Project Document Cover Sheet

Project Information

Project Acronym	DataShare		
Project Title	DISC-UK DataShare		
Start Date	March 2007	**End Date**	March 2009
Lead Institution	EDINA, Edinburgh University		
Project Director	Peter Burnhill, Mark Brown		
Project Manager & contact details	Robin Rice, Edinburgh University Data Library, Main Library Bldg., George Square, Edinburgh EH8 9LJ R.Rice@ed.ac.uk 0131 651 1431		
Partner Institutions	Universities of Edinburgh, Oxford, Southampton and London School of Economics		
Project Web URL	http://www.disc-uk.org/datashare.html		
Programme Name (and number)	JISC Repositories and Preservation Programme: Repositories Start-up and Enhancement projects strand		
Programme Manager	Andrew McGregor		

Document Name

Document Title	The Data Documentation Initiative (DDI) and Institutional Repositories		
Author(s) & project role	Luis Martinez, LSE, Project Officer		
Date		**Filename**	
URL	http://www.disc-uk.org/publications.html		
Access	□ Project and JISC internal	☑ General dissemination	

Document History

Version	Date	Comments
1	1 Feb 2008	
2	8 Feb 2008	Minor revisions
3	28 Feb 2008	Addition to acknowledgements

Table of Contents

Introduction

The Data Information Specialists Committee - UK (DISC-UK) is a group of data librarians/managers which has been sharing data support experiences and expertise since 2004. Its members are from the Universities of Edinburgh, Oxford, Southampton and the London School of Economics.

The DISC-UK DataShare project aims at improving current institutional models for dealing with curation of research generated data by exploring the possibilities of using institutional repository (IR) technologies and practices. This project brings together the distinct communities of data support and institutional repositories to bridge gaps and exploit the expertise of both to advance repository services' ability to accommodate datasets.

Purpose of the document

One of the key issues present in IRs is dealing with the description of items held in the repository. Data are not different from other digital materials and need to be described not just for discovery but also for preservation and reuse. The social science data archiving community has been working for many years on a metadata standard to describe datasets, and a new version is about to be published, the Data Documentation Initiative (DDI) 3.0.

This document reports back from the DDI 3 workshop "*Using DDI 3.0 to Support Preservation, Management, Access and Dissemination Systems for Social Science Data*" held at the Schloss Dagstuhl in Germany in November 2007. It intends to present the DDI standard to repository managers, data librarians and data managers and provide background information to help them to examine how the DDI fits with developments in their institutional repositories for research-generated data. The report discusses the appropriateness of using the different DDI versions to address the requirements of research data in IRs. It brings together some of the key questions of the DataShare project with regards to access management, linking to other materials and versioning of datasets.

Methodology

The material presented here has been gathered from presentations from the above mentioned workshop, communication with members of the DDI Alliance, and websites of the DDI and the Inter-university Consortium for Political and Social Research (ICPSR).

The document describes the DDI and its *raison d'etre* by firstly introducing DDI versions 1 and 2 and then describing version 3. The technical appendices complement those sections providing real examples of the standard. Finally, some considerations are presented for using DDI to support the management, dissemination and preservation of research data in IRs.

The Data Documentation Initiative

Introduction

The concept behind the DDI emerged amongst the data archival community as the need for a standard to describe social science datasets became apparent. In 1995 a grant funded project led by ICPSR included members of social science Data Archives and statistical data producers. But it was not until February 2003 that the DDI Alliance, an organization to develop the standard, was formed. Currently it comprises a membership of data archives and libraries worldwide, and bodies such as Statistics Canada, World Bank, WHO and Transport for London amongst others.

DDI 1 and DDI 2

The DDI structure to describe social science datasets has gone through several phases. In 2000 DDI 1 was released to deal with simple microdata surveys[1] and DDI 2 in 2003 widened the scope to aggregate data[2], also adding some geographical features to deal with data more geared towards geo-spatial capabilities. These versions of DDI were based on the dataset codebooks: *"documents with information about the structure, contents and layout of a data file"*[3].

DDI 2 makes use of XML as the meta-language to organize the information; it has five main sections:

1 The document description – describing the metadata document and sources used in its creation.

```
<titlStmt>
        <titl>EURO-BAROMETER 10 -- OCTOBER - NOVEMBER, 1978 </titl>
        <subTitl>NATIONAL PRIORITIES AND THE INSTITUTIONS OF EUROPE </subTitl>
        <IDNo agency="ICPSR">7728</IDNo>
</titlStmt>

<AuthEnty affiliation="special adviser to …">RABIER,  JACQUES-RENE </AuthEnty>
<AuthEnty affiliation="Center for Political Studies…">INGLEHART, RONALD</AuthEnty>

<biblCit>Rabier, Jacques-Rene, and Ronald Inglehart. EURO-BAROMETER 10 -- OCTOBER - NOVEMBER, 1978:
NATIONAL PRIORITIES AND THE INSTITUTIONS OF EUROPE [Codebook file]. First ICPSR ed. Ann Arbor, MI: Inter-
university Consortium for Political and Social Research [producer and distributor], 1980.</biblCit>
```

Table 1. Examples of DDI 2 tags for title, author and citation from a Eurobarometer study

[1] Microdata surveys are files containing information about individuals

[2] Aggregated databases contain information that has been aggregated to the national or regional level.

[3] Definition from ICPSR at: http://webapp.icpsr.umich.edu/cocoon/ICPSR-FAQ/0063.xml

2 The study description – containing information about the data collection: how the study can be cited, who collected the data, who distributes it, keywords and abstract.

```
<subject>
                <keyword source="archive">Common Market</keyword>
                <keyword source="archive">European Community</keyword>
                <keyword source="archive">Europe</keyword>
                      ...
                <topcClas vocab="ICPSR Subject classifications"  Source="archive">3. Attitudes
        Toward Regional Integration</topcClas>

</subject>

<abstract> EURO-BAROMETER 10 WAS CONDUCTED BY JACQUES-RENE RABIER, SPECIAL ADVISER TO THE
COMMISSION OF THE EUROPEAN COMMUNITIES, AND BY RONALD INGLEHART OF THE …</abstract>

<sumDscr>
                <collDate date="1978-10" event="start">October 1978</collDate>
                <collDate date="1978-11" event="end">November 1978</collDate>
                <nation abbr="FRA">France</nation>
                <nation abbr="BEL">Belgium</nation>
                      ...
                <geogCover>nine countries forming the European Community in 1978: France,
                Germany, Great Britain, Italy, the Netherlands, Belgium, Denmark... </geogCover>
                <geogUnit>country</geogUnit>
                <anlyUnit>individuals</anlyUnit>
                <universe clusion="I" level="study">the population, aged fifteen years or older, of
                nine nations members of the European Community: France, Germany.. </universe>
                <dataKind>survey data</dataKind>
</sumDscr>

<dataAccs>
       <setAvail media="online">
                 <accsPlac URI="http://www.icpsr.umich.edu">Ann Arbor, Mi.: Inter-university
                   Consortium for Political and Social Research</accsPlac>
                 <collSize source="archive">1 data file + 1 codebook file + Osiris data dictionary +
                 SPSS  data definition statements</collSize>
                 <fileQnty>6</fileQnty>
                 <notes> THE STUDY STAFF FOR EURO-BA ROMETER 10 DEVELOPED AN
                 EQUIVALENT FRENCH AND BRITISH QUESTIONNAIRE FOR THI…</notes>
       </setAvail>
       <useStmt>
                  <citReq> ALL MANUSCRIPTS USING DATA MADE AVAILABLE THROUGH THE
                   CONSORTIUM  SHOULD ACKNOWLEDGE THAT…</citReq>
                 <deposReq>IN ORDER TO PROVIDE FUNDING AGENCIES WITH ESSENTIAL </deposReq>
       </useStmt>
</dataAccs>
```

Table 2. Examples of DDI 2 tags within the study description section from a Eurobarometer study

3 The data files description – including information about the data files, such as format, dimensions, missing data etc…

```
<fileDscr ID="OSIRIS">
        <fileTxt>
                <fileName>odata.s7728.gz</fileName>
                <dimensns>
                   <caseQnty>8677</caseQnty>
                   <varQnty>119</varQnty>
                   <recPrCas>1</recPrCas>
                   <recNumTot>8677</recNumTot>
                </dimensns>
                <format>THERE ARE TWO COMPONENTS TO THE OSIRIS DATASET.THE OSIRIS…
                   <Link refs="NFORMAT"/></format>
                <software>OSIRIS</software>
        </fileTxt>
 </fileDscr>
```

Table 3. Examples of DDI 2 tags for the data files description from a Eurobarometer study

4 The variable description – here each of the variables is described with its values, labels and definitions.

```
<var ID="V9" name="VAR0009" qstn="Q112">
    <location StartPos="28" EndPos="28" width="1" RecSegNo="1" fileid="CARD-IMAGE"/>
    <location StartPos="26" />
    <labl level="variable">LIFE SATISFACTION</labl>
    <qstn ID="Q112" var="V9">
          <qstnLit>ON THE WHOLE, ARE YOU VERY SATISFIED,  FAIRLY SATISFIED, NOT VERY
           SATISFIED, OR NOT AT ALL SATISFIED WITH  THE LIFE YOU LEAD?</qstnLit>
    </qstn>
   <valrng><range min="1" max="4"/></valrng>
   <invalrng><item VALUE="0"/></invalrng>
    <catgry>
           <catValu>1</catValu>
           <labl level="category">VERY SATISFIED</labl>
    </catgry>
    <catgry>
           <catValu>2</catValu>
           <labl level="CATEGORY">FAIRLY SATISFIED</labl>
    </catgry>
   <catgry>
           <catValu>3</catValu>
           <labl level="CATEGORY">NOT VERY SATISFIED</labl>
   </catgry>
   <catgry>
           <catValu>4</catValu>
           <labl level="CATEGORY">NOT AT ALL SATISFIED</labl>
   </catgry>
</var>
```

Table 4. Examples of DDI 2 tags for the variable section from a Eurobarometer study

5 Other study-related materials – this section includes references to related reports and publications.

```
<relStdy> Family, Work and Leisure in the London Region, 1970 : Diaries </relStdy>
<relPubl>By Principal Investigator(s):</relPubl>
<relPubl>Young, M. and Willmott, P., <i>The symmetrical family</i> (London: Routledge and Kegan
Paul, 1973 and Penguin, 1975)
</relPubl>
```

Table 5. Example of DDI 2 tags for other materials section from the "Family, Work and Leisure in the London Region, 1970: Main Study"

DDI 2 Lite

An example of DDI 2 and the complete set of tags is shown at the DDI website http://www.ddialliance.org/DDI/dtd/version2-1-all.html. The full DDI schema features more than 300 tags, most of them optional. In addition, there is a subset of DDI tags that represents a selection of the most relevant elements - known as DDI Lite; it was developed by the Council of European Social Science Data Archives in 2001; see: http://www.ddialliance.org/DDI/related/cessda-rec.pdf. This provides a basic framework for marking up documentation; see appendix 2 for a complete list of DDI Lite elements. A DDI 2 Lite example is shown in appendix 3.

DDI 2 and Dublin Core

The Dublin Core metadata standard is a widely recognized meta-language to describe information resources. It includes two levels; Simple - which contains fifteen elements, and Qualified - which includes three additional elements, as well as a group of element refinements. The DDI Alliance provides a mapping from DDI 2 to Dublin Core, as follows.

DC Element	DDI Element	Notes
Title	<titl> 2.1.1.1	Title of Data Collection
Creator	<AuthEnty> 2.1.2.1	Authoring Entity of Data Collection
Subject	<keyword> 2.2.1.1	Keyword(s)
	<topcClas> 2.2.1.2	Topic Classification
Description	<abstract> 2.2.2	Abstract
Publisher	<producer> 2.1.3.1	Producer of Data Collection
Contributor	<othId> 2.1.2.2	Other Identification/ Acknowledgements - Data Collection
Date	<prodDate> 2.1.3.3	Production Date - Data Collection
Type	<dataKind> 2.2.3.10	Kind of Data
Format	<fileType> 3.1.5	Type of File
Identifier	<IDNo> 2.1.1.5	ID Number - Data Collection
	<holdings location="" callno="" URI=""> 2.1.8	Holdings Information - Data Collection

Source	<sources> 2.3.1.8	Sources - Used for Data Collection
Language		
Relation	<othrStdyMat> 2.5	Other Study Description Materials
Coverage	<timePrd> 2.2.3.1	Time Period Covered
	<collDate> 2.2.3.2	Date(s) of Data Collection
	<nation> 2.2.3.3	Country
	<geogCover> 2.2.3.4	Geographic Coverage
Rights	<copyright> 2.1.3.2	Copyright - Data Collection

Table 6. Mapping from DDI 2 to Dublin Core; from DDI website

DDI 2 and OAI-PMH

The Economic and Social Data Service (ESDS) has an OAI-PMH (Open Archives Initiative - Protocol for Metadata Harvesting) implementation[4] that uses DDI 2 as a metadata format; and also Dublin Core and MARC (MAchine Readable Catalogue). This implementation responds to a ListRecords request, returning a record with most of the information but not including Data Files and Variables descriptions.

As an example, the URLs below display the metadata formats available at ESDS and the DDI record for study number 2473:

- http://oai.esds.ac.uk/oai.asp?verb=ListMetadataFormats
- http://oai.esds.ac.uk/oai.asp?verb=GetRecord&identifier=oai%3Aesds.ac.uk%3AESDS%2FESDSA%2Fsn2473.xml&metadataPrefix=ddi

DDI 3

The main drawbacks of the earlier versions of DDI were that they focused on a static object - the codebook; they were designed for a limited number of uses, like data discovery; the coverage focused on a single study, rather than a collection of studies.

These disadvantages limited the scope of the standard in describing social science datasets. The metadata files were treated as an add-on to the collection, and data producers were hesitant in using DDI as it did not support the development or collection process.
DDI 3.0 came into being with a set of requirements which included the following:

- Improve the machine-actionable aspects to support programming and other information systems
- Support for computer assisted interview (CAI) instruments
- Support for the description of data series.

This represents a shift from the codebook-centric model to a data life cycle model; see figure below. Consequently DDI 3 attempts to provide metadata to support all the stages of the life cycle of a dataset. This starts with the design of the survey and data collection, going through the phases of processing, archiving, distribution, discovery, analysis and repurposing.

[4] http://oai.esds.ac.uk/

DDI 3.0 Lifecycle Model

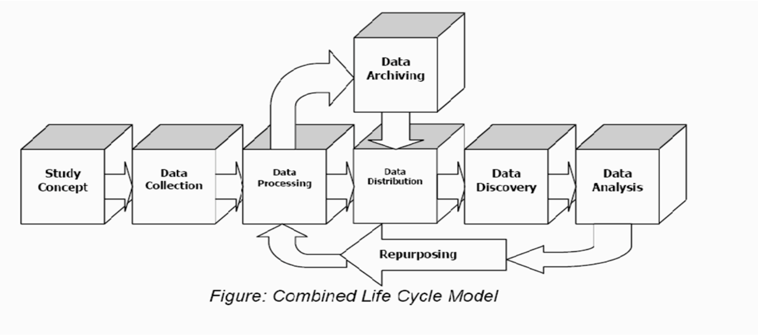


Image 1. The DDI 3.0 Life cycle model; from Dagstuhl workshop slides

To achieve these requirements, DDI 3 has adopted a modular approach (see below) which allows better maintainability as these modules can be held in different XML instances and reference can be made to them from the main instance. This modularity also allows reuse of categories for questions and variables and replacement by substitution.

Some of the advantages of life cycle orientation[5] are:

- Allows capture and preservation of metadata generated by different agents at different points in time
- Facilitates tracking changes and updates in both data and documentation
- Enables investigators, data collectors and producers to document their work directly in DDI, thus increasing the metadata's visibility and usability
- Benefits data users, who need information from the full data life cycle for optimal discovery, evaluation, interpretation, and reuse of data resources.

DDI 3.0 – Modular Structure

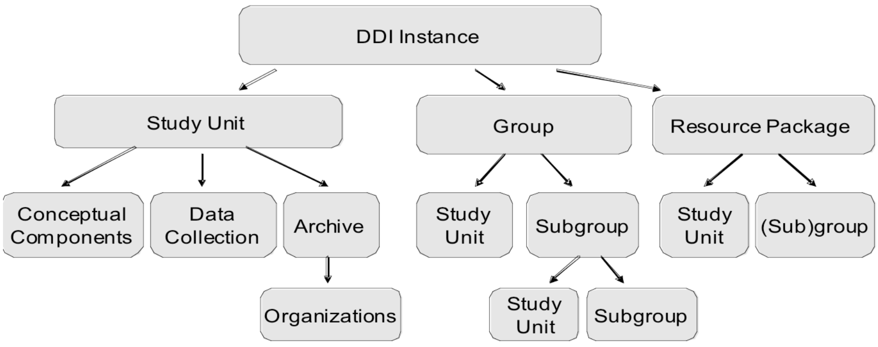


Image 2. DDI 3.0 Modular Structure. From: Introduction to DDI 3. Presentation by Sanda Ionescu, ICPSR at the CESSDA Expert Seminar, September 2007.

[5] Introduction to DDI 3. Presentation by Sanda Ionescu, ICPSR at the CESSDA Expert Seminar, September 2007.

The modules are organized in four sections, as follows.

Firstly, we have the packaging modules which will allow pointing to the different modules, grouping studies and persistent information about the data:

DDI Instance – contains pointers to all other modules and life cycle events.

```
<ns1:DDIInstance xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="ddi:instance:3_0_CR2 ../XMLSchemas/instance.xsd"
xmlns:ns1="ddi:instance:3_0_CR2"
xmlns:a="ddi:archive:3_0_CR2"
xmlns:o="ddi:organization:3_0_CR2"
xmlns:g="ddi:group:3_0_CR2"
xmlns:c="ddi:conceptualcomponent:3_0_CR2"
xmlns:d="ddi:datacollection:3_0_CR2"
xmlns:s="ddi:studyunit:3_0_CR2" >
```

Table 7. Examples of DDI 3 instance tag showing pointers to some of the modules

Group – this allows the grouping of two or more studies.

Study Unit – first module created. It contains the abstract and the purpose of the study and any information needed before the data collection stage. The module is then passed as a persistent module.

```
<s:Abstract>
     <r:Identifier>
          <r:ID>Abs1</r:ID>
     </r:Identifier>
     <r:Content>
         <xhtml:p>The 2004 American National Election Study was conducted by the Center for Political Studies of
           the Institute for Social Research, at the University of Michigan.
        </xhtml:p>
      </r:Content>
</s:Abstract>
```

```
<s:Purpose>
     <r:Identifier>
        <r:ID>Pur1</r:ID>
     </r:Identifier>
     <r:Content>The American National Election Studies are designed to present data on Americans' social
      backgrounds, enduring political predispositions, social and political values, perceptions and evaluations of
     groups and candidates, opinions on questions of public policy, and participation in
     political life.
    </r:Content>
</s:Purpose>
```

Table 8. Example of DDI 3 study unit tags from the 2004 American National Election Study

Secondly, we have a collection of maintainable schemes i.e. packages of reusable metadata maintained by a single agency:

Data Collection – contains information such as sampling procedure, collector, frequency of data collection, instrument used, cleaning, weighting and coding instructions.

```
<d:DataCollection>
   <d:Identifier>
    <r:ID>DataCollection_4245</r:ID>
   </d:Identifier>
   <d:QuestionScheme>
     <d:Identifier>
      <r:ID>QuestionScheme_4245</r:ID>
     </d:Identifier>
     <d:QuestionItem>
      <d:Identifier>
        <r:ID>A7</r:ID>
      </d:Identifier>
      <d:QuestionText>
        <d:LiteralText>
          <d:Text>How many days in the PAST WEEK did you watch the NATIONAL network news
           on TV?</d:Text>
        </d:LiteralText>
     </d:QuestionText>
     <d:CodeDomain>
      <r:CodeSchemeReference>
        <r:ID>CodeScheme_V043014</r:ID>
      </r:CodeSchemeReference>
     </d:CodeDomain>
    </d:QuestionItem>
   </d:QuestionScheme>
</d:DataCollection>
```

Table 9. Examples of DDI 3 data collection module from the 2004 American National Election Study

Conceptual Components – can be used in the Study Unit and Comparative modules. It includes concepts describing the data being documented and universes of respondents.

```
<c:Concept>
     <c:Identifier>
      <r:ID>Concept_1</r:ID>
     </c:Identifier>
     <r:Description>Exposure to national TV news</r:Description>
</c:Concept>

<c:Universe>
     <r:Identifier>
      <r:ID>Univ1</r:ID>
     </r:Identifier>
     <r:HumanReadable>All United States citizens of voting age on or before the 2004 Election
     Day. Eligible citizens must have resided in housing units in the 48 coterminous states. This
     definition excludes persons living in Alaska or Hawaii and requires eligible persons to have been
     both a United States citizen and aged 18 on or before November 2,2004
       </r:HumanReadable>
 </c:Universe>
```

Table 10. Examples of DDI 3 conceptual component module from the 2004 American National Election Study

Logical Product – describes the intellectual content and structure of the data file. It is used for both question response domains and variable representations.

```
<l:Variable>
      <l:Identifier>
        <r:ID>V043014</r:ID>
        <r:Name>V043014</r:Name>
      </l:Identifier>
      <r:Label>Days past week watch natl news on TV</r:Label>
      <l:ConceptReference>
        <r:ID>Concept_1</r:ID>
      </l:ConceptReference>
      <l:QuestionReference>
        <r:ID>A7</r:ID>
      </l:QuestionReference>
      <l:Representation>
        <l:CodeRepresentation>
          <r:CodeSchemeReference>
            <r:ID>CodeScheme_043014</r:ID>
          </r:CodeSchemeReference>
        </l:CodeRepresentation>
      </l:Representation>
</l:Variable>
```

Table 11. Example of DDI 3 logical product module from the 2004 American National Election Study

Reusable – contains a range of classes that can be used throughout the DDI structure.

Thirdly, we have the non-scheme-based modules, which include:

Physical Data Product – provides a way to describe the physical layout of the data documented in the logical product module.

Physical Instance – describes a single data file including its name, limitations of coverage and variable summary statistics. It also lists the number of cases, records and software used for production.

```
<pi:DataFileIdentification>
      <pi:Identifier>
        <r:ID>Data_ASCII</r:ID>
      </pi:Identifier>
      <pi:Location>ICPSR</pi:Location>
      <pi:URI>http://www.icpsr.umich.edu/cgi--in/bob/archive2?path=ICPSR&amp;study=4245
</pi:URI>
 </pi:DataFileIdentification>


<pi:GrossFileStructure>
   <pi:Identifier>
     <r:ID>GrossFileStructure_4245</r:ID>
   </pi:Identifier>
   <pi:CaseQuantity>1212</pi:CaseQuantity>
   <pi:OverallRecordCount>1212</pi:OverallRecordCount>
</pi:GrossFileStructure>
```

Table 12. Example of DDI 3 physical instance tags from the 2004 American National Election Study

Archive – provides information about the organization holding the DDI. It includes local information (URI, format, study class), where it came from and local restrictions such as access restrictions and confidentiality statements.

Comparative – gathers information for comparing concepts, questions and variables between and among Study Units in the same group.

DDI Profile – allows the description of the modules that have been used.

Dataset – allows the expression of the data in an XML format.

Finally, other external XML schemas, that could be used within DDI 3 include:

XHTML – this module enables the use of XHTML markup.

Dublin Core – enables the use of Dublin Core elements.

DDI 3.0 Core

As in DDI 2, there is a set of recommended elements which is a subset of the whole DDI structure - known as DDI 3.0 Core; see appendix 5 for more details.

From earlier versions of DDI to DDI 3

Converting earlier versions of DDI to DDI 3 is possible, as mapping exists between versions[6]. In addition to this, the Open Data Foundation[7] has developed an open source utility, DeXtris[8], that not only facilitates the use and understanding of XML but also allows partial conversion of DDI 2 to DDI 3.

Considerations for institutional digital repositories

Digital repositories can and should take advantage of DDI to classify and organize datasets. This, we believe, is possible in the three most common digital repository systems, DSpace, e-Prints and Fedora. Although DSpace and e-Prints use different metadata schemas, both could be modified to accept and use DDI. Fedora itself poses no problems as it can use many metadata formats as required.

The DataShare project plan has, under the Technology workpackage, a deliverable for using the DDI metadata standard to describe datasets in LSE's e-Prints repository. However, at this stage there are some open questions, such as:

- should we be using DDI 2 or DDI 3?
- how many and which tags are needed at this stage?
- do those versions address the requirements for access management, linking to other publications and preservation?
- in which ways could DDI be progressively incorporated and used in IRs?

In the following sections we will provide a comparison of DDI 2 and DDI 3, showing how they address the requirements for the DataShare project. This should inform repository managers and help them decide which version is more appropriate to their needs.

[6] This can be provided upon request
[7] http://www.opendatafoundation.org/
[8] http://www.opendatafoundation.org/tools/dextris/

DDI 2 or DDI 3?

The DDI website states that the current stable version of the DDI specification is 2; version 3.0 is in the Candidate Draft stage from July 2007 to early 2008. DDI 2 is simpler to use but DDI 3 offers fuller coverage of the phases of the data life cycle providing the flexibility that may be required in the near future.
The table below shows a detailed comparison between Version 2 and Version 3.0[9].

Version 2	Version 3.0
Inadequate representation of complex / hierarchical data	Detailed documentation for complex / hierarchical data
No instrument coverage	Full description of instrument as a separate entity
Question text appears only as part of variable description	Compatible with Computer Assisted Interviewing software
No documentation for question flow / conditions	Documents specific use of questions: flow, conditions, loops
Initially designed for microdata only	Adds support for tabular, spreadsheet-type representation of aggregate data
Aggregate data section added in V 2.1 to support limited representation (Census-type data, delimited files)	Aggregate data transport option: cell content may be included inline with the data item description
No data transport function	Inline inclusion enabled for both aggregate data and microdata
No longitudinal / time series / cross-national data comparability	Grouping structure documents studies related on one or several dimensions (time, geography, language etc.) as well as their comparability
Limited multilingual support	Support for multiple language use and translations
Single file, hierarchical design	Modular design: facilitates reuse, facilitates versioning and maintenance, supports life cycle model, allows flexibility in organizing the DDI Instance, supports grouping and comparing studies, supports creation of metadata registries

There are compelling reasons to use the modular life cycle-based version of the DDI; however, regardless of which version is selected just a few metadata tags will be mandatory. DDI 2 Lite (see appendices 2 and 3) and DDI 3 Core (see appendix 5) provide a set of basic tags which can be expanded by adding other elements as required. DDI 3.0 Core consists of the recommended elements from DDI 2 Lite; see:
http://www.ddialliance.org/DDI/ddi3/V3_MarkupTemplate.xml

[9] Adapted from: Introduction to DDI 3. Presentation by Sanda Ionescu, ICPSR at the CESSDA Expert Seminar, September 2007.

Access Management

Another deliverable of DataShare is using Shibboleth access to data in the LSE repository. Both DDI 2 and DDI 3 have tags that can be reused to deal with the access management requirements.

In DDI 2 this can be achieved through the Data Access tag (<dataAccs>), which describes access conditions. In cases where access conditions differ across individual files or variables, multiple access conditions can be specified.

In DDI 3 the relevant tags are in the Archive module under access type, with tags such as:

- **Access** - describes the aspects of access to the archive's holding
- **AccessConditions** - describes the conditions of access
- **AccessPermission** - gives the network location and identifying number of the access permission and confidentiality agreement forms, and whether they are required or not
- **AccessRestrictionDate** - provides dates for which access is restricted. Describes the date or range of dates for access restrictions to all or portions of the data.

Linking to other materials

Again, both versions of DDI will allow linking to publications. As shown previously, in DDI 2 the other materials section allows linking to related studies, including title of study, author, producer, version and physical location, amongst others. In DDI 3, linking to other materials is provided in the study unit module.

Preservation

An international group sponsored by the Online Computer Library Center (OCLC) has developed a metadata standard to support and document the digital preservation process, PREMIS[10] (PREservation Metadata Implementation Strategies). This standard helps to gather information about provenance, authenticity, preservation activity, technical environment and rights management.

Preservation was one of the main drivers for the creation of DDI 3, which likewise deals with most of with the information required for the preservation of digital materials, and there is a draft mapping from PREMIS to DDI 3[11].

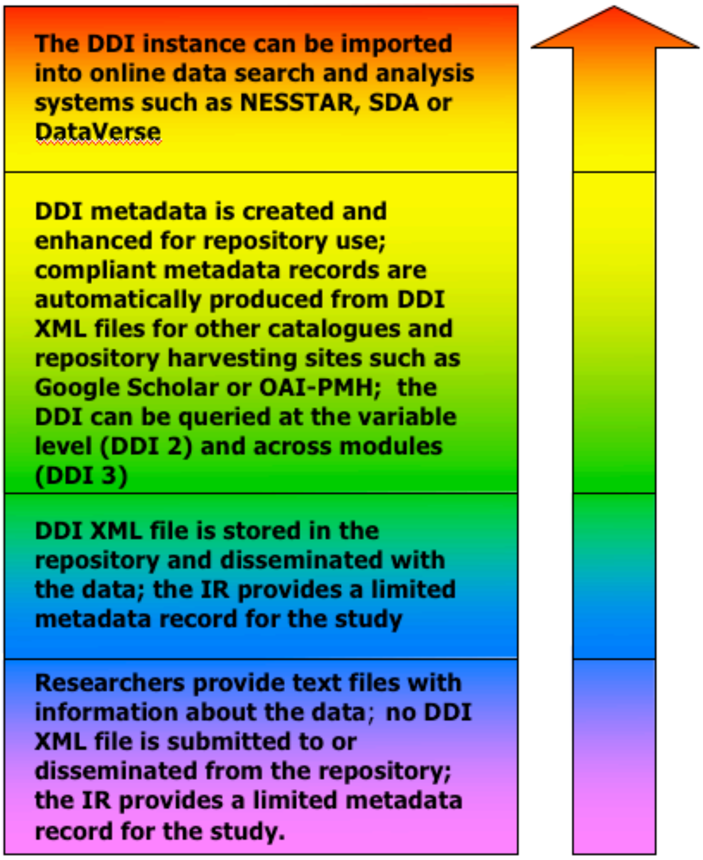
Ways in which DDI could be used in an institutional repository

The following diagram provides a sample of the scenarios in which DDI XML files could be used in IRs. The continuum starts with a situation in which there is no DDI metadata stored in the repository; researchers provide text files to describe the data and the IR produces a limited metadata record at the point of ingest. After this, the research team (or data experts working with the researchers) might create the DDI XML files and deliver them to the IR with the data files. The information in the DDI file/s could be used to populate a more complete IR record for the data. In the next step in the continuum, the DDI content could be integrated into the search system of the repository, allowing queries across the multiple levels of DDI 2 or the modules of DDI 3. In some cases, the IR may choose to integrate the numeric data

[10] http://www.oclc.org/research/projects/pmwg/

[11] This can be provided upon request

into the DDI XML instance. At the top of the continuum, the DDI instances can be imported into data search and manipulation tools such as NESSTAR[12], SDA[13] or DataVerse[14].



The DDI instance can be imported into online data search and analysis systems such as NESSTAR, SDA or DataVerse

DDI metadata is created and enhanced for repository use; compliant metadata records are automatically produced from DDI XML files for other catalogues and repository harvesting sites such as Google Scholar or OAI-PMH; the DDI can be queried at the variable level (DDI 2) and across modules (DDI 3)

DDI XML file is stored in the repository and disseminated with the data; the IR provides a limited metadata record for the study

Researchers provide text files with information about the data; no DDI XML file is submitted to or disseminated from the repository; the IR provides a limited metadata record for the study.

Image 3. Ways of using DDI in an institutional repository

Acknowledgements

I would to like to specially thank Ann Green for all her extremely valuable work on this report providing useful feedback and fantastic ideas; Wendy Thomas for all her support while writing the document and the provision of mappings between DDI and other standards, Sanda Ionescu for her presentations with invaluable information and Jane Roberts for reviewing thoroughly its various versions.

[12] http://www.nesstar.com/
[13] http://sda.berkeley.edu/
[14] http://thedata.org/

Technical Appendices

Appendix 1: Links to important technical documents

DDI 2 example
Eurobarometer 10: http://www.ddialliance.org/cocoon/DDI/SAMPLES/07728.xml

DDI 3 example
2004 American National Election Study
http://www.ddialliance.org/DDI/ddi3/NEW_4245variables.xml

XML Schema Outline - Version 2.1
From http://www.ddialliance.org/DDI/dtd/version2-1-tree.html

XML Schema Outline – Version 3.0
From http://www.ddialliance.org/DDI/ddi3/DDI_3.0_XMLSchemaDocHelp_cr.zip

Appendix 2: DDI 2.1 Lite (CESSDA recommended elements)

? Element is optional and non-repeatable
* Element is optional and repeatable
Element is mandatory

XMLSchemaXMLSchemaOutlineVersion2.1						
	Tag	Attributes				
0.0	codeBook	version				
1.0	docDscr*					
1.1	citation?	MARCURI				
1.1.1	titlStmt					
1.1.1.1	titl					
1.1.1.5	IDNo*	agency	level			
1.1.3	prodStmt?					
1.1.3.1	producer*	abbr	affiliation	role		
1.1.3.2	copyright?					
1.1.3.3	prodDate*	date				
1.1.3.5	software*	date	version			
1.1.6	verStmt*					
1.1.6.1	version?	date	type			
1.1.6.3	notes*	type	subject	level	resp	sdatrefs
2.0	stdyDscr	access				
2.1	citation	MARCURI				
2.1.1	titlStmt					
2.1.1.1	titl					
2.1.1.5	IDNo*	agency	level			
2.1.2	rspStmt?					
2.1.2.1	AuthEnty*	affiliation	type	role	affiliation	
2.1.3	prodStmt?					
2.1.3.1	producer*	abbr	affiliation	role		
2.1.3.2	copyright?					
2.1.3.3	prodDate*	date	date	version		
2.1.3.6	fundAg*	abbr	role			
2.1.3.7	grantNo*	agency	role			
2.1.4	distStmt?					
2.1.4.1	distrbtr*	abbr	affiliation	URI		
2.1.4.5	distDate?	date				

2.1.5	serStmt?	URI				
2.1.5.1	serName*	abbr				
2.1.6	verStmt*					
2.1.6.1	version?	date	type			
2.1.6.3	notes*	type	subject	level	resp	sdatrefs
2.1.7	biblCit?	format				
2.2	stdyInfo*					
2.2.1	subject?					
2.2.1.1	keyword*	vocab	vocabURI			
2.2.2	abstract*	date				
2.2.3	sumDscr*					
2.2.3.1	timePrd*	date	event	cycle		
2.2.3.2	collDate*	Date	event	cycle		
2.2.3.3	nation*	Abbr				
2.2.3.3.1	txt	level	sdatrefs			
2.2.3.3.2	concept	vocab	vocabURI			
2.2.3.4	geogCover*					
2.2.3.4.1	txt	level	sdatrefs			
2.2.3.4.2	concept	vocab	vocabURI			
2.2.3.5	geogUnit*					
2.2.3.8	anlyUnit*	unit				
2.2.3.8.1	txt	level	sdatrefs			
2.2.3.8.2	concept	vocab	vocabURI			
2.2.3.9	universe*	level	clusion			
2.2.3.9.1	txt*	level	sdatrefs			
2.2.3.9.2	concept*	vocab	vocabURI			
2.2.3.10	dataKind*					
2.2.3.10.1	txt	level	sdatrefs			
2.2.3.10.2	concept	vocab	vocabURI			
2.3	method*					
2.3.1	dataColl*					
2.3.1.1	timeMeth*	method				
2.3.1.1.1	txt	level	sdatrefs			
2.3.1.1.2	concept	vocab	vocabURI			
2.3.1.2	dataCollector*	abbr	affiliation			
2.3.1.4	sampProc*					
2.3.1.4.1	txt	level	sdatrefs			
2.3.1.4.2	concept	vocab	vocabURI			
2.3.1.6	collMode*					
2.3.1.6.1	txt	level	sdatrefs			
2.3.1.6.2	concept	vocab	vocabURI			
2.3.1.7	resInstru*	type				
2.3.1.7.1	txt	level	sdatrefs			
2.3.1.7.2	concept	vocab	vocabURI			
2.3.1.8	sources?					
2.3.1.12	weight*					
2.3.1.13	cleanOps*	agency				
2.3.2	notes*	type	subject	level	resp	sdatrefs
2.4	dataAccs*					
2.4.1	setAvail*	media	callno	label	type	
2.4.1.4	collSize?					
2.4.1.6	fileQnty?					
2.4.2	useStmt*					
2.4.2.3	restrctn?					
2.6	notes*	type	subject	level	resp	sdatrefs

3.0	fileDscr*	URI	sdatrefs	methrefs	pubrefs	access
3.1	fileTxt*					
3.1.4	dimensns?					
3.1.4.1	caseQnty*					
3.1.4.2	varQnty*					
3.1.4.3	logRecL*					
3.1.4.4	recPrCas*					
3.1.5	fileType?	charset				
3.3	notes*	type	subject	level	resp	sdatrefs
4.0	dataDscr*					
4.1	varGrp*	type	var	varGrp	name	sdatrefs
		methrefs	pubrefs	access	nCube	
4.1.1	labl*	level	vendor	country	sdatrefs	
4.3	var*	name	wgt	wgtvar	weight	qstn
		files	vendor	dcml	intrvl	rectype
		sdatrefs	methrefs	pubrefs	access	aggrMeth
		measUnit	scale	origin	nature	additivity
		temporal	geog	geoVocab	catQnty	
4.3.1	location*	StartPos	EndPos	width	RecSegNo	fileid
		locMap				
4.3.2	labl*	level	vendor	country	sdatrefs	
4.3.8	qstn*	qstn	var	seqNo	sdatrefs	
4.3.8.2	qstnLit	sdatrefs				
4.3.10	invalrng*					
4.3.10.1	range	UNITS	min	minExclusive	max	maxExclusive
4.3.12	universe*	level	clusion			
4.3.12.1	txt*	level	sdatrefs			
4.3.12.2	concept*	vocab	vocabURI			
4.3.14	sumStat*	wgtd	wgtvar	weight	type	
4.3.15	txt*	level	sdatrefs			
4.3.17	catgryGrp*	missing	missType	catgry	catGrp	levelno
		levelnm	compl	excls		
4.3.17.1	labl*	level	vendor	country	sdatrefs	
4.3.17.2	catStat*	type	URI	methrefs	wgtd	wgtvar
		weight	sdatrefs			
4.3.18	catgry*	missing	missType	country	sdatrefs	excls
		catgry	level			
4.3.18.1	catValu?					
4.3.18.2	labl*	level	vendor	country	sdatrefs	
4.3.18.3	txt*	level	sdatrefs			
4.3.18.4	catStat*	type	URI	methrefs	wgtd	wgtvar
		weight	sdatrefs			
4.3.21	concept*	vocab	vocabURI			
4.3.22	derivation?	var				
4.3.22.1	drvdesc?					
4.3.23	varFormat?	type	formatname	schema	category	URI
4.3.25	catLevel*	levelnm				
4.3.26	notes*	type	subject	level	resp	sdatrefs
5.0	otherMat*	type	level	URI		
5.1	labl*	level	vendor	country	sdatrefs	
5.2	txt?	level	sdatrefs			
5.3	notes*	type	subject	level	resp	sdatrefs
5.4	table*	frame	colsep	rowsep	pgwide	
5.5	citation?	MARCURI				
5.6	otherMat*	type	level	URI		

Appendix 3: DDI 2 Lite example (CESSDA recommended elements)

Eurobarometer 10 at: http://www.ddialliance.org/cocoon/DDI/SAMPLES/07728.xml

Appendix 4: DDI 2 example with only DC elements

Title:	EURO-BAROMETER 10 -- OCTOBER - NOVEMBER, 1978
Authoring Entity:	RABIER, JACQUES-RENE (special adviser to the Commission of the European Communities)
	INGLEHART, RONALD (Center for Political Studies, the University of Michigan)
Keywords:	Common Market, European Community, Europe, France, Belgium, Denmark, Ireland, Great Britain, Italy, Luxembourg, Netherlands, Federal Republic of Germany, European integration, European Parliament, Political issues, Political attitudes and behavior, Voting attitudes and behavior
Topic Classification:	XIV. Mass Political Behavior and Attitudes, C. Public Opinion on Political Matters, 3. Attitudes Toward Regional Integration, a. Europe
Abstract:	EURO-BAROMETER 10 WAS CONDUCTED BY JACQUES-RENE RABIER, SPECIAL ADVISER TO THE COMMISSION OF THE EUROPEAN COMMUNITIES, AND BY RONALD INGLEHART OF THE UNIVERSITY OF MICHIGAN. THIS STUDY IS PART OF AN ONGOING PROGRAM OF PUBLIC OPINION RESEARCH SPONSORED BY THE EUROPEAN COMMUNITY. THE FIELDWORK WAS CARRIED OUT BY A CONSORTIUM OF EUROPEAN POLLING ORGANIZATIONS IN ALL NINE NATIONS OF THE EUROPEAN COMMUNITY. RESPONDENTS FOR EURO-BAROMETER 10 WERE INTERVIEWED IN OCTOBER-NOVEMBER, 1978. THIS STUDY CONTAINS AN EXPANDED VERSION OF THE QUESTIONS ON EUROPEAN INTEGRATION THAT HAVE BEEN ASKED THROUGHOUT THE EURO-BAROMETER SERIES. PERCEPTIONS OF RECENT CHANGES IN THE EXTENT OF INTEGRATION AND UNDERSTANDING AMONG THE COMMON MARKET COUNTRIES ARE EXPLORED, AS ARE ATTITUDES TOWARD THE FORTHCOMING EUROPEAN PARLIAMENTARY ELECTIONS. EURO-BAROMETER 10 ALSO CONTAINS AN EXPANDED SECTION ON THE PROBLEMS WHICH RESPONDENTS FEEL SHOULD BE GIVEN POLITICAL PRIORITY IN THE COMING YEARS. THESE ISSUES ARE PROBED BOTH IN TERMS OF THEIR IMPORTANCE AND IN TERMS OF WHETHER THE NATIONAL GOVERNMENTS OR THE EUROPEAN COMMUNITY AS A WHOLE MIGHT BETTER DEAL WITH THEM. THE PERSONAL DATA SECTION OF THE INTERVIEW OBTAINED INFORMATION ABOUT THE EDUCATION, OCCUPATION, MARITAL STATUS, AGE AND SEX OF THE RESPONDENT. THIS SECTION ALSO ASCERTAINED THE OCCUPATION OF THE HEAD OF THE HOUSEHOLD AND COMPOSITION OF THE HOUSEHOLD.
Producer	INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH
Production Date	
Kind of Data:	survey data
File type	OSIRIS
Identifier	
Sources	
Language	English
Other Study Material	Commission of the European Communities. Euro-baromètre 10. Brussels: Commission of the European Communities, January 1979.
Date of Collection:	October 1978-November 1978
Country:	France, Belgium, Netherlands, Germany, Italy, Luxembourg, Denmark, Ireland, Great Britain
Geographic Coverage:	nine countries forming the European Community in 1978: France, Germany, Great Britain, Italy, the Netherlands, Belgium, Denmark, Ireland and Luxembourg .
Unit of Analysis:	individuals

Appendix 5: DDI 3 Core elements (based upon DDI 2 Lite)

		DDI 3.0 Core
Module	Sub-Module	Tag
Study Unit		Identifying_Agency_NCName
Study Unit		Version Number
Study Unit		Version Date (ISO format)
Study Unit		Name of Agency or Individual responsible for this version of the Study Unit
Study Unit		Title of Study
Study Unit		Name of Creator
Study Unit		Name of Publisher
Study Unit		Name of Contributor
Study Unit		Publication Date (ISO format)
Study Unit		Abstract of study [allows xhtml structure]
Study Unit		Universe Statement
Study Unit		Name of Funding Agency
Study Unit		Abbreviation of Funding Agency
Study Unit		Topic Keyword
Study Unit		Description of Geographic Coverage [xhtml structure allowed]
Study Unit		Name of Geographic Level
Study Unit		Date or time period covered by the data (ISO format)
Study Unit		Analysis Unit of the Study
Study Unit		Title of Other Material
Study Unit	Conceptual Component	Concept Description
Study Unit	Data Collection	Description of time method for data collection [xhtml structure allowed]
Study Unit	Data Collection	Description of Sampling Procedure [xhtml structure allowed]
Study Unit	Data Collection	Data_Collector_NCName
Study Unit	Data Collection	Description of data source
Study Unit	Data Collection	Collection Start Date
Study Unit	Data Collection	Collection End Date
Study Unit	Data Collection	Mode of Collection description [xhtml structure allowed]
Study Unit	Data Collection	Literal text of question
Study Unit	Data Collection	Description of Cleaning Operations [xhtml structure allowed]
Study Unit	Data Collection	Organization_Responsible_for_CleaningOperations_NCName
Study Unit	Data Collection	Description of Weighting [xhtml structure allowed]
Study Unit	Data Collection	Response rate
Study Unit	Data Collection	Description of Sampling Error
Study Unit	Logical Product	Category label
Study Unit	Logical Product	Code value associated with category
Study Unit	Logical Product	Name of Variable [mnemonic]
Study Unit	Logical Product	Label of Variable
Study Unit	Logical Product	Fuller description of Variable [xhtml structure allowed]
Study Unit	Logical Product	Variable Universe
Study Unit	Logical Product	Description of derivation
Study Unit	Logical Product	Valid Range
Study Unit	Logical Product	Label of Variable Group
Study Unit	Physical Data Product	Gross Record Structure - Records per case
Study Unit	Physical Data Product	Variable quantity
Study Unit	Physical Data	Character set