop.

Alerk Amin, CentERdata, Tilburg University, the Netherlands
Michelle Edwards, University of Guelph, Canada
Bryan Fitzpatrick, Rapanea Consulting, United Kingdom
Oliver Hopt, GESIS, Leibniz Institute for the Social Sciences, Bonn, Germany
Larry Hoyle, Institute for Policy and Social Research, University of Kansas, USA
Sanda Ionescu, Inter-university Consortium for Political and Social Research (ICPSR), University of Michigan, USA
Jannik Jensen, Dansk Data Archive (DDA), Denmark
Uwe Jensen, GESIS, Leibniz Institute for the Social Sciences, Köln, Germany
Mari Kleemola, Finnish Social Science Data Archive (FSD), University of Tampere, Finland
Dan Kristiansen, Dansk Data Archive (DDA), Denmark
Agostina Martinez, University of Cambridge, United Kingdom
Martin Mechtel, Institute for Educational Progress, Humboldt-Universität zu Berlin, Germany
Olof Olsson, Swedish National Data Service (SND), Sweden
Ørnulf Risnes, Norwegian Social Science Data Services (NSD), Norway
Wolfgang Zenk-Möltgen, GESIS, Leibniz Institute for the Social Sciences, Köln, Germany

APPENDIX B