QUESTASY: DOCUMENTING AND DISSEMINATING LONGITUDINAL DATA ONLINE USING DDI 3

By Alerk Amin with Michelle Edwards, Oliver Hopt, Jannik Jensen, Dan Kristiansen, Olof Olsson and Joachim Wackerow

12/07/09

DDI Working Paper Series -- Use Cases, No. 1

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

Questasy: Documenting and Disseminating Longitudinal Data Online Using DDI 3

BY ALERK AMIN WITH MICHELLE EDWARDS, OLIVER HOPT, JANNIK JENSEN, DAN KRISTIANSEN, OLOF OLSSON AND JOACHIM WACKEROW

ABSTRACT

Questasy is a Web application developed to manage the dissemination of data and metadata for panel surveys. It was primarily developed for the LISS Data Archive, but was designed to be repurposed for other surveys. The structure of the application, from the underlying database to the generated Web pages, is based on DDI 3. This paper describes how Questasy was designed and implemented.

BACKGROUND

LISS Panel

The LISS (Longitudinal Internet Studies for the Social sciences) panel is the principal component of the MESS (Measurement and Experimentation in the Social Sciences) project. It consists of 5000 households in the Netherlands, comprising 8000 individuals. Panel members complete online questionnaires every month, which take approximately 30 minutes. One member of the household provides the household data and updates this information at regular time intervals.

Half of the interview time available in the panel is reserved for the LISS Core Study. This longitudinal study is repeated yearly and is designed to follow changes in the life course and living conditions of the panel members. The other half of available interview time each year can be used to collect data for external projects. This is cost-free for researchers at universities and scientific institutes. Researchers from the Netherlands and abroad are free to submit survey proposals. The panel has been in full operation since October 2007.

To disseminate the data for the panel, a Web site was desired to manage the distribution of the data files. In addition, LISS project staff wanted the metadata about the variables to be available to researchers, in a better form than PDF/Word codebooks. For this reason, it was decided to build a data dissemination Web site that would provide advanced functionality for researchers to browse and search the metadata online. To implement this Web site, the Questasy application was created.

Questasy

Questasy is an online data dissemination tool developed primarily for the LISS panel. Questasy manages both metadata and data and provides an easy-to-use data entry module for administrators to create metadata. The Web interface allows researchers to browse and search the metadata and download datasets. The Questasy system also manages files, tracks downloads, and creates Web pages for viewing documentation including studies, concepts, questions and variables. Due to the longitudinal nature of the LISS panel, the ability to track questions and variables was a key requirement of the system. To support this, DDI 3 was chosen as the basis of the application.

Life Cycle Context

Questasy collects metadata from many aspects of the life cycle of LISS surveys. From the data production process, metadata about the Concepts, Collection, and Processing steps are collected. Additionally, the resulting data files can be stored in Questasy.

Once the metadata and data are in Questasy, it functions as a local archive, but the data can be exported to a permanent archive.

Questasy provides several services to researchers. Metadata Distribution is provided through the Web interface, Discovery is provided via the online search, and the data files are available for Analysis, usually with a statistical software package.

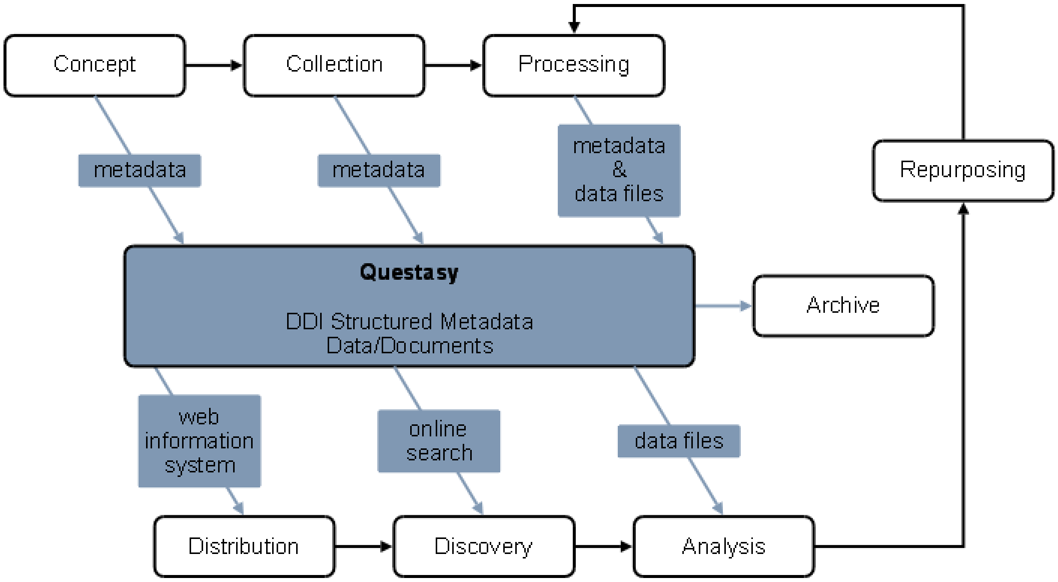


Figure 1: Structure of Questasy

USE CASE / REQUIREMENTS

Metadata Requirements

Questasy was developed with the researcher in mind, but in two capacities: one as the collector and cleaner of data (referred to as "administrator") of the LISS panel and second, as a user of the metadata and data (referred to as "researcher").

Administrators clean the collected data and prepare data files in SPSS and Stata formats, then load them into Questasy for immediate access by researchers. Variable metadata are imported from SPSS and added to the system. Additional metadata, such as questions and concepts, are added via Web forms. Efficiency and ease of adding metadata to an existing internal workflow have increased by convincing internal researchers administering the data to use DDI, knowingly or not.

Once the data are available in Questasy, researchers from around the world can search and browse the metadata for each of the surveys available. Data downloads are possible after signing a licensing agreement with CentERdata.

All surveys for the LISS panel are conducted in Dutch. In the LISS Data Archive, question text is distributed with the original Dutch text, as well as the English translation. All other metadata are distributed only in English. The Questasy application is capable of supporting multiple languages throughout its interface, but this is not used on the LISS Data Archive.

Design Process

The LISS administrators wanted a system capable of data and metadata dissemination. Questasy developers created a proposal outlining a system that met the needs of the researchers and used DDI. Administrator involvement at the beginning was crucial since implementation of the project would affect their workflow. With no DDI experience, Questasy developers approached the administrators with diagrams and "English" vs. DDI explanations. Speaking the same language is a great benefit.

During the initial planning phase, developers and administrators found fields in DDI for the metadata that were already being delivered in PDF/Word codebooks. They also looked through DDI for ideas about new metadata that could be delivered, including metadata that were already available internally, but not being distributed to users.

Word documents with tables and diagrams travelled back and forth between developers and administrators, resulting in a comprehensive list of items the administrators felt met their current and immediate future needs. Technical fields that were important for the developers to maintain were also included in the project implementation.

POSSIBLE SOLUTIONS and DESIGN CHOICES

System Design

At the beginning of the project, several options were evaluated. The developers investigated software packages, such as Nesstar, but also looked at other custom-built Web sites. The main requirement was support for longitudinal studies, which eliminated most options. The options which did support longitudinal studies were not flexible or comprehensive enough, so the decision was made to build a new system.

While the main focus for the new system was to support the LISS panel, it was designed to be customizable and extendable, to support other studies in the future.

The decision to use a PHP framework on top of a relational database was based on several factors. The most important was previous experience within CentERdata. XML databases were considered, but were rejected based on performance considerations. It was determined that a relational database could easily scale to the usage required. As DDI is a nonproprietary, system-independent standard, the developers decided that it would be easy to interoperate with other systems via a DDI import/export implementation. Thus, the choice for the internal storage would not affect other systems.

Database transactions are an important part of the system, to maintain the integrity of the database. Many input screens can affect multiple tables in the database. For example, entering a new question can create new Question Items, Response Domain, Code Schemes, and Codes. All of these data are collected via a single Web form, and then processed at once in a transaction. If the inserts are successful, the transaction is committed. If there is an error in processing the data, the transaction is rolled back, and the errors are shown to the administrator, who can then fix the errors and resubmit. This ensures the integrity of the database, especially that all of references between elements remain valid.

The search is a very important feature of Questasy. To make it easy for researchers to search for text that might appear in different tables, such as question and answer text, the developers decided to use a search engine that would index across tables. The Sphinx search engine was chosen because of PHP support, performance, and flexibility.

DDI and Relational Database

DDI 3 was chosen as the architectural basis of Questasy. The entire DDI hierarchy was analyzed to determine which elements were required to document the LISS metadata. These elements were then converted to a relational database schema.

The major DDI elements, such as Question Items and Variables, were mapped to tables in a relational database. The fields in the tables correspond to the fields in DDI.

The relationships between DDI elements are extremely important to the system. The biggest benefit is the tracking of Question Items across waves in the study, where each wave can have Question Constructs and Variables that refer to the same Question Item. Relations between DDI elements can occur in two ways: through the normal hierarchy, or through references. Both of these types of references were mapped to one-to-one, one-to-many, and many-to-many relationships between the tables. For many-to-many relationships, join tables were used.

Some fields were added to the database for internal use, and do not map to the DDI hierarchy. An example of this is in the Control Construct Schemes table, which includes the name of the source file for the questionnaire. This is used for internal purposes, but is not shown in the researcher interface, or part of the DDI standard.

Substitution groups in DDI presented some problems, specifically for Response Domains. In DDI, a Response Domain is a placeholder for a Text Domain, Numeric Domain, Code Scheme, or other domain. This type of inheritance can be resolved in several ways in the database. Single-table inheritance was chosen. The Response Domains table contains all of the fields required for all of the various Domains, and a flag to signify which type of Domain it is.

The database design does not support versions of elements, with one exception. Sometimes, errors are found in datasets after they are released. These errors are fixed, and new datasets are released. To support this activity, Physical Data Products can be versioned, to keep track of the various releases of a dataset. However, only the latest version of a dataset is downloadable by researchers, and only the metadata for the latest version are displayed on the Web site.

Some DDI elements require a tree structure, such as Groups/Study Units, Concept hierarchies, and Control Constructs. These were implemented in the database schema using left-right trees. These are supported by the application framework and have very good performance for query operations.

Once the database schema was determined, the application was implemented in PHP, using the CakePHP framework. This MVC (Model-View-Controller) framework greatly simplifies the process of building a Web site on top of a relational database. The Web forms for data entry, as well as the views for researchers, are created by running queries against the database, and creating HTML pages with the resulting data.

Overall Timelines

Developing the framework for the project with the researchers took approximately two months, followed by another twelve months of development and refinement.

A large portion of the design time was spent on determining how to use DDI as a basis for the project. It took approximately eight months to develop a working system and another four to tweak and massage the system to match everyone's requirements. During the development phase of the project there were 1.5 – 2 FTE on the project, with 1 FTE currently available for both development and support of the production system.

RESULTS

LISS Data Archive Web site

The LISS Data Archive Web site went live in early 2009, for internal administrators to begin entering data and metadata. The Web site went live for external researchers in March 2009, and has been well received. Traffic to the Web site has been better than expected, and Web statistics show that researchers are making use of all of the functionality available.

Researchers can browse studies that have been run in the panel. In the Study Unit view, they can see the metadata for the Studies, including information about the Abstracts and Data Collection. They can also download the data files and other materials associated with each study.

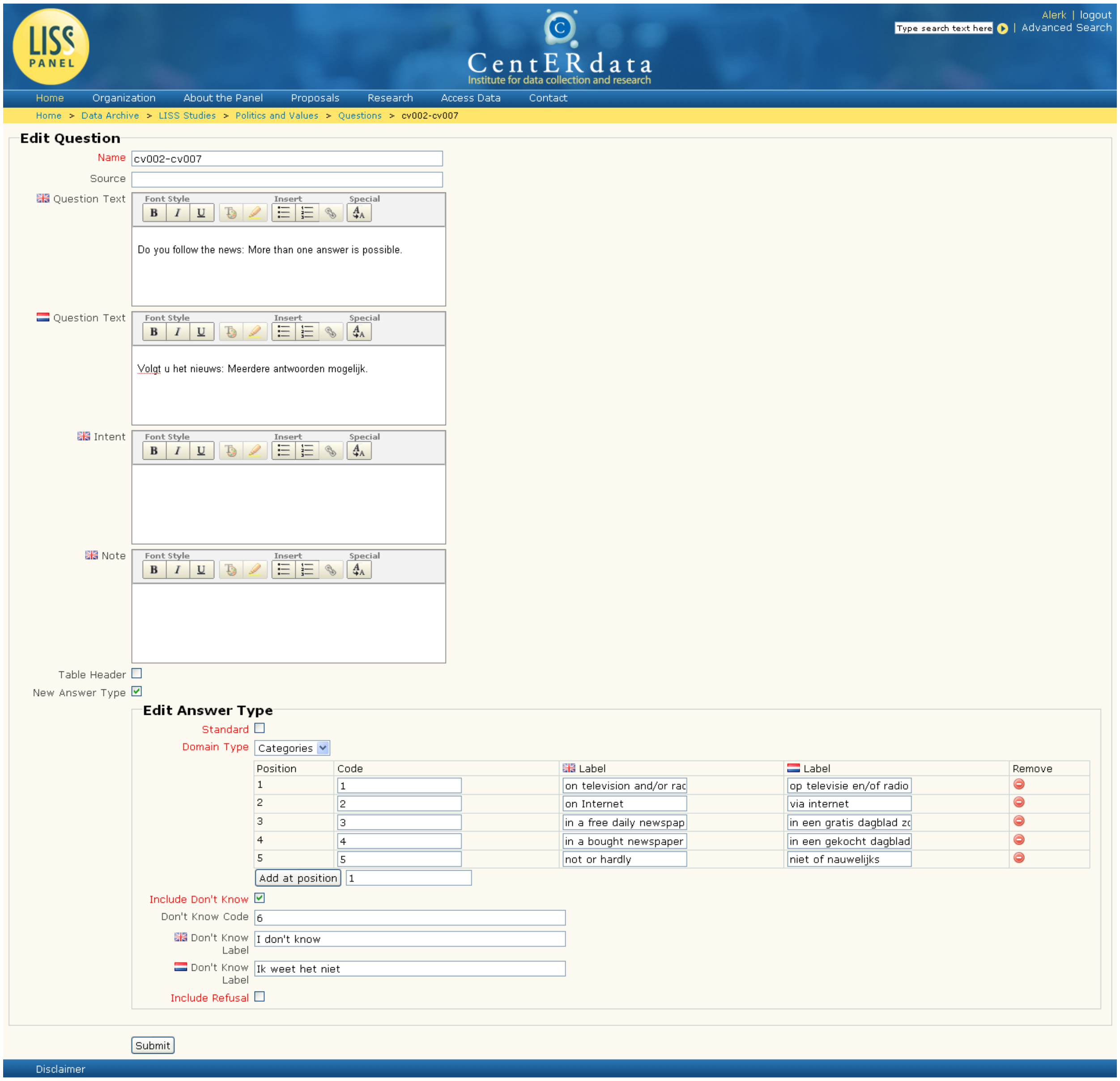
Researchers can also browse the Concepts. The Concepts are organized into a tree. From the Concepts, researchers can view the associated Variables and continue to navigate through the metadata.

When viewing Question Items, the Variables that are associated with the item are listed. For Longitudinal Studies, this includes the Variables across all the waves of the Study.

Currently, Question Constructs are shown as a list, in the order that they are asked in the Questionnaire, without detailed routing information. The "sample" field, which is equivalent to the DDI Universe field, provides information about which subset of the panel answered a particular question.

A search engine with full text indexing has been integrated. The main advantage of the search engine is that it can index complex items, including fields combined from multiple tables. The search results are ranked according to relevance, providing the researchers with the best results possible. The searches also execute significantly faster than against the MySQL database.

For the administrator interface, we have implemented a full system of Web forms to enter and manage the data and metadata. The Web forms handle the DDI relationships automatically, making it easy for administrators to enter the metadata without prior knowledge of DDI.


LISS PANEL
CentERdata
Institute for data collection and research
Alerk | logout
Type search text here | Advanced Search
Home
Organization
About the Panel
Proposals
Research
Access Data
Contact
Home > Data Archive > LISS Studies > Politics and Values > Questions > cv002-cv007
Font Style
Insert
Special
Disclaimer

Edit Question

Name: cv002-cv007

Source:

Question Text: Do you follow the news: More than one answer is possible.

Question Text: Volgt u het nieuws: Meerdere antwoorden mogelijk.

Intent:

Note:

Table Header ☐

New Answer Type ☑

Edit Answer Type

Standard ☐

Domain Type: Categories

Position	Code	Label	Label	Remove
1	1	on television and/or rac	op televisie en/of radio	⊖
2	2	on Internet	via internet	⊖
3	3	in a free daily newspap	in een gratis dagblad zc	⊖
4	4	in a bought newspaper	in een gekocht dagblad	⊖
5	5	not or hardly	niet of nauwelijks	⊖

Add at position: 1

Include Don't Know ☑

Don't Know Code: 6

Don't Know Label: I don't know

Don't Know Label: Ik weet het niet

Include Refusal ☐

Submit

Figure 2: Questasy Administrator Web Form

To speed data entry, Questasy includes an SPSS import for variable information. In SPSS, the administrator can create an OMS file from the display dictionary command. This OMS file can be imported into Questasy to automatically import the Variable metadata, including Representations. The administrators can then enter the questions and additional metadata associated with the variables.

DDI Usage

The list of DDI elements used by the Questasy system includes:

- Citation
- Code / Code Scheme
- Coding
- Collection Event
- Concept / Concept Scheme
- Conceptual Component
- Data Collection
- Funding Information
- Group
- Organization / Organization Scheme
- Other Material
- Physical Data Product
- Physical Instance
- Question Construct / Control Construct Scheme
- Question Item / Multiple Question Item / Question Scheme
- Representation
- Response Domain
- Study Unit
- Variable / Variable Scheme

The development of an export function to generate DDI XML is currently under way.

ISSUES / RESTRICTIONS

DDI

Grouping currently occurs at the top level for studies and is mainly limited to Concept and Organization Schemes. Grouping also occurs within longitudinal studies, to allow individual waves to share a common Question Scheme. Questasy takes advantage of DDI inheritance to reuse items within individual studies.

DDI metadata are currently entered into the system after the data files are prepared; CentERdata is looking for ways to harvest DDI metadata as it happens throughout the lifecycle of LISS surveys.

General

Questasy was developed primarily for one application, the LISS surveys, but with an eye towards making it reusable for other studies. Extending its application to studies beyond LISS will depend on characteristics of new studies.

Metadata are currently entered into Questasy by a student working approximately one day per week. Currently, only variable-level metadata can be automatically imported into the system, via SPSS OMS files. In the future, more of this process would be automated, especially importing of question and response domains from the questionnaire engine. CentERdata is also looking at ways to capture metadata from the data processing and cleanup steps.

Question flow from the Blaise system used in data collection is also not captured at the moment but is of great interest to the developers, internal administrators and external researchers. The current solution uses the Universe element to give some information, but investigation is under way in terms of how to implement the full questionnaire routing.

OUTLOOK / CONCLUSION

The Questasy system has been widely accepted with very positive feedback from both internal and external researchers. However, developers are looking into the future for further improvements and developments.

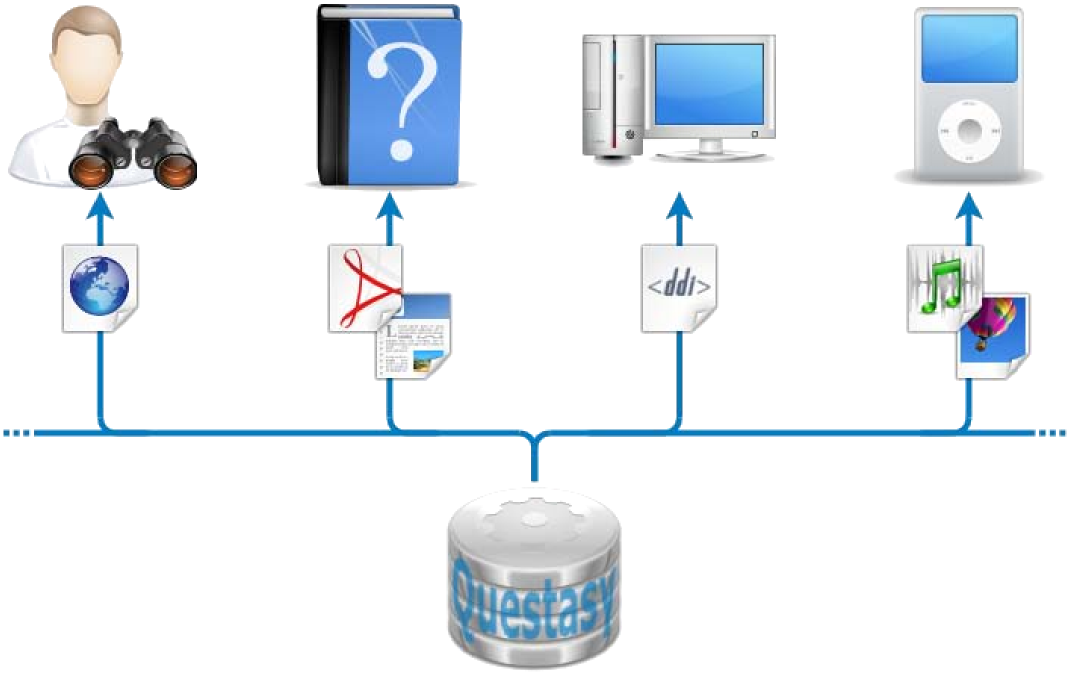


Figure 3: Future Delivery Mechanisms for Questasy

Questasy can deliver the information in several ways. Currently, the Web interface for researchers is in production. Development is currently under way on a DDI XML export, to deliver the structured metadata to other applications. Development is also under way on delivering content to portable devices such as iPods. In the future, paper codebooks in PDF/Word format may be generated directly from Questasy.

Future Developments

A customized basket for selecting variables from waves across the years is a feature that would enhance the researchers' ability to download data. Current downloads are restricted to the dataset level, but the researchers' Questasy experience would be improved by adding variable level subsetting.

Another project is currently investigating the integration of enhanced publications, with variable-level metadata about the publications. This involves giving researchers the ability to list the publications they have written based on Questasy data, and link the publication to the studies and variables they used in their research.

As the second wave of surveys are released through Questasy, harmonization and the relationship between and among wave variables will become of great interest to researchers. Additional views may be implemented to provide researchers with a clear overview of how the questions/variables are comparable across waves. The Comparison module will be investigated for this.

APPENDIX A

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

Workshop title:

Workshop on Implementation of DDI3 - Advanced Topics

Organizers:

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

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

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

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

APPENDIX B

This is a human-readable summary of the Legal Code (the full license).
http://creativecommons.org/licenses/by-nc-sa/3.0/us/

You are free:

- to Share - to copy, distribute, display, and perform the work
- to Remix - to make derivative works

Under the following conditions:

- Attribution. You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).
- Noncommercial. You may not use this work for commercial purposes.
- Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar license to this one. For any reuse or distribution, you must make clear to others the license terms of this work. The best way to do this is with a link to this Web page.
- Any of the above conditions can be waived if you get permission from the copyright holder.
- Apart from the remix rights granted under this license, nothing in this license impairs or restricts the author's moral rights.

Disclaimer

The Commons Deed is not a license. It is simply a handy reference for understanding the Legal Code (the full license) — it is a human-readable expression of some of its key terms. Think of it as the user-friendly interface to the Legal Code beneath. This Deed itself has no legal value, and its contents do not appear in the actual license.

Creative Commons is not a law firm and does not provide legal services. Distributing of, displaying of, or linking to this Commons Deed does not create an attorney-client relationship. Your fair use and other rights are in no way affected by the above.

Legal Code:
http://creativecommons.org/licenses/by-nc-sa/3.0/us/legalcode