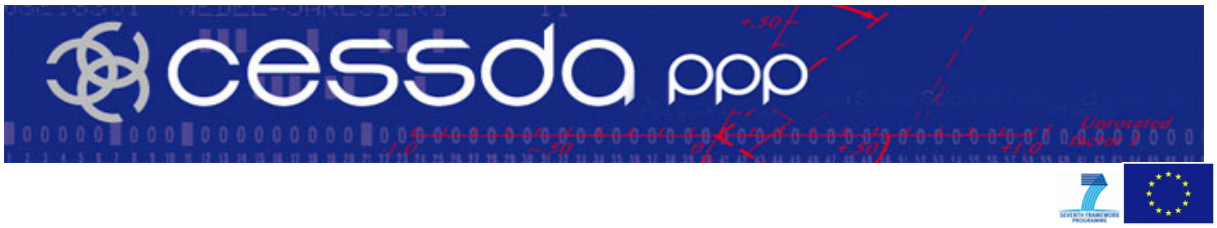

Title	**Evaluation of DDI 3 - Preservation metadata (D8.1)**
Work Package	WP8 Enhancement of data and metadata infrastructures for the CESSDA RI
Authors	Uwe Jensen (GESIS) – DDI 3 evaluation Hervé L'Hours (UKDA) – Preservation metadata
Source	• DDI 3 documentation & presentations • OAIS; METS, PREMIS
Dissemination Level	PU (Public)

Summary/abstract

Evaluation of DDI 3

This document discusses the scenarios that will evaluate the DDI standard version 3 established to document comparative social science surveys.

The bottom up approach describes four use case scenarios used for the testing of DDI 3. The example from the ISSP cross-national study research program provides the material for usage with the evaluation of the first use case to assess involved modules of the DDI standard related to the space dimension. Further bottom up uses cases are outlined for the most complex dimension space x time.

The top down approach introduced in this version covers use cases starting with the grouping approach.

Some critical aspects on use of the top level modules group / schemes and / or comparison / resource package require further examination and discussion in the new CESSDA organisation and within the DDI Alliance itself.

Definition and use of Preservation Metadata

This document addresses the definition and application of preservation metadata to the CESSDA Research Infrastructure (RI).

The distributed architecture, object model, persistent identification and versioning of the study object being preserved must all be taken into account as part of a 'shared metadata model' and, though related, are assumed to be beyond the scope of this document

Table of contents

1 The Data Documentation Initiative (DDI) and the new Version DDI 3 1
2 DDI 3 evaluation procedure 3
2.1 Bottom up approach in evaluating DDI 3 3
2.2 Top down approach in evaluating DDI 3 top level modules 4
2.3 Key findings from DDI 3 evaluation 5
2.4 Summary of important DDI 3 innovations 8
3 Core recommendation and implications 9
3.1 Technical implications 9
3.1.1 DDI 3 compatible base technical infrastructure 9
3.1.2 DDI 3 compatible tool and service developments 10
3.2 Specifications and developments of technical metadata standards 11
3.2.1 DDI 3 related issues 11
3.2.2 Preservation metadata - OAIS - Quality standards & long term preservation 11
3.3 Organisational implications 12
4 Definition and use of Preservation Metadata 14
4.1 Introduction 14
4.2 Core conclusions 15
4.3 cessda-ERIC Implications & Recommendations 16
4.3.1 Organisation 17
4.3.2 Portal 17
4.3.3 Registry 17
5 References 19
5.1 DDI 3 evaluation 19
5.2 Preservation metadata 19

1 The Data Documentation Initiative (DDI) and the new Version DDI 3

The Data Documentation Initiative (DDI) is an effort to create an international standard for describing social science data. The DDI project was initiated by ICPSR in 1995 to develop a metadata specification to replace the obsolete OSIRIS standard. The development of a first DDI version was published March 2000. A first external evaluation of the DDI took place in 2001 and the four evaluators summarized their core findings (beyond others) as follows[1]:

> "The four evaluators were in agreement that the DDI is a worthwhile scientific effort and that it fills an urgent need for standardization of social science technical documentation and interoperability. One termed it "a strategic component of the infrastructure necessary to support the exchange of structured social research survey data. … In terms of the substance of the emerging standard, the evaluators concurred that the DDI does an "excellent job for survey research," but each noted that more work is necessary to extend the standard to more complex types of data – most notably, time series and aggregate/tabular data."

With the formation of the DDI Alliance in 2003 it became a membership-based organisation that allows for improved sustainable development among the international network of data archives. The Alliance published DDI version 2.0 in 2003 and version 2.1 in 2005.

Based on a three year community based effort the new version DDI 3 responds to core needs of comparative research in the social sciences as well as to providing with the modular structure a basis for innovations in tools and functions which support the whole study and data life cycle. Technically it applies the advantages of an XML Schema format compared to the former XML DTD format and thus eased machine actions and software programming in many ways.

DDI 3.0 was officially published at April 28 2008. The present version DDI 3.1, published on October 18, 2009, reflects resolution of the final URN structure to ensure persistent URNs for all identified elements and the correction of bugs.

Applying the new DDI facilities will provide new options for efficient metadata production and management and enhances the potential for exchange and re-use of metadata between data producers and data providers along the whole lifecycle of studies and research data.

However to realise DDI 3 compliant workflows and seamless production, exchange, publication and finally public retrieval and access by European social science and humanities (SSH) researchers still require adequate development of modular editing and managing tools for daily practice to provide and interact with a DDI compliant and interoperable technical infrastructure within a future cessda-ERIC.

The challenge is to roll-out and to implement a DDI 3 compliant infrastructure at large in step-wise fashion to transform the new potentials to advanced and new data products for the scientific community.

This report contributes to a general endeavour of several work packages in the CESSDA-PPP to plan the future technical developments and software upgrades of the CESSDA RI. The

[1] Daniel Greenstein, et al (2001): Results of the Evaluation of the Data Documentation Initiative (DDI) http://www.ddialliance.org/sites/default/files/evalsummary.pdf

different technical and organisational challenges including designing a general infrastructure, the tool developments for particular purposes (like the portal facilities, harmonisation platform, question database and a thesaurus management system) and the related specifications of the future use of DDI 3 and further standards like those for preservation provide the basis of an integrated technical roadmap.

The particular contribution of this report summarizes first the work on the “Evaluation of DDI 3” (based on internal work paper DR. 8.2.2) with particular focus on handling comparative surveys by applying core DDI3 modules. Secondly it reports in summary on the “Definition and use of Preservation metadata” (internal work paper DR 8.1.1).

2 DDI 3 evaluation procedure

The work undertaken in this field was to answer questions on the capacity of DDI 3 to support life-cycle processes of complex datasets, including preservation.

Representation of questions from the questionnaire, their relation to variables, comparative data representation, dataset management, versioning, and preservation metadata had been regarded as key issues to specify.

The work methods are basically grounded on

- ISSP Module on Role of Governments (RoG) as the general test case
- A bottom up approach to evaluate DDI 3 in relation to the documentation of Question / Variable using the Finnish RoG 2002 case as an example on country level
- A top down approach to evaluate first the management of the whole module with four study instances with different numbers of participating countries and second to extend this approach to manage survey series in total, like the ISSP with ten modules, over 500 samples from 44 different countries.

The results made available for the top down approach are based on evaluating 'concepts at papers & presentations' from public sources as provided by the DDI Alliance, core presentations at IASSIST conferences, and other reports from the public domain. Major work undertaken in that direction was to:

- To become familiar with the new standard on the base of the documentation of DDI 3.1
- To explore and integrate materials from presentations providing information to enhanced understanding of the major concepts and mechanism.
- Evaluate particular use cases to integrate practical solutions with underlying concepts.

Beyond in some cases very specific questions on certain elements (see respective work paper DR. 8.2.2) the Report on DDI 3 focuses on general principles, modules and mechanism in understanding the innovations made in DDI 3 to support the roll-out and future implementations of DDI 3.

2.1 Bottom up approach in evaluating DDI 3

The starting point of the documentation of the Finnish ISSP data 2006 (FSD2248) was to form a DDI 3_0 instance considered as part of a group named "All Finnish ISSP studies". The following tools were used for the documentation:

- CESSDA core template from the DDI Alliance website;
- Extensive manual additions using the Oxygen XML editor.

The modules and sub entities (listed only for higher levels) applied in the tagging work are provided with the resulting XML file:

- FSD2248_ddi3_based_on_cessda_core_instance_DRAFT20090918.xml.

This file comprises major content placed at the modules:

- DDI 3_0 instance
    - Group module (Finnish ISSP as a whole)

- Study module for FSD2248 (Finnish role of government 2006)
  - Data collection (Methodology & Instrument)
  - Logical data product (Variables)
  - Physical instances (Data file)
- Archive module with organisational information

The outcomes were basically summarized in the work paper DR 8.2.2 "Evaluation of DDI 3" under the header "Use case 1 Bottom up: Finnish case of RoG 2006". It formed the base for intensive discussions and in-depth work with DDI experts during a workshop organised by the DDI Alliance (Dagstuhl 2-5.11.2009). Further conclusions also in relation to the needs of future use of DDI 3 at FSD (Finnish Social Science Data Archive) were presented at the first annual EDDI in Bonn (4.12.2009). Final results will be published at the beginning of 2010 under the title “DDI 3: An Archive’s Perspective to DDI 3” at the DDI Working Paper Series.

2.2 Top down approach in evaluating DDI 3 top level modules

With respect to the first group of innovations (see above) the question is how to make use of these improvements and in which aspects of the lifecycle or more precisely: How to use them specifically, for which processes, at key phases during the lifecycle.

One of the major questions of the top down approach regards the role and the strategy on how and why to use the various options.

- Major top level module (and respective sub modules) group, resources package, and comparison to represent, manage and version comparative studies, related questionnaire content, and variables and data files.

As an intermediate step a systematic structure of four scenarios were deployed describing basic principles, options and rules to organise the complexity of comparative surveys:

- Scenario 1: Group and Comparison
- Scenario 2: Resource Package and Comparison
- Scenario 3: Combination of case 1 and 2
- Scenario 4: Standalone use of Resource Package or Group

The final expected outcome from the combinations of modules is to provide recommendations and guidance on the strategy in applying these core modules along lifecycle processes to achieve and even more to improve the daily work processes for data providers and beyond.

Of particular interest are their basic function and efficiency in documenting metadata from complex repeated international surveys for:

- A country, being part of a single survey instance (like a topical module or a wave) and the related country question / variable metadata in the common project language as well in the native language(s);
- Integrating the single countries to a common survey instance with common and deviating metadata for the whole context of that instance, the respective integrated questions and variables and the relationships among many countries, each serving one or more national languages;
- Representing several survey instances (like replications of a topical module with basically one common documentation language) as physically integrated version or as

(virtually integrated) set of country instances, which are physically independent but referenced with each other;
- Cumulation of survey instances (like replications of a topical module) integrating all those country data sets (representing one sample / measurement point in time) according to the respective survey programme regulation. The ISSP cumulation of four instances Role of Government integrated those countries that participated at least twice;
- Organising a survey series as a whole like ISSP with ten different, independent modules or other large research programmes with complex topical issues organised like time series, cumulations of single trend questions or other topical and methodological designs;
- Finally and overall to consider the optimized design of DDI modules to allow agencies like data archives and other agents economic and efficient re-use of available metadata in their internal processes as well as between them along a common workflow.

The scenarios and further questions are basically described along with four versions of the work paper DR 8.2.2 "Evaluation of DDI 3". Intermediate conclusions and outcomes were presented first time at the work package leader (WPL) workshop in Essex, August 2009. It formed the base for further intensive discussions and in-depth work with DDI experts during a workshop organised by the DDI Alliance (Dagstuhl 2-5.11.2009).

Further work was undertaken focusing on the re-use of metadata in internal archival workflows. First conclusions on principles in designing and applying the modules Group and Resource Package in practice to organise metadata from comparative survey series were presented at the first annual EDDI in Bonn (4.12.2009). Final results will be published at the beginning of 2010 under the title “GROUPING OF DATA SERIES USING DDI” at the DDI Working Paper Series.

2.3 Key findings from DDI 3 evaluation

The outcomes of the practical and conceptual evaluation of DDI 3 undertaken in WP8 can be summarised as follows according to the tasks 8.2., where a working group:

- Cooperated intensively with the DDI people at ICPSR who provided their expertise to solve questions arising during the course of the PPP such as Persistent Identifiers. Conceptual issues occurring during the particular work in WP8 on DDI 3 could sufficiently be solved due to extended contacts to the DDI TIC and in particular by the DDI expert workshop in Dagstuhl, Germany November 2009;
- Identified key issues regarding metadata representation of comparative data on study, question and variable level. The work on top level modules in particular Group and Resources packages provides comprehensive and quite positive conceptual insight on how to organise and to manage repeated international comparative surveys.

Of particular importance for future metadata work and documenting processes is the result that these modules provide several options to re-use or to share already existing metadata in the data producer’s and the data provider’s workflows applying a combined set of both Group (inclusion by inheritance) and Resource package (inclusion by reference).

The metadata transfer between different instances or agents allows the re-use of locally captured metadata (like for field data, translation of the instrument, data collection events on a national level) at the agency that is in charge of integrating the country specific data and

metadata. Thus international research projects organised by such a typical business case could principally benefit in applying respective DDI 3 modules.

Furthermore the combined Group/Resource package approach supports the cumulative integration of successive waves. This particular workflow process still needs more investigation.

Both modules provide the capacity for adequate representation of study context information (study level metadata), the questionnaire (question wording and flows; national languages), the logical and physical representation of fielded data (variables; data files), and related documents including versioning.

However it must be considered that the work on Group and Resource Package express so far a static top down view on an existing complex survey series 'as it is'. Further conceptual modelling has more explicitly to apply the rules and requirements of the metadata processes in (present and future) daily work flows according to the technical and organisational impacts. Additionally the Comparison module needs to be integrated to achieve a fully fledged ensemble to document international comparative surveys under the aspect of what is comparable and where deviations occur (with more precision compared to the group approach).

- The use of single items of particular DDI 3 modules has been examined in more detail for a cross-section study (Finnish ISSP 2006) to perform a critical assessment of the DDI singular items for the base module Group and the subordinated modules Study Unit, Data collection (Methodology, Instrument), Logical Product (Variables) and Physical Product (data file) and Archive (organisational information and regulations like data access).

As a general result all necessary information (metadata) could be expressed within the several DDI schemas so far. The particular use case demonstrated with DDI conformant XML files the respective success of that endeavour.

- Considering the work from the top down and the bottom up approach in summary, the results on a conceptual and practical work level provided already a broad insight to the potential and capacity of DDI 3. Its capacity was demonstrated to support the major life-cycles processes until the phase to prepare studies and data for publication;
- Further results and considerations are provided with portal specific outcomes (WP5) and related issues in WP12. Particular results on DDI 3 and preservation metadata are provided in chapter 4 of this report;
- Critical remarks.

In addition to quite positive results critical aspects of the evaluation regard the following:

There is manifestly a large gap between the potentials of the new DDI 3 standard on the one side and the existence of adequate tools to capitalize these advantages on the other side in testing and all the more for applying them to daily work routines at large.

Considering the available dedicated tools for the testing of a cross-sectional study such as one ISSP country instance requires still some tagging by hand (this regards also other test case and becomes more and more critical if the complexity grows according to a complete cases (from study level to question / variable documentation) like a single Eurobarometer trend or even a

subset of the trend file at all). Although this discrepancy is significant at present there are on the other side arising prototypes leading towards solutions in tools development like the DDI editor with a layered framework presented at the first EDDI in Bonn. Related to available tool developments and in compliance with recommendation on further data and metadata extensions WP8 propose the development of an accompanying modular DDI 3 editor (see Report D8.1; D8.3) that provides the basic functions to meet the practical needs of comparative survey documentation and the respective functions to manage such study collections in total.

The cooperation with DDI experts was sufficiently extended during the CESSDA-PPP in reviewing DDI 3 under several perspectives. However it will be essential for the future cessda-ERIC to ground further DDI developments on a sustainable cooperation with DDI experts (in terms of time and resources) to allow significant progress in the roll-out of DDI in a dedicated timeframe that allows step-wise implementation of the technical infrastructure and the adequate metadata production and management software both required to provide finally high quality data and metadata from cessda-ERIC members and cooperating research projects to SSH community.

This perspective must additionally consider that advanced metadata needs from research domains and new technical challenges require continuous adjustment and ongoing development of this standard as well as standards related to further domains and growing relevance (Preservation, Aggregate Data, extended context like Geospatial metadata, persistent citations of data, advanced methodological information, concept metadata etc.).

Compared with diverse daily practices, workflows and organisational regulations on archival data management at present among the single archives (ingest processes, data processing needs, long-term preservation; added value data production for complex multi-lingual comparative surveys) the new cessda-ERIC infrastructure requires harmonised standard procedures and a common data model to achieve and sustain interoperability for DDI based data products and beyond.

2.4 Summary of important DDI 3 innovations

Beyond some more details to explore during the roll-out of DDI 3 the new version solves some substantial open needs from the past for the documentation of complex data from social science research. Core advantages and important innovations are summarised in the following section.

1. **General documentation needs on metadata from complex survey series**
   - Adequate documentation and management of comparative study instances on spatial and temporal dimensions by modular structures and referencing mechanism
   - Full documentation of the survey measurement instrument as separate entity
   - Managing question flows (conditions; loops: machine actionable)
   - Grouping of study series to manage comparative data collections at large
   - Mechanism for identification, versioning and persistent citations of data
2. **Specific needs on harmonisation, comparison, multilingualism and concepts**
   - Extended support for general variable comparison applying the Comparison module
   - Support of harmonization projects to create harmonized data and applied concepts
   - Enhanced support to document surveys with multiple languages
3. **Support of extended scope of Data structures and Data types**
   - Support for qualitative and quantitative data

- In-line inclusion of micro data (information about individuals) and aggregate data (information that has been aggregated to spatial level like nations or regions)
- Support for tabular, spreadsheet-type, representation of aggregate data
- Aggregate data transport (inline inclusion of cell content with data item description)
- ISO 11179 compliant data registries such as question, variable, and concept banks

4. **Efficient lifecycle management and economic study and data documentation**
   - Metadata workflows (capture, exchange, enhance, manage) at and among different agencies and agents are completely supported from research project planning, data production and added-value knowledge products to dissemination and analysis of studies with high-quality research data
   - Extension to more metadata from the study life-cycle like concepts, instruments and data-collection, versioning and persistent identification.
   - Facilities to eliminate redundancies by core modules and referenced (scheme based) elements (group; resource packages for question and variable documentation; DDI profile) at each agency and related life-cycle events
   - Creation of repositories like question data banks for public re-use
5. **Archival management issues and alignment to further standards**
   - Improved capture of archival information for data organization and management
   - DDI profiles for specific use at an individual service provider instance
   - Alignment to OAIS (Open Archival Information System) and related standards PREMIS and METS.
   - Expanded alignment with other metadata standards like Dublin Core, MARC, ISO11179 (metadata registries standard), SDMX data exchange, and Geospatial Metadata Standards such as FGDC (Federal Geographic Data Committee) and ISO 19115 (Geographic Information Metadata)
6. **Support in programming and to manage manifold relationship**
   - Use of XML schema improves machine driven actions and respective programming
   - Compatible with Computer assisted software
   - Modular design and referencing mechanism allows to host structures from complex survey series and to manage relationship among and between data collections, variables and questions and its respective versions.
7. **Publication of metadata from comparative or comparable data for public re-use**
   - Grouping model and the comparative module improve public discovery and re-use of comparative and comparable data by the scientific community for comparative research
   - Improved support to create registries like question data banks enhance public re-use in several fields like to support new research or questionnaire developments
   - Modular design and referencing mechanism allows relationship and thus cross-referencing and browsing among collections, variables and versions and across surveys and collections
   - Versioning enhance up-to date information for proper data analysis, data citation and controlled replications of survey results

3 Core recommendation and implications

The WPL workshop in Essex August, 2009 commonly agreed on three core recommendations:

1. DDI 3 is as mandatory target the future CESSDA documentation standard considering strong needs in further DDI evaluation in cooperation with DDI experts and DDI TIC;
2. Future tools development must be compatible with DDI 3;
3. Plan transformation process from DDI 2 to DDI 3 supported by incremental technical developments.

CESSDA impacts regard three central areas to be considered along with the development of an underlying strategy and business plan by the CESSDA management.

3.1 Technical implications

This area regards three different levels of the distributed infrastructure:

3.1.1 DDI 3 compatible base technical infrastructure

To set-up the general technical infrastructure a Requirement, Design and Implementation plan is required. The outline of such a distributed architecture (repository / registry approach) in the tender report "Technical Specifications for a European Question Data Bank"[2] provides a considerable blue-print to define detailed functional requirements and technical specifications for a larger and DDI 3 compliant CESSDA technical infrastructure system.

> "The proposed architecture expands beyond the scope of the QBD and addresses broader issues related to the federated CESSDA environment, the technology diversity, variations in metadata structures and ownership. It was necessary to solve some of these in order to design a solution for the QBD.
> …we propose for this backend metadata storage to be driven by a web services based architecture composed of a registry and multiple repositories. This provides support for future CESSDA applications as well and allows for metadata currently not directly needed by the 3CDB and QBD to be stored in the system. The exchange of information between the different systems will take place using web services and based on a on a common XML metadata model." (p. 12 Architecture, Overview)
> "…using XML, we need to create a model that can independently represent these objects with the ability to maintain their referential integrity and ensure their unique identification. The above list is remarkably, but not surprisingly, very close to the model of the latest version 3 of the DDI specification. This should therefore be used as a base for the CESSDA model which would then at the same time ensure compliance with the DDI (an optional requirement of the QDB)." (p. 13 Architecture, Metadata Model)

An internal evaluation paper[3] of the tender report in work package 9 says:

---

[2] Technical Specifications for a European Question Data Bank. Final Version May 2009
Arofan Gregory, Pascal Heus, Chris Nelson (Metadata Technology) and Jostein Ryssevik (Ideas2evidence)
http://www.cessda.org/project/doc/CESSDA_PPP_QDB_May09.pdf
[3] WP9 QDB Tender Evaluation, Maarten Hoogerwerf, 12.08.2009 Draft

> "The design-principles as used in DDI3 fulfils these requirements, making DDI3 a logical choice to implement or to use as the basis for an agreed metadata model.
> Web services are a logical choice, and its lightweight variant REST[4] might be a practical alternative. The question is whether this doesn't lack valuable features for scalability, security, etc." (p. 4 Background and Overview)

Both DDI 3 compliance and web-based Service Oriented Architecture appear finally necessary to run (particular) services like the harmonisation platform or the question databank at the portal.

The whole process towards step-wise practical DDI 3 implementations will be a transformation effort over many years which recommends phased implementation plans. This has to consider priorities in terms of (new or updated technical) services at the CESSDA portal and the accompanying local workflows and tools to submit the data products for public access, resource implications and technical constraints. Therefore the Metadata Tech-model requires in-depth work (considerably in cooperation with the authors) to:

- Focus functional and technical requirements, constraints and options related to data documentation, management, and dissemination needs;
- Estimate required respective resource and time frames in relation to setup a list of priorities in a phased development approach;
- Formulate overall standards, technical developments, and resources and maintenance issues to support the CESSDA management in developing the future cessda-ERIC.

Such core developments and implementations must be efficiently planned, organised and managed within the cessda-ERIC. General recommendations in this context were formulated in the report **D8.4 "Models for future organisation of technical R&D developments and training for the metadata technicians".**

3.1.2 DDI 3 compatible tool and service developments

To serve content and services in capitalizing on the new DDI features and options requires the development of specific software and functions for internal archival purposes and / or public use. The following work packages submit therefore particular recommendations, specifications and plans on four central applications:

- Thesaurus management system (WP4);
- Portal functions on publishing & browse, search, access to data and metadata (WP5);
- Technical solution for a Persistent Identifier (PID) and Versioning system (WP12);
- Online Harmonisation platform and Question database (WP9).

Modular DDI 3 compatible editor

WP8 recommended the development of a modular DDI 3 compliant editor to make use of:

"the advanced DDI 3 features and to bridge the gap in missing a metadata production tool for daily data documentation work ... The proposal aims to develop and implement as soon as

[4] Representational State Transfer (REST), http://en.wikipedia.org/wiki/Representational_State_Transfer

possible a fully functioning editor to work with DDI 3 in the daily production and for the dedicated purpose as specified." (D8.1 p.8.).

The Report **D8.1 "Data and Metadata Extensions of the CESSDA RI"** (chap. 3.1;.p. 7) formulates a set of functional needs for advanced DDI 3 compatible metadata processing and outlines considerable modules to serve these needs (chp.3.2).

3.2 Specifications and developments of technical metadata standards

This aspect regards the needs on presently applied standards (in particular DDI 2) and issues on further standards to apply in the future, and which are supported by DDI 3 as well.

3.2.1 DDI 3 related issues

As it regards DDI 3 the technical work packages opted for further DDI evaluation under conceptual and technical aspects (like effects on programming issues) in preparing a roadmap to roll-out, implement, and support and maintain DDI 3.

As an integrative part in moving towards a business plan for the technical development an end-to-end evaluation of DDI3 issues with DDI expert was recommended to:

- Analyse the future pay-off to practically apply DDI 3 and / or DDI 2 for cross-sectional to complex data and metadata holdings. Aim is to provide conclusions and constraints to come to detailed recommendations on the future organisational strategy and impacts on technical production line in applying DDI3 and / or DDI 2 to daily work procedures and workflows within a distributed CESSDA infrastructure. This includes comparison of estimated resource requirements to potential benefits of migrating the particular types of studies.
- To develop a migration strategy for advanced use of DDI 3 in documenting, publishing and public re-use of metadata from complex social science surveys (and beyond) and to support new or extended types of metadata beyond traditional social science domains.
- Of importance is furthermore to apply mappings between DDI 3 and other standards like Dublin core, SDMX (Statistical Data and Metadata eXchange) or Geospatial Metadata Standards.

It is an ongoing requirement to maintain and develop implemented standards over time to keep compliance with new metadata needs by extending the scope, range or type of social science data. As such, it necessitates sustainable resources both for daily support and proactive actions to improve or extend the scope and the functionality of standards and tools. Further strategic implications are provided with the Report **D8.3 "Funding models for future development of metadata standards and software tools".**

3.2.2 Preservation metadata - OAIS - Quality standards & long term preservation

As it regards standards for Preservation Metadata the WP8 work paper DR. 8.2.1 "Definition and use of preservation Metadata" provides an overview on OAIS the ISO reference models for an Open Archival Information System and references the respective technical standards PREMIS and METS as well as initiatives working on quality standards regarding long-term preservation.

Beside the roll-out of DDI 3 the cessda-ERIC will apply several standards to serve and certify particular activities of the infrastructure. In concluding that a common CESSDA data model is therefore of central importance DR 8.2.1 underlines furthermore to

> "consider a movement towards OAIS conformant terminology to be an important step in improving data management and preservation planning. Agreement on appropriate terminology within the study lifecycle and for the study object model will facilitate some of the planning challenges that face CESSDA in a multi-national and multilingual goals environment."

The outcomes and conclusion on preservation metadata are provided in chapter 4 in this report.

3.3 Organisational implications

The roll-out of the new DDI 3 standard and already recognised requirement for technical developments also implies changes and adjustments on the organisational level of the cessda-ERIC and its' members. However final decisions and overall regulations to implement the cessda-ERIC infrastructure are still under development. This concerns as well the need to integrate the work package specific implications on the use of DDI 3 to one strategic umbrella. However the following aspects on particular organisational issues are considerable:

The diversity of present local regulations on standards and systems is significant and to achieve the CESSDA goals requires extended harmonized practise on the

- Applied standards for technical and substantial documentation;
- Local workflows and the publishing regimes.

Such key challenges are closely related to the quality standards and best practise governed by the cessda-ERIC as well as the requirements of the CESSDA portal. Core aspects of the present practices which are central for future transformations are issues to solve along with the roll-out and implementation process of DDI 3 at both the members and the total infrastructure of the cessda-ERIC:

- What standards are applied to all studies? How look the particular standards within series?
- What is to document and to what detail (study, question, instrument, and variable - data processing: formal standards, substantial harmonisation - data file production: simple files, integrated file, cumulation, time series - Versioning - Publication (Zitation rules)?
- Where is it documented and to what granularity applying what DDI elements (1/2.0/2.1) or other technical standards?

In fact the present DDI practises of the archives needs to be compliant with the strategies of the future cessda-ERIC. Central concern is to bridge the gab of the present use of DDI 2 and the roll-out of DDI 3 to transform its potentials into future benefits for the daily practices along the whole life-cycle. Following core issues are central for a successful transition roadmap:

- For what scope of studies and new data types is it necessary to apply DDI 3 and for what scope of studies is it sensible to retain them in DDI 2?

- What are the organisational and technical impacts to serve both standards in parallel for a certain period of time?
- What are the organisational and technical impacts, the required resources (cessda-ERIC members, the hub, external tenders) and time horizon to migrate and transform (selected scope of) studies or complex survey collections to DDI 3.
- How can a phased roadmap look in detail considering the requirements to bridge the gap of current practices with DDI 2 and expected options in applying DDI 3 in the future?

Implementation of new technical standards implies updated standards of operational processes, new tools and adequate detailed procedures. As an important accompanying mean for these organisational transformations qualified guidance, advice and training is to provide to the members of the cessda-ERIC and to the particular needs of the daily operators.

As this is part of the general strategic planning “Training activities” WP6 and WP7 provided respective recommendation on related programmes and a Virtual Centre of Competence providing an information and management platform. In reviewing these proposals the WPL coordination workshop in Essex (August 2009) agreed on the importance of implementing interacting expert groups, with one having particular responsibility for standards-related issues.

4 Definition and use of Preservation Metadata

This chapter summarize major aspects in the definition of and the application of preservation metadata to the CESSDA RI as addressed with the work paper DR 8.2.1.

4.1 Introduction

The document addresses the scope and overall element set required for preservation management in two key areas of the CESSDA environment:

- Harmonised practice across CESSDA member archives;
- Requirements for the CESSDA Portal.

In this regard it examines the possibilities for the collection and management of preservation metadata in a variety of forms including DDI 2.1, DDI 3.0, PREMIS and METS.
Some assumptions have had to be made including that:

- A distributed web services architecture will be required;
- Local systems and practice will not align at the same rate;
- A common interchange format with an agreed element set will be required;
- That any metadata standard agreed will be capable of sufficiently granular versioning to support ongoing preservation events.

The conceptual problems are less about the requirements for and the collection and management of preservation metadata, the complexity arises more in the area of the object model and versioning issues encountered with a DDI3 model implemented in a distributed architecture.

Some of the key challenges in implementation are applicable far beyond preservation metadata and fall within the scope of maintaining all relevant metadata (e.g. coordinating updates) across a distributed architecture. Any final recommendations for the implementation of preservation metadata must align with a ‘shared metadata model’ ([1])

A clear minimal requirement for preservation metadata and an ongoing increase in the application of preservation metadata standards throughout the lifetime of CESSDA will be vital in creating and supporting the evolving technical infrastructure.

A number of goals for the CESSDA model and DDI 3.0 evaluation work package, including question bank administration and complex comparisons of datasets are not addressed within this document. The assumption is that the preservation metadata approach and elements described here should be sufficient to support any technical implementation including a distributed solution. The primary dependencies for successful preservation metadata management are:

- an agreed object model encompassing all relevant studies and their content;
- a shared metadata model across the CESSDA infrastructure;
- documentation of which elements must be harvested from a contributing archive;
- what constitutes sufficient change to warrant a new harvest;
- what is the minimal level of description that must accompany an amended study;
- what, if any, elements must be re-ingested into the contributing archive from the distributed CESSDA architecture.

4.2 Core conclusions

Harmonisation of practice, metadata collection and interoperability of machine-actionable metadata must take place within a framework which is aligned with the OAIS Reference Model.

The approach to preservation and to preservation metadata collection and management must start from the OAIS perspective and continue with more specific preservation metadata implementation strategies designed to align with the OAIS Reference Model.

Common understanding, interpretation and terminology (or terminological mappings) are pre-requisites to common practice and eventual implementation of, or integration with, common technologies.

The existing disparity of technical archival solutions indicates that harmonisation will be best achieved through a phased approach.

Any phased approach is likely to necessitate different archives adopting agreed common practices and related infrastructure at varying speeds dependent on existing resources and legacy implications; with this in mind the initial requirements should be carefully set at the minimal practical level.

A roadmap from this initial minimum to greater harmonisation of practice and eventual true interoperability must be flexible and responsive to changes in the wider archival approach to preservation.

The PREMIS Preservation Metadata Implementation Strategies standard provides a stable and accountable basis for preservation metadata planning but is not sufficient for all preservation and metadata administration purposes. The PREMIS standard explicitly states that detailed implementation in some areas, including format registries, technical/administrative metadata extraction and creation are beyond the scope of the standard.

In line with the related goals of ‘harmonisation’ and ‘interoperability’ it is perfectly possible for an archive to be PREMIS compliant by ‘capturing’ and ‘knowing’ the information required by PREMIS while not being capable of exporting a valid PREMIS XML file for its Archived objects, or even being able to provide the relevant information in a ‘machine actionable’ form.

The ongoing mapping of DDI3.0 to PREMIS elements will form an important part of the overall ‘metadata model’ required by CESSDA technical infrastructure but initial requirements for preservation metadata delivery from local archives to the portal may be very limited. The metadata model must also take account of those areas beyond the scope of PREMIS.

Any metadata model for CESSDA must take account of both DDI 3.0 and PREMIS preservation metadata elements until a full evaluation has been made and best practice agreed. Both a top down (broad object modelling) and a bottom up (agreed best practice on the representation of studies in DDI3.0) approach will be required for successful harmonisation of preservation metadata in later phases of CESSDA cooperation.

Initial requirements may be limited to agreement on which aspects of change to a locally held study are deemed sufficiently significant to trigger the creation of a new Dissemination Information Package (DIP) for harvest by the CESSDA infrastructure. Of the numerous 'events' surrounding a study during the archival phase of its lifecycle many will only be of relevance to the maintaining archive; the initial requirement for preservation metadata exchange should focus on amendments to dissemination information packages harvested by the CESSDA portal. Increased harmonisation of preservation metadata practices is desirable and any increase in the provision of machine-actionable preservation metadata strengthens the trust relationship between CESSDA and its member archives; it also simplifies oversight of compliance with CESSDA membership requirements.

The primary goal of initial preservation metadata sharing must be to allow end users of the portal to identify changes to studies over time at an appropriately granular level.

Necessary dependencies for final decisions on appropriate granularity for preservation metadata exchange include object models, DDI 3.0 best practice and Persistent Identifiers, all designed within a clear framework for the desired distributed implementation model ([0])

4.3 cessda-ERIC Implications & Recommendations

Preservation metadata capture, management and exchange must form a key part of the agreed CESSDA metadata model. The general approach may be aligned with the recommendations of this document but the particulars of implementation will depend on the final technical infrastructure agreed. If a distributed modular architecture is agreed for CESSDA each participating 'node' (whether an 'archive' or a 'bank') will need to evaluate the requirements for capturing, managing and exchanging any metadata they are responsible for.

CESSDA implications may be broadly divided into two areas. Those which impact the CESSDA organisation:

- Standards for membership;
- Best practice for members;
- Alignment of practice.

Those which impact the Portal:

- Agreed object model;
- Minimal metadata set for harvesting by the portal;
- Tools for exporting in appropriate formats;
- The resubmission and ingest of new or amended study information from other nodes.

Timescales may be divided into:

- Standards which must be in place for the initial creation of the infrastructure and workflows/dataflows;
- Agreed flexible phases for:
  - Adoption of improved local practice;
  - Greater integration with the technical infrastructure.

The issue of one or more central registries providing a reference point for multiple archives is relevant at both organisation and portal level.

4.3.1 Organisation

There is no a priori requirement for local archives to capture all metadata relating to every preservation action. It is recommended that any 'events' which occur in a node are recorded in as much detail as possible. Such a generic approach to events occurring to objects will allow us to expand the granularity over time and respond to changes in best practice for recording preservation-specific events.

This issue would benefit from simple event tools capable of recording the completion of common OAIS phases and exporting details of those event (with associated timestamps, agents and rights information) to PREMIS format.

Such a record of events over time would facilitate review and management of:

- Disparate Practices;
- OAIS alignment;
- Trust relationships between archives and end users;
- Local approaches, extensions, requirements.

Any eventual harmonisation must consider that common understanding, terminology and interpretations are a pre-requisite for common practice and common technology. Any tool would need to take into account the object and versioning model agreed. Essentially the recommendation has got to be a supported phased approach.

4.3.2 Portal

Note that 'exchange' is a two-way process with an implication that an exchange format must be sufficient to permit controlled harvesting from participating archives (a specialist Dissemination Information Package or DIP) and to act as a form of Submission Information Package (SIP) for later ingest (as new or revised studies) back into the archives.

Discussions to date do not indicate that there is a requirement to exchange Preservation Metadata for each action taken by the Archive.

For harvest by the portal we require a DDI 3.0 compliant format with a clear minimal element set and possibly more advanced options to take account of the functionality of various 'banks' as they come online. A number of tools may be required to transform metadata from disparate legacy formats.

From the preservation metadata perspective the first phase of interoperability only requires that any changes between harvests are recorded and harvested in a suitably granular manner. It is a reasonable assumption that CESSDA will increase the level of granularity required over time.

4.3.3 Registry

The concept of a format registry may usefully be expanded to encompass other standard reference points for multiple archives. Many of these would extend logically from file format definitions though some apply to the higher level information Object/intellectual entity.

File Formats:

- Signatures for format identification (to support automated format identification with 'levels of confidence' in the identification using a particular method)
- Potential embedded metadata (to support automated metadata extraction)
- Recommended embedded metadata (for producers)
- Significant properties of a format (to permit the selection of which properties are significant in the preservation of a particular object)
- Creating Applications (possible application which may have created a file format; see software)
- Environment:
  - Hardware (recommended, minimal and known to work for rendering a particular format, including dependencies; see hardware/software)
  - Software (recommended, minimal and known to work for rendering a particular format including dependencies; see hardware/software)

Hardware/Software:

- For reference by file formats: a collection of hardware recommended, minimal or known to work in rendering a particular format with a particular subset of significant properties.
- For reference by file formats: a collection of software recommended, minimal or known to work in rendering a particular format with a particular subset of significant properties.

5 References

5.1 DDI 3 evaluation

DDI 3 Standard
Technical Specification, DDI 3.1 (2009-10-18), full download package including:

- Technical Specification Part I: Overview Version 3.1, October 2009
- Technical Specification Part II: User Guide Version 3.1, October 2009
- Part IV XML Schema files and field-level documentation

Presentation of DDI 3 uses cases and recommendations

- DDI Publications: Working Paper Series, Papers, Presentations, Reports
- CESSDA PPP - Outputs and Publications: Tendered Report:
  Technical Specifications for a European Question Data Bank
  (Metadata Technologies Ltd. 2009)
- IASSIST 2009 - Thu 28 May
  Session D1 "Tag - You're it! DDI Applications and Experiences "
  Presentation
  "Managing the Metadata Life Cycle: The Future of DDI At GESIS and ICPSR" slide 14
- CESSDA Expert Seminar, UK Data Archive, September 2007
  "Introduction to DDI 3.0." (PPT 2.1MB). Presentation by Sanda Ionescu, ICPSR

ISSP use case material sources

- ISSP modules in general
- ISSP Module Role of Government I-IV and the Cumulation

Literature and Presentations

- DataShare project - Luis Martinez
  "The Data Documentation Initiative (DDI) and Institutional Repositories" 28 Feb 2008
- Ann Green, Yale University
  A tour of the DDI held at Cornell University 27.2.2009

5.2 Preservation metadata

Technical Specifications for a European Question Data Bank. Final Version May 2009
Arofan Gregory / Pascal Heus / Chris Nelson (Metadata Technology) and Jostein Ryssevik (Ideas2evidence)
http://www.cessda.org/project/doc/CESSDA_PPP_QDB_May09.pdf

Reference Model for an Open Archival System (OAIS)
Consultative Committee for Space Data Systems - CCSDS 650.0-P-1.1 (May 2009)
http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206500P11/Attachments/650x0p11.pdf

METS - Metadata Encoding and Transmission Standard
www.loc.gov/standards/mets/

PREMIS - PREservation Metadata: Implementation Strategies;
Data Dictionary version 2 (March 2008)
http://www.loc.gov/standards/premis/