IMI2 821520 – ConcePTION

ConcePTION

WP7 Information and data governance, ethics, technology, data catalogue and quality support

D7.9 Test report for FAIR data catalogue (1st)

Lead contributor	Morris Swertz & Eleanor Hyde, University Medical Centre Groningen (17 – UMCG)
	m.a.swertz@rug.nl
Other contributors	Marianne Cunnington, GlaxoSmitKline (41 – GSK) Rosa Gini, Agenzia Regionale di Sanità (10 – ARS) della Toscana

Document History

Version	Date	Description
V0.1	19 Oct 2020	Draft 1
V0.2	23 Oct 2020	Draft 2 – author review
V0.3	28 Oct 2020	Draft 3 – WP7 review
V0.4	06 Nov 2020	Draft 4 – Managing Board Review
V1.0	22 Nov 2020	Final version

Abstract

The objective of this report is to summarize initial user assessments of the first prototype of the ConcePTION FAIR data catalogue (see Deliverable 7.6 for description). The first prototype focuses on presentation of meta-data descriptions of population-based healthcare and surveillance datasources. It also captures details of tools and documentation used to allow data harmonization across these datasources.

A questionnaire-based tool was developed to allow Data Access Providers (DAPs) to directly describe characteristics (meta level data) of healthcare and surveillance datasources (D7.6). This aimed to allow researchers using the ConcePTION network to better understand the datasources available and able to answer their research questions and study needs. This questionnaire tool was deployed to 20 DAPs within the ConcePTION project. A follow up user survey was then deployed to a subset of DAPs covering different countries and data types to understand how this tool could be further streamlined and optimised. Significant overlap was identified between the questionnaire tool and the process rolled out to DAPs to design the ConcePTION Common Data Model (CDM) and to support data harmonization efforts. The second catalogue prototype will therefore seek simplifications in the questionnaire tool with the addition of automated meta-data collection where possible.

To enhance transparency of the design of the ConcePTION CDM and data harmonization/ transformation steps; detailed datasource information was captured and catalogued for the same 20 DAPs. Informatic requirements to enhance presentation and querying of this information were sought through interviews with WP7 leads. These requirements were assessed against the existing catalogue software tools originally developed for another European project (LifeCYCLE), which form the basis of the ConcePTION catalogue. These tools will be modified for the second catalogue prototype to accommodate these new requirements from ConcePTION. Modifications will be designed to be used by other consortia to increase the sustainability of the final catalogue.

Introduction

The objective of the ConcePTION data catalogue is to allow researchers to identify relevant Data Access Providers (DAPs) and datasources for the conduct of studies to generate evidence on the safety of medicines used in pregnancy within the ConcePTION network. It also aims to enable ConcePTION partners to manage and transparently share meta-data on the individual data items and their harmonisations as used in ConcePTION studies. These elements will be delivered according to FAIR Findable, Assessible, Interoperable and Reuseable principles

The catalogue will consist of four main areas hosting different data types and fulfilling different needs:

1. Meta-data describing the DAP Organisation, the datasource(s) the DAP has access to, and the individual data items contained within datasources. Population-based healthcare or surveillance data will be captured in a separate section from pregnancy report datasources. As described in Deliverable 7.6, meta-data capture in the initial catalogue prototype was based on a specially designed survey/questionnaire tool. Researchers will be able to query the meta-data.
2. For ConcePTION studies based on population-based healthcare or surveillance datasources, data are harmonised in a two-step process; first syntactically against the ConcePTION Common Data Model (CDM) and then on a study-by-study basis into a semantically harmonised set of study variables. A second part of the catalogue will house the data dictionaries of the original data, the Extract, Transform and Load (ETL) process and scripts for transformation into the CDM and the study-based decisions for semantic harmonisation.
3. The Catalogue will host the results of the Data Characterisation with the results organized to enable querying by researchers to understand the detailed features of the data whose description is contained in the Catalogue.
4. Finally, the Catalogue will host a negotiation service for researchers to engage consortiumpartners and DAPs.

The first catalogue prototype focused on section 1) of the catalogue capturing population-based healthcare and surveillance datasources. This report describes the user test of the Catalogue for ConcePTION, built using MOLGENIS tools and software. For the purpose of this report, the users are the DAPs who provided information on their datasources. Future prototypes and user tests will focus on the end users who are researchers wanting to conduct research studies through the ConcePTION network.

In addition, new user requirements related to section 2 of the Catalogue were charted specifically with regard to describing and being able to query the 'common data elements' in the Catalogue, i.e., the ability to capture common data elements for pooled statistical analyses across datasources, and the description of mapping from the original datasources onto these common data elements. For further context please see 'Template of ETL Specification' in ConcePTION Deliverable (D) 7.5. In particular, it was assessed whether ConcePTION could benefit from an existing tool developed and currently in use within the Lifecycle project (MOLGENIS software catalogue module, https://MOLGENIS88.gcc.rug.nl/menu/main/app-MOLGENIS-app-lifecycle). Detailed requirements were captured in interviews between Rosa Gini, Morris Swertz and Eleanor Hyde in July, August and September 2020. Some requirements were missing and the potential for MOLGENIS software to add new, desired functionality was assessed using a prototype as the basis for the next version of the Catalogue.

This report first discusses evaluations and feedback of the meta-data section of the Catalogue and separately discusses new user requirements captured during the period since D7.6 to be incorporated in the second Catalogue prototype (to be described in D7.7).

Meta-Data Catalogue Section

Methods

The initial ConcePTION FAIR Data Catalogue prototype aimed to sustainably capture key meta-data for population-based datasources and pregnancy report datasources and to allow researchers to query and identify datasources to answer specific study questions on the safety of medications used in pregnant and breastfeeding women. Due to the heterogeneous nature and structure of the datasources, the initial Catalogue prototype focused on a meta-data collection solution based on a specifically designed questionnaire tool that could be sent to DAPs for their completion (see D7.6 for questionnaire and prototype description). The questionnaire for population-based data was designed according to specifications outlined by Work Package One (WP1 see D7.1) and DAP self-completion was favored to increase accuracy of data capture as well as possibility of updated information capture by sustainable means.

The questionnaire data collection tool was initially tested with 20 WP1-aligned-DAPs with access to population-based healthcare or surveillance datasources (See Appendix 1). A link to the questionnaire tool was sent to DAPs through the Task Management System (TMS - see D7.5). A set of instructions for questionnaire access and completion, as well as a key set of definitions related to terms used in the questionnaire (Source population, datasource population, datasource) to drive consistency in meta-data capture, was made available through the TMS to DAPs.

Following questionnaire completion, a user survey was sent to a convenience sample of DAPs chosen to represent a range of data types (electronic healthcare data and population-based congenital anomaly registries) and countries. The user survey aimed to identify areas for further improvement as well as assessing user experience which would enable an assessment of sustainability of the data capture approach. In this case the catalogue user was defined as the DAP providing meta-data. The questions from the user survey are described below. Some focused on basic usability (questions 1-4). However, some (questions 5-7) specifically probed overlap with the process that had been put in place to design the ConcePTION CDM and further developed as part of the pipeline for data transformation. This involved the collection of data dictionaries from DAPs as

well as interviews with the DAPs covering the datasource structure and content.

1. How long did it take you to complete the questionnaire?
2. How did you find the experience of logging into the questionnaire? Any particular challenges?
3. How did you find navigating through the questionnaire? Any particular challenges?
4. Were the definitions of source data population, datasource and datasource population useful guides for answering the questions?
5. Were there key areas of overlap with the DAP interview?
6. Which details of the questionnaires were not captured in the DAP interview?
7. Could some parts of the questionnaire be condensed?
8. Any other comments?

Results

The questionnaire was marked complete within the system by 11 out of 20 DAPs within the timeframe allotted for the task. Five of the DAPs who completed the survey were additionally contacted and asked to complete the short feedback survey detailed above to understand strengths and limitations of the initial catalogue approach. The survey results are described in Table 1.

Table 1: Overview of User Survey on Meta-Data Collection from DAPs

Question	CPRD	Tuscany Registry of Congenital Defects-RTDC	EFEMERIS/ CHUT	FISABIO	Malformation Monitoring Centre Saxony-Anhalt
Time to complete questionnaire	90-120 mins	180 mins	120 min	240 mins.	120-180 min
Experience completing questionnaire	Logging in was clear, given the instruction sheet,	Logging in was clear.	OK	Logging in was clear and easy through the website. The instruction sheet was really clear and useful	For logging in it was helpful to have the instruction sheet! Then it was clear and straightforward.
Navigation through the questionnaire	Hard to gauge how long it would take to complete because it wasn't possible to navigate between pages once a survey section was completed	Should be allowed to navigate provisionally to the next pages without having completed mandatory fields. Useful to be able to download the pdf or Excel version of the questionnaire before navigating.	Would have been useful to navigate between pages	Maybe the structure of the different parts was not clear and not easy to know how long it would take to fill	No particular challenges
Was the definition guide useful?	Yes	Yes	Not necessary	Yes	Yes
Areas of overlap with the DAP interview?	Timeframe from application for data to availability Variables section First half of the provenance of data Parts of healthcare settings and codes	The questionnaire could be partially completed using many of the answers already given during the DAP interview, e.g. what data are available.	Timeframe from application for data to availability Variables section First half of the provenance of data Parts of healthcare settings and codes	Yes	DAP organization information (but is needed) and variables

Details of questionnaire not captured in the DAP interview	Governance and self-report for meds/diagnoses	The category 'infants <1 year' was not reported separately but embedded in the category 'infants with toddlers 28 days-23 months'. This category is important for the registries of congenital anomalies that mainly cover the first year of age.	None	All questions could be answered through the information filled in the FISABIO's interview answer sheet	Yes, self-report for meds/diagnoses
Possible to condense some parts of the questionnaire	Some of the data elements could be put together into a longer list with a single question	Some of the data elements could be put together into a longer list with a single question.	None	Some of the data elements could be put together into a longer list with a single question	No, I would not put these different question elements together.
Other comments?	Despite overlap, this format may be easier to screen compared with the data dictionary	None	A lot of overlap with the different data we have already provided before	Maybe the format is easy to screen, but the interview answer sheet contains more specific information regarding the characteristics in each DAP.	None

The overall questionnaire experience was positive: The DAPs surveyed reported positively on the questionnaire tool layout, instructions and navigation. Some small improvements, such as ability to navigate back and forth between pages of the questionnaire, were noted. It was clear from the feedback that there was significant overlap between the meta-data collection questionnaire and the data dictionary collection and DAP interviews related to the data transformation steps (CDM and ETL design and execution). In addition, the meta-data questionnaire took between 90 and 240 minutes to complete with longer times driven by additional checks in the local language which is unlikely to be compatible with sustainable means of meta-data collection across a large number of DAPs, the majority not directly involved with the ConcePTION project (See D1.1).

Discussion

These findings were discussed with WP7 Leads and Task 7.4 Leads, the latter with significant experience across different catalogue models including BBMRI Biobank Directory. This led to several recommendations for the meta-data section of the ConcePTION FAIR data catalogue:

- Simplify the meta-data collection focusing on organisational details and a few key datasource characteristics essential for pregnancy studies e.g. availability of mother-child link. This simplification is likely to lead to a more sustainable catalogue model that can be maintained across many potential DAPs. WP1 will be consulted to identify key datasource characteristics during the design phase for the second prototype of the catalogue.
- Explore automated options for the meta-data collection including automated publication searches based on organisation names and key words. This option may further reduce the burden on the DAP while enabling an efficient means of maintaining current information on potential datasources.
- Consolidate the detailed specifications for meta-data collection with the process developed through data dictionary collection and DAP interviews.

Datasource Documentation, CDM and Data Transformation Catalogue Section

Methods

To enable the design of the ConcePTION CDM a consistent stepwise process was put in place and tested with the same 20 population-based healthcare and surveillance DAPs (See D7.5). This involved collection of data dictionaries from DAPs, extraction of information from the data dictionaries according to a standardized framework and interviews with DAPs to confirm accurate representation of their data. The collection of detailed information, and desire for transparency across the data transformation pipeline (original data via ETL to ConcePTION CDM) led to discussions around additional requirements for the CDM documentation part of the catalogue. These requirements were collected through extensive virtual interviews between Eleanor Hyde and Morris Swertz (UMCG) and Rosa Gini (ARS, WP7 co-lead). These requirements were captured through a shared spreadsheet (presented in the Results section) which was also used to capture technical questions. A detailed prototype of the part of the catalogue housing and visualising the CDM documentation was created by Morris Swertz and evaluated against original CDM documentation (primarily Word documents) provided by Rosa Gini to assess which requirements could be implemented against existing MOLGENIS software tools being deployed for the ConcePTION catalogue versus those requirements that would need modifications to existing tools or even design of bespoke tools.

Results

Spreadsheet overview of requirements

The results from the interviews were captured in a spreadsheet summarised in table form (Note: the full detailed spreadsheet is available on request). We used a 'traffic lights' system to indicate whether a requirement could be implemented using existing tools within the MOLGENIS software/catalogue, and also provided an estimate of implementation complexity.

'Traffic light' key:

	Possible, can require extra work
	Possible, but some further detailed specification will be required before development
	Not possible (or advisable) in MOLGENIS

High-level requirement	Traffic light	Implementation complexity
4 tabs on the home page: 1) CDM tab 2) Datasources tab 3) Study variables tab 4) ETL tab		Small
CDM tab: Left: CDM tree Right: High-level description of the CDM		Small
CDM tab: Versioning of CDM		Medium to major, depending on granularity of the versioning. This has not been decided. In practice we assume that CDM will not change but will only extend thus limiting version issues in practice. This would define new version as old version + additions.
CDM tab: Drill down to groups of tables (data banks)		Add topical groups to tables
CDM tab: Drill down to a single table		Medium size modifications needed
CDM tab for a single table: Left: The columns in the table are displayed Right: Description of the header level		Existing functionality
CDM tab for a single column in a single table: Left: Selected column Right: The 'conventions' for that variable		Medium size modifications needed

Conventions: Variable name (just a short string) Mandatory (yes/no/conditional statements) Description (short box of text) Format (character; character yyyymmdd; int; float) Vocabulary (this is complex) Comments (longer box of text) Example1 Example2		
Datasources tab: Left: List of trees (1 tree per datasource)		Medium size modifications needed
Datasources tab for a single DAP: Left: list of data banks in the datasource Right: Description of the datasource		Medium size modifications needed
Datasources tab for a single data bank in a datasource: Left: The tables in the data bank are displayed Right: Answers to 5 questions: 1) What triggers the creation of a record of the table? 2) Is the table collected for all the population of your database, or only for a subpopulation? 3) Can you comment on the completeness and quality of the table? If you don't have formal measurements, feel free to convey the assumptions you commonly make 4) What is the time span of the table, how often it is refreshed, and which is the lag time between the data creation and the time when the data has the potential of being available to your organization? 5) Include other comments you may want to share about this table		Medium size modifications needed. Probably needs a few iterations with the user group to get right.
Datasources tab for a single column in a single table in a single data bank in a datasource: Left: selected column Right: Original name (short string) Meaning (longer string) Vocabulary in English (see comments) Comment (long string)		Minor changes needed
Versioning of DAP		Medium to major depending on granularity and interaction with CDM versioning. In practice we assume that DAP will not change, or only extend information thus limiting version issues in practice.
Study variables tab and underlying functionality		Needs further detailed specification

ETL tab to report details on the ETL applied. Can create, edit ETL online. Can indicate which tables of data provider where joined/filtered and how records are created to feed the CDM table, and can indicate mapping of variables from this query to CDM (either selecting a column, or applying a conversion rule); this will mirror the content of a 'specification table' from the ETL template (see Deliverable 7.5)		Requires major refactoring as compared to LifeCycle use case because the transformation has to be described to incorporate the level of 'Table' (instead of on variables, as in LifeCycle).
ETL status		Can indicate if a ETL is 'draft' or 'final'.
Underlying functionality to execute ETL		Would require development of new tool. It was therefore decided to keep the execution of the ETL processes outside the catalogue. Catalogue will only report current state.
Study variables		

The CDM tab and its underlying required functionality coincide with existing MOLGENIS functionality in which a user's choice on the lefthand side in a menu tree determines the information displayed on the righthand side. The structure of the menu tree is also in line with MOLGENIS software functionality; 'grouping' of tables is, however, a potential issue if grouping is not to be hard-coded.

The study variables tab requires more detailed clarification.

The ETL tab requires transactional functionality with complex screen flow and interaction: Unfortunately this does not match the MOLGENIS user interface and it is inadvisable in the short term (and also for budget reasons) to custom-build to this extent within MOLGENIS. Therefore, it was decided that the ConcePTION catalogue will be limited to displaying the ETL results as provided.

Evaluation of requirements in prototype

To assess difficulty of implementation, UMCG created a small proof of concept of only the data structure changes needed to accommodate the CDM needs of the ConcePTION project. The prototype is only a data model into MOLGENIS which delivers (a) Excel file format for data import/export and (b) basic forms for editing/listing data. What has NOT been implemented are bespoke user interfaces to show different trees and tabs as listed in the requirements above, as these will require user interface changes. These will require significant work towards the next catalogue release but has been evaluated as feasible to implement.

The following features have been tested in this prototype:

- [DONE] Can register variable (aka 'data element','column', 'attribute')
- [DONE] Can group data element into a table (aka dataset, class, dataframe, …)
- [DONE] Can attach table + data element to a collection (aka 'data bank', 'cohort', 'study', 'registry', 'CDM', 'DAP')
- [DONE] Can indicate label (aka 'question' or 'measurement')
- [DONE] Variable names are unique within a collection or collection+table
- [DONE] Can indicate data format (aka type)
- [DONE] Can indicate codelist (or ontologies)
- [DONE] Can create tree for common data elements and individual data providers
- [DONE] Can indicate collection event (aka 'assessment')
- [DONE] Can deal with age repeated measures in wide format
- [DONE] Can create a tree of variables by topic(s)
- [DONE] Can create a tree of tables by topic(s) (aka sections)

- [DONE] Can see difference between a harmonized variable and 'raw' data provider collected(cohort/raw) collection
- [DONE] Can see more details on table (like role, constraints of table)
- [DONE] Can indicate if a variable is mandatory in a table
- [DONE] Can indicate how source table(s) should ETL to which CDM table
- [DONE] Can indicate how origin columns map to CDM table, given source table statement
- [DONE] Can see example data in origin column or harmonization target variable
- [DONE] Can indicate that a table is in LONG format (e.g. because of non-standard collection events such as a survey)
    - [DONE] Can indicate which column is participant id
    - [DONE] Can indicate which column indicates which event
- [DONE] Can define foreign key relations from one variable to another
- [DONE] Can see if an ETL is 'draft' or 'final'
- [DONE] Can indicate version on level of whole collection (i.e. for a DAP, CDM, cohort, study, etc)
- [DRAFT] History: Created, Updated, Change by person etc
- [DRAFT] Can see versions history of variables, variables collection, ETLs based on versioning on level of collection (i.e. CDM and DAP sources need to be copied completely to create a new version)
- [DRAFT] add information on the organisation behind the collection. In ConcePTION these are the DAPs
- [OUT OF SCOPE] Can give permission to individual users to edit/create specific variables, ETLs, etc.

Discussion

Given the complexity, it was decided that the 'LifeCycle' catalogue will be generalized based on ConcePTION use cases but in a way that the catalogue can also be used by other consortia (EU Longitools and EU Athlete have plans to also use this MOLGENIS catalogue system), thus increasing potential for future sustainability of this section of the catalogue.

Conclusion

For the second catalogue prototype (D7.7) several changes will be implemented:

- The meta-data section of the catalogue will be simplified with data collection focused on DAP information. Automations based around DAP publications will be explored. This aims to allow sustainable upscaling of the number of DAPs captured in the catalogue.
- More detailed information on the data held by DAPs will be available and will be focused in the CDM section of the catalogue. This section will be further developed based on the LifeCycle catalogue.

Appendix 1: Data Access Provider List

1) University of Oslo (UOSL)
2) University of Aarhus
3) University of Dundee
4) University of Ulster (ULST)
5) Centre Hospitalier Universitaire de Toulouse (CHUT)*
6) University of Bordeaux
7) University Medical Center Groningen (UMCG)
8) PHARMO Institute
9) Leibniz Institute for Prevention Research and Epidemiology (BIPS)
10) Fundacion para el Fomento de la Investigacion Sanitaria y Biomedica de la Comunitat Valenciana (FISABIO)
11) Foundation University Institute for Primary Health Care Research Jordi Gol I Gurina (IDIAPJGol)
12) Universita degli Studi di Ferrara (FERR)
13) Consiglio Nazionale delle Ricerche CNR Tuscany (CNR-IFC – Instituto Fisiologia Clinica)
14) Agenzia Regionale di Sanita della Toscana (ARS)
15) University of Messina
16) Malta Congenital Anomalies Registry
17) Malformation Monitoring Centre Saxony-Anhalt
18) National Institute for Health and Welfare, Finland
19) University of Swansea
20) GlaxoSmithKline