Questasy: Online Survey Data Dissemination Using DDI 3

Marika de Bruijne and Alerk Amin[1]

Introduction
Based on the DDI Working Paper 'Questasy: Online Data Information and Dissemination Using DDI 3'. Special thanks to the co-authors of this paper: 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 survey projects. It was primarily developed for the LISS Data Archive, but was designed to be repurposed for other archives as well. The application has been operational since 2009 and its external web interface can be viewed at: www.lissdata.nl.

Questasy manages both metadata and data and provides an easy-to-use data entry module for administrators to create metadata. The external 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 many of the LISS panel studies, the ability to track questions and variables throughout a study was a key requirement of the system. To support this, DDI 3 was chosen as the basis for the structure of the application, from the underlying database to the generated web pages.

This paper describes the main functionality of Questasy and how we designed and implemented the system.

1. 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, totaling approximately 30 minutes per month.

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 per year is used to collect data for external research projects in different disciplines within social sciences. This is cost-free for purely scientific research. Researchers from both the Netherlands and abroad can submit survey proposals. The panel has been in full operation since the end of 2007.

Next to the LISS Core Study, many of the other LISS studies are longitudinal. In these studies, questions are repeated in new measures (waves) to the same respondents in order to measure changes over time.

One of the goals of the LISS panel is to make the collected data available to the international scientific community. Due to the longitudinal character of the data, we needed to pay special attention to the way the data and metadata of all measures of a study would be presented.

2. The Project
Questasy was developed with the researcher in mind, but in two capacities: one as the internal employee (referred to as "administrator") and second, as an external 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 onto Questasy for immediate access by researchers. Once the data are available in Questasy, researchers from around the world can search and browse the metadata for each of the surveys available.

2.1 Start
In order to disseminate the data and metadata for the LISS panel surveys, we started with relatively simple application requirements. Soon we realized that if we wanted to provide researchers with a better tool for browsing and searching the metadata of our largely longitudinal studies, we would need a more advanced solution than, for instance, simply publishing PDF or Word codebooks.

At the beginning of the project, we evaluated several existing applications. We investigated software packages, such as Nesstar, but also looked at other custom-built websites. Our requirement to support longitudinal studies eliminated most options. The options which did support

longitudinal studies were not flexible or comprehensive enough for our needs, so the decision was made to build our own system: Questasy.

2.2 Choice for DDI 3

The choice for DDI 3 was initially not an obvious one. The main reason that DDI 3 piqued our interest was its support for longitudinal studies. When we started studying the DDI 3 way of managing each detailed entity of a survey project as its own controllable element, this first seemed to cause a large amount of extra work in terms of data entry, which we hadn't expected in advance.

In particular, the separation of question and variable metadata, and full documentation of both, required 'out-of-the-box' thinking from our administrators who were used to think about data dissemination in terms of codebooks, containing mainly variable metadata.

Although we believed the structure of DDI 3 would support our needs, it was in fact only later through

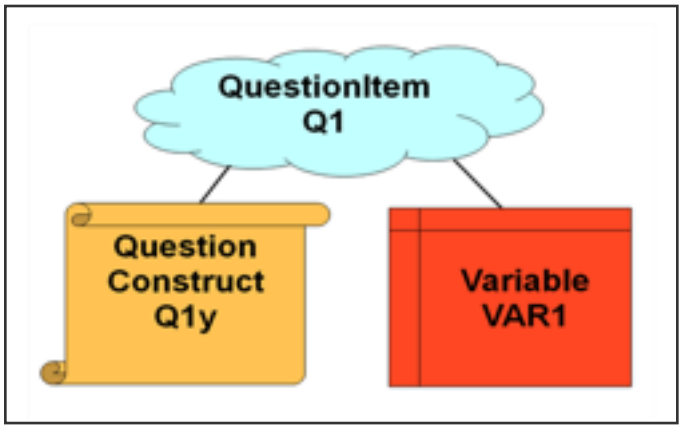

Figure 1. Figuring out relations between questions and variables

practical examples of more complex survey data, that the full advantage of this structure was truly experienced. Finally, we had found a system that enabled documenting the questions the way they had been presented to the respondents and the data variables the way they were contained in the final dataset, without losing information from either side.

2.3 DDI Elements in Questasy

After having made the decision to use DDI 3, Questasy developers created a proposal outlining a system that met the internal user requirements 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 translations. Talking the same language is a great benefit.

During the initial planning phase, we found fields in DDI for the metadata that were already being documented and included in codebooks. We also looked through DDI for ideas about new metadata that could be delivered, including metadata that was already available internally, but not being distributed to data users.

Word documents with tables and diagrams travelled back and forth between developers and administrators to result 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.

The DDI elements that were chosen to be used by Questasy are:

- 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

2.4 DDI 3 as Architectural Basis

DDI 3 was chosen as the architectural basis of Questasy. After having analyzed the DDI hierarchy to determine which elements were required to document the LISS metadata, the chosen 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 were 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 2 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.

A couple of 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, to which we added the name of the original source file for the questionnaire. This working name is used only for internal purposes and therefore not shown in the researcher interface, nor is it part of the DDI standard. As these extra fields will not be part of a DDI export, it was important to find a DDI equivalent for all the fields we wanted to publish.

Substitution groups in DDI presented some problems, specifically for Response Domains. In DDI, a Response Domain is 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. We chose single-table inheritance. 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, Variable Schemes 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 is displayed on the website.

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.

2.5 System Design

Once the database schema was determined, the web application was built using a PHP framework. The application queries the database to save/retrieve data (such as questions or variables). It then formats the data as HTML, which are then delivered to the end user's web browser. Both the web forms for data entry as well as the views for researchers are created this way.

We decided to use a PHP framework on top of a relational database due to several factors. The most important was previous experience within CentERdata. XML databases were considered, but were rejected based on performance considerations. We determined that a relational database could easily scale to the usage we required. As DDI is a file-format standard, we decided that we could easily interoperate with other systems via a DDI import/export implementation. Thus, our 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 data entry 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. We wanted to make it easy for researchers to search for text that might appear in different tables, such as question and answer text. For this reason, we decided to use a search engine that would index across tables. The Sphinx search engine was chosen because of PHP support, performance, and flexibility.

2.6 Overall Timeline

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 8 months to develop a working system and another 4 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.

3. Website Usage

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

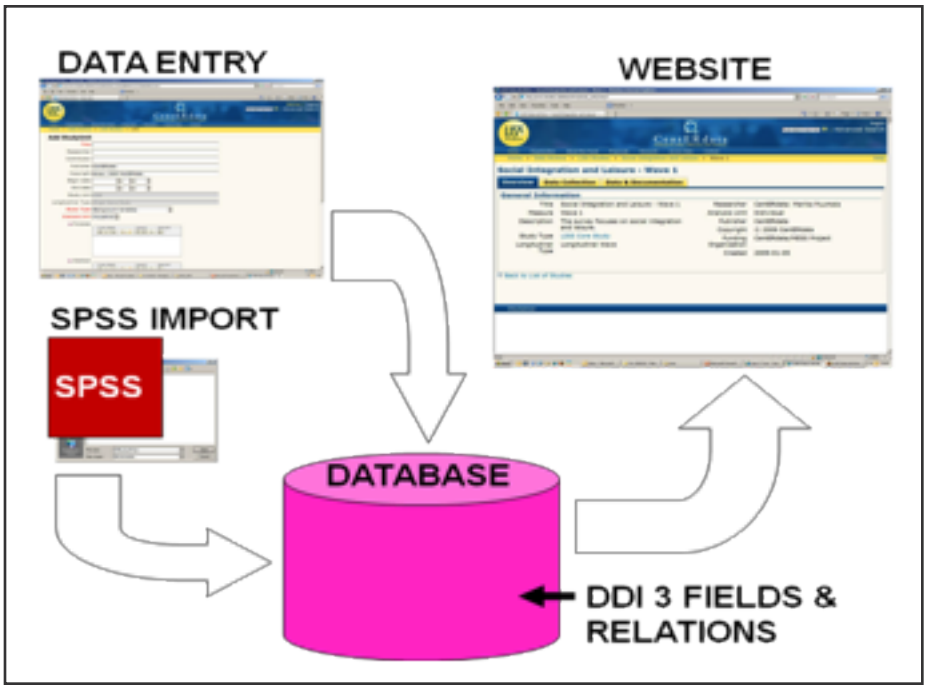

Figure 2. Questasy supports data entry via webs forms and automatic SPSS metadata import

3.1 Data Entry

In Questasy, one can capture metadata from many aspects of the lifecycle of LISS surveys. From the data production process, metadata about the Concepts, Collection, and Processing steps are collected. And of course, the resulting data files can be stored in Questasy.

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. Metadata is currently entered into Questasy by a student working approximately one day per week.

To speed data entry, we have implemented an automatic import for variable metadata captured in SPSS. In SPSS, the administrator first creates an xml file in an OMS session (Output Management System) for the SPSS dictionary information. This can be imported into Questasy to import the Variable metadata, including Representations. The administrators then enter the questions and additional metadata associated with the variables via web forms.

3.2 External Website

On the public website, researchers can browse studies that have been conducted 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. To download datasets it is first required to fill out an agreement about the use of data.

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.

While the Questasy database contains Questions Items, Control Constructs and Question Constructs, we don't present all these relations on the external website. Here, we have simplified the information about questions showing an integrated view for Question Item and Question Construct information. In fact, Question Constructs, and related Question Item information, are shown as a list, in the order that they are asked in the Questionnaire.

We have also integrated a search engine with full text indexing. The 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.

All surveys for the LISS panel are conducted in Dutch. In the LISS Data website, question texts are 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, although this is not used in the LISS data website.

4. Issues and Restrictions

Looking at ways in which our use of DDI is somewhat restricted, the use of Grouping could be mentioned. 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.

Currently, only variable-level metadata can be automatically imported into the system, via SPSS OMS files. In the future, we would like to automate more of this process. Especially importing of question and response domains from the questionnaire engine would make a big difference in required data entry effort.

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

Finally, while Questasy was developed primarily for one application, the LISS panel, it was designed to be customizable and extendable, so that it could support other studies in the future.

it relatively easy to create new forms of DDI compatible exports.

A customized basket for picking out variables from waves across the years is a feature that would enhance the researchers' ability to download data. While current downloads are restricted to the dataset level, the researchers' Questasy experience would be improved by adding variable level sub-setting.

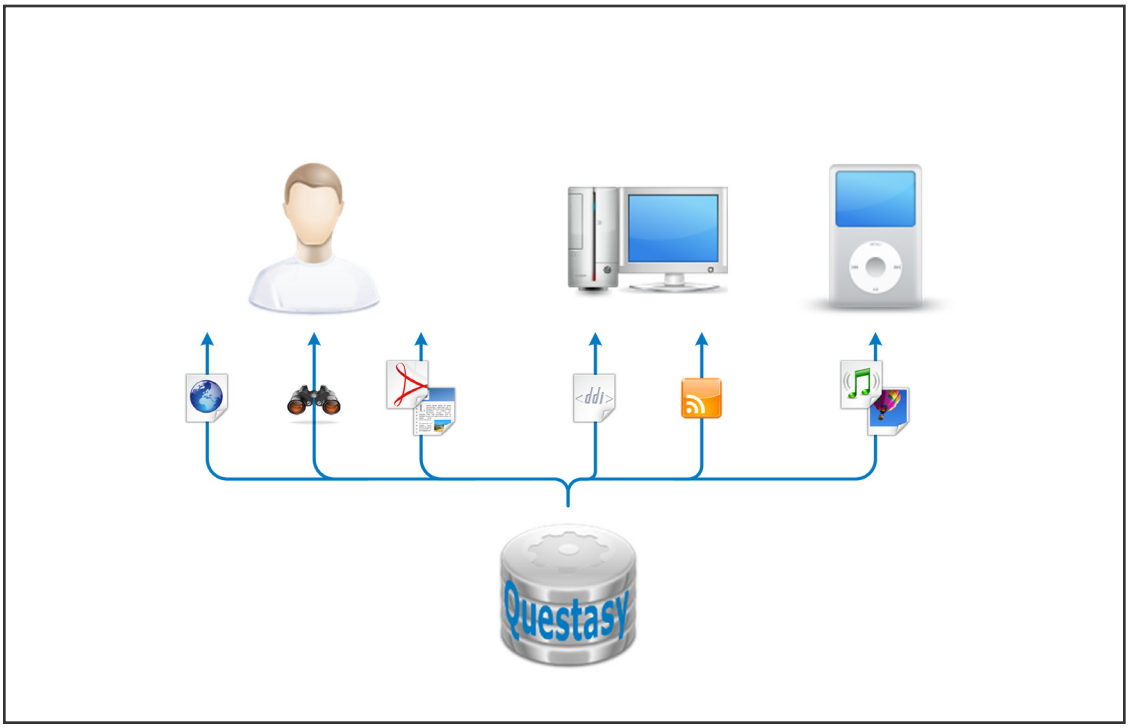

Figure 3. Questasy provides possibilities for delivering metadata in various ways

5. Outlook and Future Developments

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

Questasy can deliver the information in several ways. Currently, the web interface for researchers is operational. Once the metadata and data are in Questasy, the system functions as a local archive, but the data could be exported to a permanent archive. The development of an export function to generate DDI XML is currently underway, to deliver the structured metadata to other applications. There have also been some experiments on delivering content to portable devices such as iPods. In the future, paper codebooks in PDF/Word format might be generated directly from Questasy. The great benefit of a relational database in which the DDI structure is embedded is that this makes

Current projects underway are looking at the integration of enhanced publications, with up to 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.

Finally, as the second wave of the LISS Core Study is being released through Questasy, including changes to some questions, harmonization and the relationship between and among similar variables have become of great interest to us. Additional views may be implemented to provide researchers with a clear overview of how the questions and variables are comparable across waves. The Comparison module will be investigated for this.

6. Summary: Main Benefits

To summarize, the main benefits we have experienced

while using Questasy and DDI 3 for disseminating our data are:

1. Separating the documentation of questions and variables and enabling many-to-many relations between these
Fully documented variables only point to the survey questions they are based on. This enables full documentation of both entities within a survey project, without losing any information such as original question formulation. For example: The data based on a question where multiple answers are possible are often processed into several 'dummy' (0,1) variables, one variable containing the answers to one answer option. In Questasy this can be presented by creating a single question and several variables which each point to the same question.

2. Longitudinal data comparison via questions within longitudinal studies
Questions can be reused within a longitudinal study. This enables creating a connection between variables in different waves that are based on the same question.

3. Several options for searching data on detailed level
Not only information on study level, but also metadata of both questions and variables can be searched by entering keywords. The website visitor can choose in which fields to search: study units, concepts, Dutch or English question text or variable metadata, or all of these. It is also possible to browse the contents of the database in different ways: by studies or concepts, for example, and in the future by topics and year.

4. Relational database enables presentation of metadata in many ways
Thanks to separating pieces of metadata on a detailed level into their own manageable fields in a relational database, Questasy provides a flexible basis for creating various different views on the website to data users. Further, since the metadata of Questasy is structured according to DDI 3, it is relatively easy to create new forms of output that are DDI compatible.

Notes

1. Marika de Bruijne (m.debruijne@uvt.nl), Alerk Amin (a.amin@uvt.nl) . CentERdata, University of Tilburg .PO Box 90153, 5000 LE TILBURG, The Netherlands.Phone: 013 466 8325 / 8326. Fax: 013 466 2764