

Data Service Infrastructure for the Social Sciences and Humanities

EC FP7

Grant Agreement Number: 283646

**DASISH WP 3.2.c: [Question Variable Data Base]**

Nature: R

Responsible: Knut Kalgraff Skjåk, Norwegian Social Science Data Services (NSD)

Work Package Leader: Eric Harrison, City University (ESS)

Nature: R

Contributing Partners and Editors:

Taina Jäskelainen, FSD

Håvard V. Bakkmoen, Hilde Orten, Knut Kalgraff Skjåk, NSD

Joachim Wackerow, Metadata expert consultant

Version: 2.0

**CONTENTS**

[1 Introduction 3](#_Toc361914595)

[1.1 Purpose of the system 3](#_Toc361914596)

[1.2 Objectives and success criteria of the project 3](#_Toc361914597)

[1.3 Acronyms and Abbreviations 4](#_Toc361914598)

[2 Current system 4](#_Toc361914599)

[2.1 The question data base of the ESS 4](#_Toc361914600)

[2.2 Other systems 5](#_Toc361914601)

[3 Stakeholders 5](#_Toc361914602)

[3.1 ESS and other DASISH projects 5](#_Toc361914603)

[3.2 Other social science projects and institutions that are collecting, archiving and/or disseminating data 7](#_Toc361914604)

[3.3 Networks of projects within DASISH 7](#_Toc361914605)

[3.4 Users within the research community (researchers and students) 7](#_Toc361914606)

[3.5 Administrators and developers of other tools and systems 7](#_Toc361914607)

[4 Proposed system 8](#_Toc361914608)

[4.1 Overview 8](#_Toc361914609)

[4.2 Functional requirements 8](#_Toc361914610)

[4.3 Nonfunctional requirements 9](#_Toc361914611)

[5 System models 11](#_Toc361914612)

[5.1 Example Use Cases 11](#_Toc361914613)

[5.2 System architecture 13](#_Toc361914614)

[5.3 Connections with other systems 15](#_Toc361914615)

[References: 16](#_Toc361914617)

# 1 Introduction

## Purpose of the system

Access to survey questions in their original languages is often severely limited.The scope of the WP3, task 2c Question Variable Data Base (QVDB) is to create a searchable database with a broad public profile, providing user access to survey questions in their original languages, the concepts they are based on, additional material used in the data collection, the resulting variables and their generation, coding classifications and more.

The primary aim of the tool under development aims to serve the ESS and other DASISH projects in their business processes[[1]](#footnote-1), by facilitating reuse of metadata components at the different stages and sub-stages of these processes.

The tool will also allow users from within the research community to browse the database (question, concept and variable mining) in order to design new research or to explore existing projects.

Developed initially with the European Social Survey (ESS) in mind, the Question Variable Data Base will be designed to service projects external to DASISH, to interoperate with other systems and tools, and to act as a reusable model for other question databanks.

## Objectives and success criteria of the project

The aim of WP3, task 2.c in the DASISH project is to create a system with capacities to serve internal business processes of the ESS and other DASISH projects, as well as to provide access to the research community of questions in original languages, concepts, variables etc.:

The project is successful if:

* system requirements, user needs, functional and non-functional requirements for the QVDB have been identified and described in a suitable document (= this requirement analysis document);
* a reusable database model has been developed;
* compatibility with the DDI-Lifecycle is achieved, to allow for interoperability with other tools and systems;
* a system architecture that enables easy exchange of metadata in DDI format to the other task 3.2 tools has been built;
* this should comprehend file-based import and export, as well as communication by web-services;
* a technical solution/software will be selected that will allow the functional requirements specified at point 4.2 and 4.3 to be fulfilled, and will be approved by the DASISH EB;
* an implementation schedule has been drawn up;
* the technical platform for the QVDB has been set up in accordance with the requirements analysis;
* the QVDB has been populated with content in accordance with the requirements of the DASISH DoW;
* the QVDB has been tested for compliance with the identified use cases by members of WP3.2 as well as by selected external persons;
* the QVDB has been launched;

## Acronyms and Abbreviations

CAPI – Computer Assisted Interview

CESSDA – Council of European Social Science Data Archives

DASISH – Data Service Infrastructure for the Social Sciences and Humanities

DDI – Data Documentation Initiative

DoW – Description of Work

EB – Executive Board

ESS – European Social Survey

GLBPM – Generic Longitudinal Business Process Model

NSI – National Statistical Institute

PAPI – Paper and Pencil Interview

PPP – Preparatory Phase

QDT – Questionnaire design team

QDDT – Questionnaire Design and Development Tool

QVDB – Question Variable Data Base

RDF – Resource Description Framework

SQP –Survey Quality Prediction

TMT – Translation management tool

WP – Work Package

XML – Extensible Markup Language

# 2 Current system

The QVDB will be developed and set up in DASISH WP 3.2, and will replace the current ESS question database. A selection of existing tools will be explored with the purpose of getting ideas for the new tool. The possibilities of reuse of components from these tools will also be explored.

## The question data base of the ESS

The existing ESS question data base is one of several databases serving the ESS Data Archive in their business processes, first and foremost for the purposes of documenting datasets resulting from the survey. The data base stores metadata components related to questions and variables used in the ESS, and is also used by some other surveys.

The existing ESS data base is a legacy system that is not generic and is difficult to change. In order to achieve the goals and purposes of the planned QVDB (see 1.1), a new system will thus need to be developed.

## Other systems

Other systems as specified below will be explored to learn from their functions and get ideas for the new tool, and will also be evaluated with the purpose of possible reuse of components. Several of the projects specified below are developed or utilized within CESSDA and NSI/Eurostat institutions and are based on DDI-Lifecycle:

* The Questacy database: Questasy is a web application developed to manage the documentation and dissemination of data and metadata for panel surveys. The Questacy system is freely available and developed by CentERdata.
* DdiEditor: DdiEditor is the key tool in a framework of data processing tools and processes composing data processing of survey datasets. The end product is data documentation in accordance with international metadata standards. The DdiEditor is developed by DDA and is freely available under a LGPL licence.
* MISSY: The Microdata Information System, MISSY, provides online access to data-related information essential for the analysis of official microdata. MISSY 3 is currently under development at GESIS, and aiming at supporting official surveys like EU-SILC and EU-LFS.
* Rogatus QMMS: Rogatus is an open source toolset currently in development at DIPF with support of GESIS, TBA21 and OPIT. Rogatus consists of different DDI compliant applications (e.g. Questionnaire Builder, Translation Builder, Metadata Builder, Rogatus Portal) to support a multitude of survey processes.

# 3 Stakeholders

The ESS and other DASISH projects are expected to form the core user group of the QVDB, as the aim of the tool, as expressed above, is to service these projects in their internal business processes.

Second, the QVDB is aiming at serving the research community, like researchers and students, in exploring metadata from existing projects or in designing new research.

Third, projects and institutions external to the DASISH, that collect, archive and distribute data from the social (and possibly also the humanist) sciences, should be given access to the services of the QVDB for similar purposes as those of the DASISH projects.

Further, while interoperability is an aim of the system architecture applying to the three tools of WP3, task 2, administrators and developers of tools and systems external to these tasks may seek interoperability between their systems and the QVDB.

At last, projects or institutions developing tools similar to the QVDB could be interested to reuse the QVDB model as a whole, or particular modules or layers.

## ESS and other DASISH projects

Stakeholders:

The core user group of the QVDB is expected to consist of the ESS and other DASISH projects.

Usages:

The main usages of the QVDB by the core user group are expected to be use in internal business processes, as well as its storage function.

### 3.1.1. Business processes:

The use of the QVDB in the ESS and other DASISH projects is closely related to their internal business processes. A generalized view of the steps and sub-steps in longitudinal data collection can be found in the Generic Longitudinal Business Process Model as presented in Barkow et al 2013, p.7.

In **designing** new research, content from the QVDB can be used in the instrument development, by export of questionnaire related metadata elements to questionnaire development software and translation tools. At this stage content from the QVDB can also be used in specifying data elements, e.g. in designing and redesigning variables and coding standards, outputting command setups for statistical packages, producing textual variable specifications like data protocols etc. The QVDB could also work as a tool for project managers and administrators to get control over the content and versions of metadata elements like questions and variables that have been used in their project over time.

While **building** instruments, content from the QVDB could be exported via web-services to questionnaire delivery tools that understand DDI, e.g. CAPI tools and Web-survey tools. Content from the QVDB could also be used to output PAPI questionnaires, export to coding tools etc..

Software-specific generation instructions for the construction of derived variables produced at the **data processing** stage and included in the QVDB can be used in reports, or reused in later/other data collections, exported to harmonisation tools etc..

While **preparing dissemination products**, variable definitions for the data files for publication, enhanced reports etc. could be outputted from the QVDB.

At the **dissemination** stage, metadata elements from the QVDB could be exported to data analyses, browsing and download systems etc.

DDI-L support for generating enhanced publications based on content from the QVDB can be useful at the **evaluation** stages prior to- and after a data collection, while preparing **dissemination** reports for publication, for validating instruments etc.

### 3.1.2 Storage:

The storage function and structure of the QVDB makes up the basis for the above mention usages. To maximise their usage and re-usage possibilities, metadata elements should as far as possible be stored as detailed study-independent metadata components, that can be reused by many, and for a variety of purposes. To achieve this, the storage structure should be based on the DDI-Lifecycle model, where a set of the available components will be used.

The multilingual metadata storage function of the QVDB will also serve as a memory for cross-sectional and longitudinal projects, and will give users access to metadata from a variety of projects.

## Other social science projects and institutions that are collecting, archiving and/or disseminating data

Stakeholders:

Social science projects and institutions that are collecting, archiving and/or disseminating data, but that are not part of the DASISH project.

Usages:

The usages by social science projects and institutions that are not part of the DASISH project, are expected to be the same as those described at 3.1.1, that is use in internal business processes and storage.

## Networks of projects within DASISH

Stakeholders:

Potential usage could also include services related to the business processes of networks of projects within DASISH.

Usages:

Serve internal business processes of networks of projects within DASISH, like possible metadata harvesting by CESSDA and interoperability with possible CESSDA systems (see 5.3).

## Users within the research community (researchers and students)

Stakeholders:

Researchers (incl. Principal Investigators) and students.

Usages:

### 3.4.1 Support of research based on data:

Users within the research community, like researchers and students are expected to be interested in browsing the content of the database, in order to explore the metadata related to particular projects and datasets.

### 3.4.2 Support of research design:

Users within the research community, like researchers and students are expected to be interested in finding and accessing metadata in order to design new research.

Create enhanced reports:

Researchers and students are also expected to be interested in creating enhanced reports based on the content of the QVDB, e.g. for evaluation purposes.

## Administrators and developers of other tools and systems

Stakeholders:

Administrators and developers of other tools and systems may seek operability between their system and the QVDB.

Usages:

### 3.5.1 Import/request of content from the QVDB, export of content to QVDB from external systems:

Administrators and developers of other tools and systems are expected to be interested in importing content from the QVDB to their system, or in exporting material from their system to the QVDB.

### 3.5.2 Full or partial reuse of database model:

Administrators and developers of other tools and systems are expected to be interested in reusing the full database model, or parts of the model, or reuse all or some of the components of the tool.

# 4 Proposed system

## 4.1 Overview

The system to be implemented has to fulfill a number of functional and non-functional requirements deriving from the needs of the stakeholder groups, that is, the intended users of the system as much as the persons/institutions responsible for creating, developing, reviewing, and hosting the system.

## 4.2 Functional requirements

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Category** | | **Requirement** | **Priority[[2]](#footnote-2)** | **Applicability to other tools** |
| **Support internal business processes of ESS; other DASISH projects; projects and institutions that are not part of DASISH** | | Storage of concepts, questions, variables, coding standards and other metadata components structured in a standardised system at a high level of granularity (DDI-Lifecycle) | Must have | QDDT TMT (selected elements) |
| Storage of multilingual content, structured in separate languages | Must have | TMT, QDDT |
| Storage of study-independent metadata components | Must have | QDDT, TMT |
| Support for ID and versioning | Must have | QDDT, TMT |
| DDI Resource Packages | Must have | QDDT, TMT |
| Support import of machine readable content | Must have | QDDT, TMT |
| Interface for manual entry of content/registration | Must have | QDDT, TMT |
| Support for questionnaire delivery (PAPI, CAPI, Web-survey) | Should have | TMT, QDDT? |
| Support for user-defined enhanced publications | Should have | QDDT, TMT? |
| Support of export of command setups for statistical packages | Should have |  |
| Support comparison of metadata elements like question lists, concept lists, questions, concepts etc.. | Should have | QDDT, TMT? |
| **Support of the research community (researchers and students)** | Boolean field level search, with possibilities of combining words and phrases using Boolean operators to limit, widen, or define the search | Must have | QDDT, TMT? |
| Boolean field level search in all elements and languages of the database | Must have | QDDT, TMT? |
| Download and print of search results in human readable formats (pdf, others) | Must have | QDDT, TMT? |
| Programme access to search results in machine readable formats (DDI-XML, DDI-RDF) | Must have | QDDT, TMT |
| Support for user-defined enhanced publication | Should have | QDDT, TMT? |
| Support autocomplete as well as truncation in search | Should have | QDDT, TMT? |
| Possibility to search with keyword/thesaurus terms | Explore | QDDT |
| **DDI Export/import possibilities** | DDI-Lifecycle and DDI-Codebook import from- and export to other tools/web-services | Must have | QDDT, TMT |
| Communication with Questionnaire Design and Development Tool (QDDT) | Must have | TMT |
| Communication with Translation management tool (TMT) | Must have | QDDT |
| Communication with SQP | Explore possibilities | QDDT, TMT |
| Communication with other DASISH tools, e.g. occupation coding tool | Explore possibilities on how content possibly could be reused in QVDB | QDDT |

## Nonfunctional requirements[[3]](#footnote-3)

4.3.0 Data model / Communication

|  |  |  |
| --- | --- | --- |
| **Reusable database model** | Technical specification and documentation of system | Must have |
| Modular system | Must have |
| Core module based on core DDI components | Must have |
| DDI profile | Must have |

### 4.3.1 Usability

The system should have satisfactory level of usability both for end user that wish to search and browse the database, and for users that work with the tool, i.e. users responsible for maintaining the metadata stored in the system.

Any users familiar with web applications should be able to use the system.

**4.3.2 Reliability**

The system should be able to support the expected number of concurrent users. Furthermore, since it’s an online system, it is always available. The system may be taken down when maintenance is needed on the server. This should be done infrequently. Any problems will be dealt with within normal office hours.

**4.3.3 Performance**

Response times should be acceptable. Operations performed daily, like searching and browsing the database, and by users working with the tool, should be as fast as possible. Infrequent or complex requests should have adequate response times.

The system should be able to handle the expected number of users.

**4.3.4 Supportability**

It’s an online system. No software has to be installed on the client side (other than a web browser).

Documentation will be created on how to install the system (on a server). The software is open source.

**4.3.5 Security**

The system must allow different access rights to different types of users. Access to browse and search the database could be freely available. Users that update any data/metadata should be required to log in with username and password. Depending on the security needs the system may or may not run on a secure web server (i.e. the communication between client and server is encrypted).

# 5 System models

## Example Use Cases[[4]](#footnote-4)[[5]](#footnote-5)

|  |  |  |  |
| --- | --- | --- | --- |
| **Category** | **Scenario** | **Use Case** | **Actor: Person or program** |
| **Support internal business processes of ESS; other DASISH projects; projects and institutions outside DASISH** | QDT Head/convenor wishes to design a new version of a repeat module of a questionnaire that has been fielded in an earlier round or wave of the survey | **Export module structured in detailed DDI from QVDB, import to QDDT** | Programs talk to each other |
| QDT Head/convenor wishes to design a new questionnaire | **Find relevant questions/questionnaire related metadata elements** | Person interacts with system |
| **Export of study-independent metadata components from the QVDB, import to QDDT (e.g. of questions without question names/numbers)** | Programs talk to each other |
| A translator wishes to view other translations of the same question in a particular language with the purpose of verifying a translation or reuse of an existing translation | **Find translations of the question in the chosen language in the QVDB** | Person interacts with system |
| Print a report of the searches | Person interacts with system |
| **Export of any selected metadata elements for reuse into the translation tool** | Programs talk to each other |
| A project manager wishes to see how particular variables and questions have developed over time in a survey | **Find variables and questions fielded in more than round or wave of the survey** | Person interacts with system |
| **View comparisons made between different versions of the metadata element, e.g. variables** | Person interacts with system |
| Print a report of the searches | Person interacts with system |
| A data archive team is making new variable names and labels and variable representations to be used for the required data deliverables for a survey. | **Enter study-dependent variable related information manually into the QVDB** | Person interacts with system |
| **Enter study-independent information related to the variable manually into the QVDB** | Person interacts with system |
| A metadata manager wishes to link the variable information to other metadata elements like questions and concepts | **Establish references to other metadata elements like questions , concepts etc.** | Person interacts with system |
| A data archive team wishes to reuse variable information created at earlier stages in the work process later in the work processes of the survey i.e. for use in data protocols, as variable specifications for statistical packages etc. | Create and print a user defined enhanced report with the selected variable information from the QVDB | Person interacts with system |
| Variable information structured in detailed DDI is exported from the QVDB as command setups for statistical packages | Programs talk to each other |
| A QVDB content manager wishes to import a final questionnaire module, structured in detailed DDI elements from the QDDT for archiving | **Find final module** | Person interacts with system |
| **Import module structured in detailed DDI to the QVDB, export from the QDDT** | Programs talk to each other |
| A QVDB content manager wishes to import a final questionnaire module translated into a particular language for a particular country and structured in detailed DDI elements from the translation tool for archiving | **Find translated module** | Person interacts with system |
| **Import translated module structured in detailed DDI to the QVDB, export from translation tool** | Programs talk to each other |
| A QVDB content manager wishes to import questions and variables from a DDI-C or DDI-L based question data base with an export function | **Find questions and variables in the database** | Person interacts with system |
| **Import DDI-XML instance or fragment to QVDB from the data base and make edits if necessary** | Programs talk to each other |
| A project manager wishes to import questions in multiple languages into a different system than the QVDB for reuse | **Find questions in multiple languages** | Person interacts with system |
| **Export of DDI-XML instance or fragment from the QVDB** | Programs talk to each other |
| **Support of the research community (researchers and students)** | A researcher wishes to make an evaluation of the generated variables of a survey. He wishes to view all generated variables, as well as the syntaxes used to produce them | Find generated variables and the syntaxes they are based on | Person interacts with system |
| Make PDF of search results | Person interacts with system |
| A student wishes to see which questions, variables and concepts have been used in a particular survey | **Find questions, variables and concepts used in study, e.g. question lists, variable lists, concept lists, as well as information on single questions, variables and concepts and relations between them** | Person interacts with system |
| Make PDF of search results | Person interacts with system |
| A methodologist wishes to see how repeated metadata components in a survey, like questions and variables have developed over time | **Find all versions of the metadata element of interest (question; variable)** | Programs talk to each other |
| Views any comparisons made between different versions of the metadata element, e.g. variables | Person interacts with system |
| Make PDF of search results | Person interacts with system |
| A researcher wishes to explore the coding standards used in a survey | **Find and view the summary of DDI Resource Package** | Person interacts with system |
| **Find Resource Package used in survey** | Person interacts with system |
| Create user defined enhanced report based on the content of the RP | Person interacts with system |
| A researcher wishes to import questions in multiple languages into a different system than the QVDB for reuse | **Find questions in multiple languages** | Person interacts with system |
| **Export DDI-XML instance or fragment from QVDB to different system** | Programs talk to each other |
|  | A researcher wishes to import questions in multiple languages into a questionnaire delivery tool | Questions in multiple languages structured in detailed DDI is exported from the QVDB to questionnaire delivery tool | Programs talk to each other |

## 5.2 System architecture[[6]](#footnote-6)

The task 3.2 team will develop three independent web applications, a questionnaire design and development tool (QDDT), a translation management tool (TMT) and a question, concept and variable data base (QVDB).

### 5.2.1 A common metadata model for the three tools

The three independent tools will be designed to communicate with each other in an efficient way. Therefore a common metadata model is needed. Work is on-going with the aim of specifying a common metadata model for the three tools, which will provide the backbone for exchanging metadata between the three tools via web services. The possibility and relevance of exchange of metadata components between the applications of task 3.2 and the SQP system will also be explored.

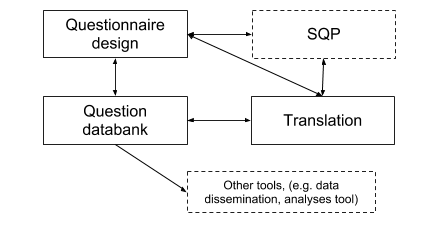
A core aim of the task 3.2 tools is to serve surveys at different stages in their business processes. Communication between the three tools will allow for metadata components developed at earlier stages of the work processes to be reused at later stages.

At a very general level, one could say that the three tools cover the work processes of questionnaire design and development, translation as well as archiving and providing metadata for reuse. The QDDT would typically serve questionnaire development related issues, while the TMT would serve the translation process. The QVDB is planned as a multipurpose archiving tool, mainly useable in almost stages of the GLBPM[[7]](#footnote-7), with the possibility to serve a variety of tools and web-services.

The aim is to achieve a common metadata model which enables the metadata flow between the three tools. Additionally the common metadata model could be used internally by the tools. Then it would be possible that software components could be reused by the three tools.

To achieve the common metadata model, the following issues need to be resolved:

1. Which metadata elements will be used in the transfer between the three tools;
2. mapping between the metadata elements and the DDI;
3. the direction for the flow of metadata elements between the three tools, as well as the steps in the work processes at which metadata components are exchanged;
4. administrative ownership of metadata;
5. a common identification and versioning system, including a versioning policy;
6. how the exchange of DDI metadata takes place, which type of DDI instances or fragments that will be transported, and which type of web-service will be used.



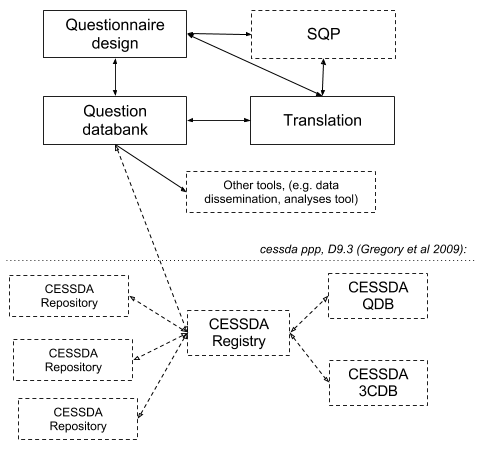
*Figure 1: Task 3.2 possible architecture and its relationship with other tools*

## Connections with other systems[[8]](#footnote-8)

The QVDB model will be DDI-based and the system will provide DDI-Lifecycle and DDI-Codebook export and import functionalities. The aim is to construct an application suitable for integration in a CESSDA network, as described by Gregory et al 2009 for WP9 of the CESSDA PPP. Both the QDDT and the TMT are regarded as important content providors for the QVDB.

The CESSDA network is described as a set of distributed applications, where the QVDB of task 3.2.c could act as one of several CESSDA repositories. The idea of the CESSDA network is to create a central registry that would facilitate search across surveys. The registry that is described in Gregory et. al. 2009 is, however, not a task 3.2 deliverable.

Figure 2 describes the planned task 3.2 applications and their possible relationship to SQP and a possible CESSDA network.



*Figure 2: Task 3.2 possible architecture and its relationship with a possible CESSDA network*

The system should also be constructed with the aim of serving projects and institutions within the social science community that are not in the DASISH project.

# References:

Barkow, I. et al., ‘Generic Longitudinal Business Process Model, Documenting the Helix’ *Data Documentation Initiative* *DDI Working Paper Series – Longitudinal Best Practice, No. 5* [Online] Available: <http://dx.doi.org/10.3886/DDILongitudinal05> [02 May 2013]

Gregor**y, A.** et al. (2009) Technical Specifications for a European Question Data Bank [Online] Available: <http://www.cessda.org/project/deliverables.html> [15 Oct 2012]

Description of Work - Data Service Infrastructure for the Social Sciences and Humanities (DASISH) 2011-08-15

1. For a generalized view of steps in longitudinal data collection, see <http://dx.doi.org/10.3886/DDILongitudinal05> , p. 7. [↑](#footnote-ref-1)
2. Ranking of priorities: Must (top priority) – Should – Could – Would (lowest priority). [↑](#footnote-ref-2)
3. Content from this section is planned to be included in a common requirement document for the three tools. [↑](#footnote-ref-3)
4. Most important features marked in bold [↑](#footnote-ref-4)
5. For ranking of priorities of functionalities, please see point 4.2 [↑](#footnote-ref-5)
6. Content from this section is planned to be included in a common requirement document for the three tools. [↑](#footnote-ref-6)
7. Generic Longitudinal Business Process Model (Barkow et. al 2013) [↑](#footnote-ref-7)
8. Content from this section is planned to be included in a common requirement document for the three tools. [↑](#footnote-ref-8)