Deliverable 8.2: Overview Document on Already Existing Data Harmonisation Tools

Sebastian Mate[1], Ines Leb[1], Christina Schüttler[1], Kaisa Silander[2], Timo Miettinen[3], Juha Knuuttila[4], Niina Eklund[2], Prof. Dr. Hans-Ulrich Prokosch[1]

[1]Chair of Medical Informatics, Friedrich-Alexander-University Erlangen-Nuremberg, Germany
[2] Department of Health, National Institute for Health and Welfare, Helsinki, Finland
[3] Institute for Molecular Medicine in Finland, University of Helsinki, Helsinki, Finland
[4] Department of Information Services, National Institute for Health and Welfare, Helsinki, Finland

1 Preambel

The aim of this document is to analyse the state of the art and existing tools for data harmonisation, considering their applicability within CS-IT. In addition, it will include information on the different levels and versions of various standardised and non-standardised ontologies and data dictionaries and schemas. The major focus of this document is to present some examples of already ongoing harmonisation projects in the field of biobanking and biological sample collections, as well as aiming to support the project by providing a literature review on the biobanks' needs regarding the harmonisation tools and processes described in BBMRI-ERIC CS-IT's Requirements Specification document (Eklund 2016). Since data harmonisation, however, is not only an important issue in biobanking, but a key component in all data sharing/data projects, our focus is also on such projects outside the biobanking area. This review has been done as a deliverable of the Biobanking and Biomolecular Resources Research Infrastructure – European Research Infrastructure Consortium (BBMRI-ERIC) Common Service IT (CS-IT) as well as the Horizon2020 INFRADEV project ADOPT.

The current version is based on a comprehensive literature review and experiences from own work in this field as well as based on communication with colleagues in related interdisciplinary multi-partner research projects.

2 Introduction

The Biobanking and BioMolecular resources Research Infrastructure - European Research Infrastructure Consortium (BBMRI-ERIC) aims to establish, operate and develop a pan-European distributed research infrastructure in order to facilitate the access to biological resources as well as facilities and to support high quality biomolecular and biomedical research. The BBMRI-ERIC Common Service IT (CS-IT) has the mission to deliver expertise, services, and tools relevant to the pursuance of tasks and activities of BBMRI-ERIC. The project proposal is divided between eight work packages (WPs) that are presented in figure 1.

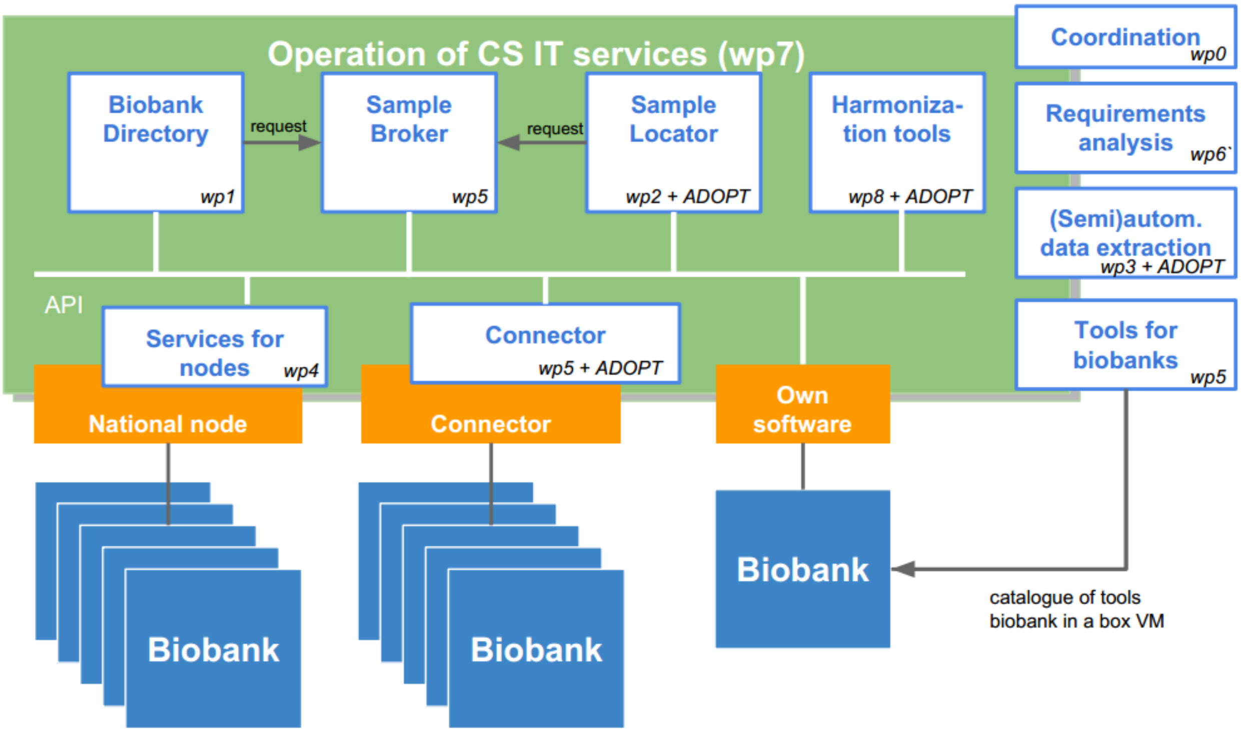

Figure 1: Overview of CS-IT services and work packages

The Directory (WP1) will enable interested parties to identify biobanks that potentially have samples and data of interest based on aggregate level information, but not at the level of samples. It will build upon stable components such as the MIABIS 2.0 standard. For authorised users, in contrast, the Sample Locator (WP2) will provide a controlled access environment to explore fine-grained information on a sample level, while complying with the general data protection regulation. Its architecture and implementation will support both a data cube model (warehouse) and federated queries to the national level databases and/or to the local biobanks. Personal private communication with biobanks is facilitated using the Negotiator (WP2) that will have workflows to support sample request scenarios. Such a communication is expected to run in repeated cycles to address the researcher's needs. To facilitate integration of all these systems, tools for the harmonisation of data will be provided (WP8 in collaboration with ADOPT).

In addition to the central services, BBMRI-ERIC will also offer diverse hosting services for national nodes and required interface tools (WP4), e.g. website hosting and operations of national node directory and simple interface for the Sample Locator, and offer diverse tools for use by biobanks (WP5), e.g. delivery of reference connectors for the biobanks to connect them to the national node (from ADOPT), a catalogue of tools and a ‘virtual machine’ with tools such as LIMS preinstalled. To ensure there is a substantial consumer base for CS-IT services, WP6 will provide a continuous monitoring and analysis of users’ needs and an analysis of usage and usability of existing services in order to propose shutting down or revamping services. All BBMRI-ERIC services will be operated by WP7, including supporting sustainability of the tools developed in related projects (e.g. enabling access to BiobankCloud, BBMRI-LPC, BioMedBridges, BioSHaRE/P3G, MIABIS and RD-connect resources) and providing basic collaborative tools for BBMRI-ERIC ecosystem, such as mailing lists, forums, and remote collaborative tools like video conferencing and web conferencing systems.

The service to be developed in WP8 relates to the work in ADOPT WP2, WP3 and related projects. The act of delivering data to the centralised Directory or the Sample Locator to form data availability statistics will necessitate the processing of requests over data derived from different biobanks, where a common language or ontology will be essential. For the purposes of data harmonisation, relevant terms and ontologies with respect to the attributes and value lists present in the data of a participating biobank will need to be agreed upon. Typically, local data items and value lists are fixed, but the tools specified within this service shall support mapping on an agreed upon common terminology. This includes, but is not limited to, artifacts such as biobank metadata characterisation, quality indicators for biosamples and terminologies/ontologies for diseases. The scope and depth of the ontology used to describe and classify these artifacts need to be defined first in terms of minimum and optional data sets. As reference knowledge, SPREC, Biobank lexicons and MIABIS 2.0 can be the starting points for this, and ontology mapping experiences from other European projects with similar challenges shall also be considered. After initial analysis, depending on where the realistic scope of the work is set, further focus could be made on providing translation specification for clinical information as well, available via HL7, DICOM, and other standards under the IHE umbrella and to utilise standards common in medicine such as ICD-10 and SNOMED CT. At the same time, advanced clinical formats such as openEHR (ISO 13606-2) will be considered, given their ability to describe and query structured and precise clinical information in an implementation-independent way.

Once initial specifications are laid out, the national nodes and biobanks could follow the specification in their harmonisation work, making for smoother delivery of data to the Directory and Sample Locator. The resulting common ontologies themselves should be described using standard technologies such as RDF and OWL.

In the following initiatives, projects and tools will be described, which are currently used in data sharing projects around the world. When details of the harmonisation process are available in the literature or at project websites, those are illustrated in this document. Even though some of the projects have not yet published details of their harmonisation concepts and tools, they are at least mentioned in this document, since it might be helpful to follow-up on their future work during the course of the CS-IT/ADOPT project.

3 Projects and Tools

3.1 OHDSI/OMOP

The Observational Health Data Sciences and Informatics project (OHDSI) is a multi-stakeholder, interdisciplinary collaboration to bring out the value of health data through large-scale analytics. All of the project solutions are open-source. This project builds on the definition of a Common Data Model (CDM) for comparative effectiveness in the Observational Medical Outcomes Partnership (OMOP) and has recently described that "the vision of creating accessible, reliable clinical evidence by accessing the clinical experience of hundreds of millions of patients across the globe is a reality" (Hripcsak 2015). The concept behind this approach is to transform data contained within disparate databases into a common format (data model). In OHDSI this is used to then perform systematic analyses using a library of standard analytic routines that have been written based on this common format. Such open-source analysis tools currently are ACHILLES (Automated Characterization of Health Information at Largescale Longitudinal Exploration System), HERMES (Health Entity Relationship and Metadata Exploration System), PLATO (Patient-Level Assessment of Treatment Outcomes, providing predictive models that assess probability of a patient experiencing any outcome following initiation of any intervention, given his or her personal medical history), HERACLES (Health Enterprise Resource and Care Learning Exploration System, helping to build and explore cohorts to assess a specific clinical population across a wide-variety of clinical dimensions), and HOMER (Health Outcomes and Medical Effectiveness Research, enabling risk identification and comparative effectiveness studies, with real-time exploration of the effects of medical products) (Hripcsak 2015).The OMOP CDM model has been successfully evaluated in the last four years (FitzHenry 2015, Voss 2015, Rijnbeek 2014, Matcho 2014, Makadia 2014, Zhou 2013, Overhage 2012) and also been replicated for six European EHR databases (Schuemie 2013). The OMOP CDM is currently available in its fifth version and its developers found that it enables harmonising disparate coding systems — with minimal information loss — to a standardised vocabulary.

To represent the relevant domains, the CDM contains 39 tables from which some major ones are briefly described in the following. A CONCEPT table contains records that uniquely identify each fundamental unit of meaning used to express clinical information. Such concepts are derived from source vocabularies, which represent clinical information across different domains (e.g. conditions, drugs, procedures) through the use of source codes and associated descriptions. A VOCABULARY table includes a list of the vocabularies collected from various sources or newly created by the OMOP community. Records in the Standardized Vocabularies tables are derived from national or international vocabularies such as SNOMED CT, RxNorm, and LOINC. Furthermore, custom concepts have been defined to cover various aspects of observational data analysis. The DOMAIN table includes a list of OMOP-defined domains where the concepts of the standardised vocabularies can belong to. Mapping between local institution-based source codes and OMOP concepts was based on the SOURCE_TO-CONCEPT_MAP table (a legacy data structure within the OMOP Common Data Model, recommended for use in ETL processes to maintain local source codes which are not available as Concepts in the Standardized Vocabularies, and to establish mappings for each source code into a Standard Concept as target_concept_ids that can be used to populate the Common Data Model tables). The SOURCE_TO_CONCEPT_MAP table however, is no longer populated with content within the Standardized Vocabularies published to the OMOP community. Instead the CONCEPT_RELATIONSHIP table is now used for this purpose. The CONCEPT_RELATIONSHIP table contains records that define direct relationships between any two concepts and the nature or type of the relationship. Each type of a relationship is defined in the RELATIONSHIP table (which provides a reference list of all types of relationships that can be used to associate any two concepts in the CONCEPT_RELATIONSHIP table). (http://www.ohdsi.org/web/wiki/doku.php?id=documentation:cdm:single-page)

3.2 OBiBa, Public Population Project in Genomics, and the Maelstrom Research programme

The following description captures a set of very closely related developments in which a set of tools around data harmonisation and distributed data analysis have been developed. For this purpose Maelstrom Research (https://www.maelstrom-research.org/) brings together an international team of epidemiologists, statisticians, and computer scientists to answer some of the challenges of cross-cohort research collaborations. Its activities are founded on previous projects, such as the DataSHaPER project, the OBiBa software application suite (http://www.obiba.org), and the federated data analysis methodology developed under the DataSHIELD project (http://www.datashield.org). In order to summarise those activities descriptions from the respective websites have been brought together in the subsequent text.

OBiBa is an international project committed to build open-source software for epidemiological studies and biobanks. OBiBa software is developed in close partnership with large-scale studies and supports the entire data management lifecycle including data collection, integration, harmonisation, sharing, and analysis. OBiBa was launched in 2007 as a core project of the Public Population Project in Genomics (P3G) (http://www.p3g.org/). The project grew rapidly and obtained substantial financial support from its main partners to build a new generation of open-source software for data collection and management (https://www.drupal.org/node/2565475). In 2012, OBiBa and the DataSHaPER project, also from P3G, joined their efforts to create the Maelstrom Research (https://www.maelstrom-research.org/) programme whose mission is to develop tools and methods for epidemiological data harmonisation (http://www.obiba.org/pages/about/). There are three general approaches to analysing harmonised data across collaborating studies: pooled data analysis (https://www.maelstrom-research.org/about-harmonization/environment-for-co-analysis#Pooled data analysis), summary data meta-analysis (https://www.maelstrom-research.org/about-harmonization/environment-for-co-analysis#Summary data meta-analysis), and federated data analysis (https://www.maelstrom-research.org/about-harmonization/environment-for-co-analysis#Federated data analysis). The first two approaches, pooling individual-level data in a central location and meta-analysing summary data from participating studies, are commonly used in multi-centre research projects. In addition, Maelstrom Research and its partners are proposing a new method for co-analysing harmonised data across multiple studies: performing federated analysis of geographically-dispersed datasets. DataSHIELD (www.datashield.org), the methodology developed to achieve this, essentially coordinates parallelized simultaneous analysis of the individual-level data hosted on geographically-dispersed servers. DataSHIELD acts as an interface module between Maelstrom's Opal software and the R statistical environment. The federated data analysis approach enables collaborating studies to participate in combined analyses in a secure, scalable and sustainable manner. Unlike data sharing initiatives based on central data deposition, the federated data analysis approach allows studies to remain in complete control of their data. Unlike meta-analysis of study-level estimates, the federated approach allows investigators to safely and remotely analyse data at their convenience and in real time avoiding the important delays related to waiting for each individual study to produce and provide the required summary statistics. Harmonisation involves achieving or improving comparability of similar measures collected by separate studies or databases for different individuals (https://www.maelstrom-research.org/about-harmonization/environment-for-co-analysis).

Some research programmes foster prospective implementation of harmonised measures to collect data across studies, while others turn their efforts to retrospective harmonisation and co-analysis of existing datasets. For data harmonisation the Maelstrom Research programme applies step-by-step guidelines which were developed by the DataSHaPER (Data Schema and Harmonization Platform for Epidemiological Research) team (https://www.maelstrom-research.org/about/data-harmonization). They aim to provide a means for investigators to ensure quality, reproducibility and transparency of the results of multi-centre research. Such guidelines describe a five-step process and are available at the Maelstrom Project web page (https://www.maelstrom-research.org/what-we-offer/guidelines). In its core harmonisation step the approach used to process data under a common format (https://www.maelstrom-research.org/what-we-offer/guidelines/data-processing-methods) will vary depending on the variables to be harmonised, the data collected by each study and the possibility to pool data. The Maelstrom Research team recommends algorithmic transformation, simple calibration model, standardisation model, latent variable model and multiple imputation models as possible approaches. Further on the Maelstrom website an overview of literature on general harmonisation guidelines and approaches as well as applied harmonisation publications (called methods library, see: https://www.maelstrom-research.org/repository/methods_library) are provided, but no software tools supporting such harmonisation processes are available. The OBiBa software tools comprise

1. Onyx, a web-based application that manages participant baseline interviews at assessment centres or clinics,
2. Opal, OBiBa's core data warehouse, providing all the necessary tools to import, transform and describe data,
3. DataSHIELD enabling advanced statistical analysis of individual-level data from several sources without actually pooling the data from these sources together,
4. Mica, a powerful software application used to create data web portals for large-scale epidemiological studies or multiple-study consortia and
5. Agate, a web application that offers users related services to the OBiBa software stack: user authentication, user profile management, user notifications.

In collaboration with the DataSHIELD team from the University of Bristol, the OBiBa team built the complex software infrastructure required to run securely DataSHIELD analyses on data stored in Opal (http://www.obiba.org/pages/products/opal).

3.3 i2b2 and SHRINE

Early U.S. developments of hospital-based research data repositories go back to the beginning of this century with the i2b2 toolkit being one of the most prominent developments (Murphy 2010). With this development (which even builds on very early work of Safran (Safran 1989)) Zak Kohane and his team have suggested to instrument the health care enterprise for discovery research in the genomic era by using EHR data to drive discovery in disease genomics (Kohane 2011, Murphy 2009). To extend such capabilities beyond single hospitals the i2b2 team has set up a Shared Research Informatics Network (SHRINE; see http://catalyst.harvard.edu/services/shrine/ for details) which enables the distribution of i2b2 queries across multiple instances, increasing the ways to leverage EHR data for research first with four other Harvard hospitals, and later in many other U.S. data sharing and comparative effectiveness research networks (McMurry 2013, Kohane 2012).

The SHRINE system is an extension to i2b2 that uses a modified query client and a highly customisable query distribution system, which is based on SPIN (Shared Pathology Informatics Network) (McMurry 2007), to distribute queries across multiple i2b2 sites. Results are then aggregated centrally in the system and presented to the investigator. To achieve semantic mappings, a mapping file stores 1:1 concept mappings, which are translated on-the-fly at query runtime.

Two exemplary projects, a pediatric rheumatology consortium (CARRAnet, a registry of patients across 60 institutions), and a pediatric inflammatory bowel consortium of at least 40 institutions, each represented by its own i2b2 repository and a connecting SHRINE systems illustrate the success and the scalability of the i2b2/SHRINE approach (Kohane 2012).

The Scalable Collaborative Infrastructure for a Learning Health System (SCILHS, pronounced “skills”) is a Clinical Data Research Network that participates in the Patient-Centered Outcomes Research Institute (PCORI) Network, PCORnet. SCILHS is comprised of 12 health centres across the United States that cover over 10 million patients. Each site uses i2b2 to store and analyse patient data for clinical research. Since PCORnet-affiliated networks are required to adopt the PCORnet Common Data Model (CDM; see http://www.pcornet.org/pcornet-common-data-model/ and below) the SCILHS team has developed a process to generate a PCORnet Common Data Model (CDM) physical database directly from existing i2b2 systems, thereby supporting PCORnet analytic queries without new ETL programming. This involved a formalised process for representing i2b2 information models (the specification of data types and formats), an information model that represents PCORnet CDM Version 1.0, and a programme that generates PCORnet CDM tables, driven by this information model. As an important result, this approach is generalisable to any logical information model (Klann 2016).

In todays healthcare and research environments it will be unavoidable that different projects or regulatory rules require the usage of different types of information models.Thus, transfering data elements from one information model to another, as e.g. described above for the i2b2 to PCORnet CDM transformation, will become a typical task for research institutions or research networks. Another example of such a development has recently been described in Mo (2016), who presented a decompositional approach to transfer querying algorithms defined for the Quality Data Model (an established standard for representing electronic clinical quality measures on EHR repositories) into messages/queries for i2b2-based data repositories.

3.4 PCORnet

A major U.S. funding initiative is the PCORnet initiative launched by the Patient-Centered Outcomes Research Institute (PCORI). PCORnet aims at establishing an effective and sustainable national research infrastructure that will advance the use of electronic health data in comparative effectiveness research (CER) and other types of research. In December 2013 funding for 11 clinical data research networks (CDRN) and 18 patient-powered research networks (PPRN) has started and meanwhile been prolonged for a second funding phase (Fleurence 2014, Collins 2014). Just recently a major result from the cooperation of all PCORnet CDRN has been published. It is the initiation of a pragmatic trial which aims at recruiting 20.000 patients who are at high risk for ischemic events. In all PCORnet institutions patients are recruited based on EHR data and respective routine care documentation is regularly transferred into the PCORnet Common Data Model (CDM; compare http://www.pcornet.org/pcornet-common-data-model/).

The PCORnet CDM is based on the Mini-Sentinel Common Data Model (MSCDM; www.mini-sentinel.org) and has been informed by other distributed initiatives such as the HMO Research Network, the Vaccine Safety Datalink, various AHRQ Distributed Research Network projects, and the ONC Standards & Interoperability Framework Query Health Initiative. The PCORnet CDM is positioned within healthcare standard terminologies (including ICD, SNOMED CT, CPT, HCPCS, and LOINC) to enable interoperability with and responsiveness to evolving data standards.

3.5 EHR4CR

A major project in Europe to reuse routine clinical data is Electronic Health Records For Clinical Research (EHR4CR). Composed out of 11 EFPIA pharmaceutical companies, 22 public partners (academia, hospitals and SMEs) and 5 subcontractors (Lastic 2011), the project aims to accelerate various aspects of clinical trial execution. The project is working on four use-cases (Fritz 2015):

1. Clinical protocol feasibility
2. Patient identification and recruitment
3. Clinical trial execution
4. Serious adverse event reporting

The project developed a distributed architecture similar to SHRINE/i2b2, but surpassed such previous platforms in various aspects.The project has developed its own conceptual reference model (EHR4CR information model), based on a HL7-based UML model (Ouagne 2013) that also includes a set of common data elements that are used across the EHR4CR network (Doods 2014). The project aimed to use the Object Contraint Language (OCL) to map between the central and local information models (Ouagne 2013). However, the final prototype uses an SQL code generating approach to handle query transformations (Bache 2013), which also includes its own human-readable (internally used) query language, ECLECTIC (Bache 2013). The EHR4CR system supports two types of data stores, the EHR4CR 'native' schema and i2b2 (Doods 2014). This could be achieved by implementing syntactic and semantic mappings between the EHR4CR and i2b2 architecture. The first was implemented by using "mini SQL" queries by retrieving data records from the i2b2 schema and transforming them according to the EHR4CR internal schema. The latter could be achieved by using 1:1 conceptual mappings at a low-granularity level. Another major achievement is that the project managed to sort out legal issues, such as patient data protection and data ownership, across Europe. The prototypical initial software platform was afterwards redeveloped by one of the project's industry partners (Custodix) and is currently marketed under the name InSite. As one of the key sustainable entities arising from the Electronic Health Records for Clinical Research Project in collaboration with several other European projects and initiatives supported by the European Commission the European Institute for Innovation through Health Data (*i*~HD, www.i-hd.eu) has been formed at the end of 2015. It shall specifically address obstacles and opportunities to using health data by collating, developing, and promoting best practices in information governance and in semantic interoperability (Kalra 2015).

3.6 MIABIS Connect

MIABIS Connect has been developed as part of the BioMedBridges project (Building data bridges from biology to medicine in Europe) that has the objection to facilitate data integration in different domains by developing shared technical infrastructure (http://www.biomedbridges.eu/).

One of their work packages, WP10 (Integrating disease related data and terminology from samples of different types), created for this reason a prototype which is based on MIABIS 2.0 and serves as “a sample-centred biobank federation framework that aims to facilitate sample discovery among biobanks members of a federation” (Merino 2015). Since MIABIS Connect is an open-source software framework, it can easily be used by biobanks and research communities. Moreover, the effort to maintain the software is minimal. The data harmonisation process of the data is designed to adjust to different semantic models by allowing the biobanks to keep their local semantic, while simultaneously using the MIABIS semantic for data sharing. The shared data itself does not leave the biobank but can only be viewed as a search result through the query interface (Merino 2015).

The two main modules are MIABIS Server and MIABIS Client. The server is a software component that has to be installed locally in the biobank. After uploading the biobank data in a file repository and mapping it to MIABIS, the data can be queried. For this purpose the client – a web application based on Kibana (https://www.elastic.co/de/products/kibana) – enables the researchers to search for appropriate samples in all biobanks in the federation (Kovalevskaya 2016). The MIABIS Server itself also consists of two components:

1. The Mapper takes over the task to map the attributes in the uploaded data files against the MIABIS attributes. Following this process a map file is saved in MIABIS server. All sub-

sequent data exports are done by reference to this map file, unless changes to the biobank's internal data model are made.
2. The Converter "reads the biobank files together with the produced map and indexes the results in an instance of ElasticSearch" or translates it into the MIABIS sample exchange format, MIABIS-TAB, which will be introduced to the biobank community in the future (Merino 2015).

The next steps of WP10 will include a closer collaboration with BBMRI-ERIC not only regarding ELSI CS and Quality CS but also by extending MIABIS Connect "to the European level through BBMRI Common Services IT" (Merino 2015).

3.7 MOLGENIS

3.7.1 Metadata Model and Mapping Registry (MMMR)

The Metadata Model and Mapping Registry (MMMR, part of the MOLGENIS project) is a registry that simplifies the search and selection of existing data models, formats, and guidelines. The MMMR uses a Google-like search to enable researchers to find a collection of meta-data artefacts of use to the biomedical science community (see: http://www.biomedbridges.eu/making-sure-standards-are-fit-purpose). The registry consists of five components:

1. Meta-data model: framework to capture descriptions of the structural elements
2. Content: actual meta-data
3. Registration mechanisms: systems to enter meta-data via user interface
4. Query interfaces: alternative views to enable human users to search across all collected meta-data
5. Mappings: system to view and curate entity/ attribute mappings

An integrated mapping tool is used to generate mappings across data models. This enables data integration by establishing harmonisation rules. In addition, it displays how other data models map onto searched parameters. This mapping view is useful regarding already existing standards to facilitate the mapping between them. Moreover, it can serve as template for new standard proposals. And finally, the end-user has the opportunity to upload their local data and get assistance to convert it in order to upload it to public repositories.

The tool is implemented in collaboration with EU-BioSHaRE project using the BiobankConnect software. The registry content is open access and the software open-source (McMurry 2014).

3.7.2 BioSHaRE / BiobankConnect

The mission of the European FP7 BioSHaRE Project (see https://www.bioshare.eu/) is to ensure the development of harmonised measures and standardised computing infrastructures (compare Doiron (2013) and respective YouTube Videos (https://www.youtube.com/watch?v=rfOP7C_odNs&ebc=ANyPxKqrdglGLynT3BZioOu0vgUukFo4sBIXjhrSAu-tQwn2QGhtUZQn6f9_AVNXOkJ8dK2Vn480JaOarOmbsHcDj1nl1RJQgQ)).

It has been designed to contain nine interrelated work packages from which the WP on “Data repository and epidemiological/clinical harmonization” (WP2: UMCG Morris Swertz) and “Biospecimen harmonization/standardization" (WP5: HMGU Melanie Waldenberger) might be ones to look at, in terms of biospecimen/clinical data harmonisation tools. As described by the BioSHaRE team “a variety of tools and methods are developed in BioSHaRE for retrospective and prospective harmonization, facilitating full valorisation of the database contents for the scientific community”. The respective tool developed for this purposes is MOLGENIS BiobankConnect (Pang 2015a, https://www.youtube.com/watch?v=Gc1VKRCmTWU) which implements a three-step harmonisation strategy. First, data elements of interest, which are defined based on the research question, are manually annotated with ontology terms. Then, these ontology terms are used to automatically scan the descriptions of the thousands of available data elements from each biobank to find potential matches. Finally, all candidate matches are sorted from ‘best’ to ‘worst’ so researchers can quickly decide on a useful match. Such steps among others rely on the application of Opal (OBiBa's core data warehouse; compare e.g. http://www.obiba.org/pages/products/opal/)

In BioSHaRE retrospective harmonisation led to the generation of common format variables for 73% of matches considered (96 targeted variables across 8 studies). This enabled authenticated investigators to perform complex statistical analyses of harmonised datasets stored on distributed servers without actually sharing individual-level data using the DataSHIELD method (Gaye 2014).

3.7.3 SORTA

The System for Ontology-based Re-coding and Technical Annotation (SORTA) was developed to semi-automatise the process that is needed to perform a retrospective standardisation of biomedical data. This process includes "matching of original data to widely used coding or ontology systems such as SNOMED CT" (Pang 2015b) and ICD-10. "For each data value, SORTA provides a list of the most relevant standard codes based on the lexical similarity in percentage. Users can then pick the correct matches from the suggested list" (http://molgenis.github.io/documentation/guide-SORTA). SORTA uses Lucene and n-gram based matching algorithms and works in the following three steps:

1. Coding systems or ontologies are uploaded and indexed in Lucene to enable fast searches.
2. Users create their own coding/recording project by uploading a list of data values.
3. Users receive a shortlist of matching concepts for each value from which they can select the suitable mapping.

As described in (http://molgenis.github.io/documentation/guide-SORTA), "standard codes (ontologies) can be imported using the EMX format, the model can be browsed and viewed as a UML diagram as well as a flat list in the webbrowser".

3.8 GEN2PHEN

GEN2PHEN is a project that "aims to unify human and model organism genetic variation databases towards increasingly holistic views into Genotype-To-Phenotype (G2P) data, and to link this system into other biomedical knowledge sources via genome browser functionality" (see: http://gen2phen.org/about-gen2phen). In the scope of this project a system was established that is partially centralised and partially federated, supports direct and automated data submission, and enables powerful comprehensive searching.

Work package 3 ("standard data models and terminologies") was in charge of formulating standards that were necessary for the G2P database development and the data exchange within G2P, since no structured information representation of the biomedical data existed. In order to achieve this goal, the group developed a system for storing and sharing data. One component is the Observ-OM, a minimal model, and Observ-TAB, a spreadsheet format based upon Observ-OM. Observ-OM (object model) describes the principle data elements and their relationships by capturing the core information about scientific observations. This was a result of interactive workshops involving experts from model organism and human resource communities. It also provides a common language that is able to harmonise representations and support software implementations. Observ-TAB is a tabular exchange format, that was developed to be able to share observation data and metadata without the need for complex informatics support. Existing ontologies are also compatible with the Observ-OM/TAB system by allowing specific references to ontology terms (Adamusiak 2012, GEN2PHEN Final Report 2014).

3.9 PhenX Toolkit

PhenX (consensus measures for Phenotypes and eXposures) aims to facilitate cross-study analyses by promoting the consistent use of measurement protocols. This proceeding makes it easier to validate results and combine study data (see: https://www.phenx.org/). In order to achieve this goal, the toolkit provides "a core set of well-established, low-burden standard measures to collect phenotypic and environmental data for large-scale genomic studies" (Pathak 2011). By using the toolkit users can select measures of interest and request further information, e.g. how to collect data on these measures. Then a report is provided that includes a description of the measure and other related information. Afterwards the user can review the protocol that was used to obtain the requested data on the measure and assess if the protocol was useful for their study. This way measures can be identified that are suitable for a certain study population (Hamilton 2011).

3.10 RD-Connect

RD-Connect is a European project, aiming at establishing an integrated platform to host and analyse genomic and clinical data from research projects as well as clinical bioinformatics tools for analysis and integration of molecular and clinical data to discover new disease genes, pathways, and therapeutic targets (Thompson 2014). The platform has a special focus on rare disease research. Unfortunately, until today no detailed results and platform/tools for sample querying (directly on the sample level) and data harmonisation have been published. Currently rare disease biobanks are invited to make samples accessible to the wider scientific community via a searchable and dynamic sample database (see http://rd-connect.eu/platform/biobanks/publishing-rare-disease-sample-collections-on-the-rd-connect-catalogue/), which however, does not seem to al-low searches on a direct sample level. The RD-Connect online catalogue is rather made up of a collection of existing databases, registries and biobanks called ID-Cards. Even though the term “Common Data Elements” (CDEs) was introduced by the NIH/NCATS Global Rare Diseases Patient Registry Data Repository (GRDR, https://grdr.ncats.nih.gov/) programme to define those database fields (elements) which could be used in any rare disease registry and RD-Connect promotes such CDEs, no such CDE list/description is available on the website yet. It currently only tells, that "CDEs are necessary to ensure that data are defined in the same way, use the same standards, and use the same terminologies. The use of CDEs facilitates the standardisation of data entry and allows for harmonisation, sharing and exchange of information across registries and diseases, and various analyses and studies. These CDEs are designed to capture the information at a minimum level of detail that is needed for all or most clinical research on rare diseases. Based on the comparative analysis among the existing lists of CDEs, a minimum data set for patient data entry to be used in the RD-Connect framework has been developed.

3.11 European Genome-phenome Archive (EGA)

Also focusing on the distribution and sharing of genetic and phenotypic data, the European Genome-phenome Archive (EGA) shall be established as a permanent archive (Lappalainen 2015). It was originally launched in 2008 by the European Molecular Biology Laboratory’s European Bioinformatics Institute (EMBL-EBI) to support the voluntary archiving and dissemination of data requiring secure storage and distribution only to authorised users. It has later expanded from an exclusively EMBL-EBI project to a collaboration with the Centre for Genome Regulation (CRG) in Barcelona, Spain. This may be a first step toward a larger distributed network of data archiving and dissemination services. Since the launch of the EGA, researchers from around the world have deposited and accessed data from over 700 of its studies of various types and the EGA has grown from about 50 TB to 1,700 TB in 2015. In its current version however, no user-friendly querying tool is provided in order to efficiently retrieve available data sets. Currently the EGA websites (through which data sets can be requested) are arranged around the study concept, where a study is typically an experimental investigation of a particular phenomenon, for example, a genome-wide association study or a matched tumor-normal cancer genome project. An EGA study page describes how the study was conducted and all the associated data sets. Currently however, phenotype data are most often provided only at the level of the data set, rather than the individual. Only few submissions included individual-level, detailed phenotypes.

Thus, not having any ontology-based querying support today, in future the EGA project team shall also work on several new added-value services that will increase the usability of the submitted data. For example, submitted sample phenotypes shall be described using ontology-based terms to facilitate better search functionality and assist users looking to merge data across studies. Links shall further be established with literature databases such as Europe PubMed Central to more closely track secondary publications based on data from the EGA (Lappalainen 2015). Nevertheless, all such developments currently seem to be available only on the planning level and no detailed description for ontology or data harmonisation tools have been found in the literature yet.

3.12 bioCADDIE

bioCADDIE (biomedical and healthcare Data Discovery Index Ecosystem) is an NIH BD2K (Big Data to Knowledge initiative) Data Discovery Index Coordination Consortium that seeks to develop a data discovery index (DDI) (see: https://biocaddie.org). The purpose of such a tool is to index data that are stored in other places. It is not a storage environment. This way information about the availability of data objects, their access requirements and the way how these different objects are related to each other can be provided. On one hand it helps researchers to find relevant data and on the other hand it guides data producers to share their data efficiently (Sansone 2015).

bioCADDIE works in close collaboration with the NIH Commons. “The Commons is defined as a shared virtual space where scientists can find, deposit, manage, share and reuse data, software, metadata and workflows - the digital objects of biomedical research” (see: https://datascience.nih.gov/commons). In regard of the DDI, bioCADDIE can benefit from a cloud-based storage or access controls that the Commons can provide.

The development team for the DDI considered several technology approaches for the prototype:

1. Relational database technologies like MySQL or Oracle, since many data repositories are based on these technologies and are well known
2. Semantic web technologies or Linked Data (e.g. DataOne, NIF) in order to connect different databases
3. Search technologies that can be utilised to create data indexes as well as connect data repositories

The particular characteristics of the prototype, e.g. how the data query works, haven’t yet been described in detail. But they intend to refer to already existing strategies or approaches that are pursued in other projects like ELIXIR with which they want to exchange experience in regard of identifier systems or metadata specification. All considerations of the components needed for a DDI are described in the project’s white paper (Ohno-Machado 2015).

The bioCADDIE prototype DataMed (version 0.5) has been running since the early month of 2016. Working group 3 is in charge of delivering the DATS (DatA Tag Suite) model for DataMed as a scalable way to index data sources. It describes the metadata and the structure of the datasets. The initial phase of this work package has already been completed. The result was a specification document with a core metadata requirement list as deliverable (see: https://drive.google.com/folderview?id=0B5XliQRRTcWIfmllZlNDb2F4NHJ4ZnJUV1F0akNVZnV2cVBBM2lIU1dJelk4WVNYQmZoZWc&usp=sharing). The elements of the core metadata set were derived from the combination of a bottom-up mapping and a top-down use case approach. By analysing preassigned use cases, questions were identified, whose answers should be supported by DataMed. Then the questions were abstracted in order to match them with the results of the bottom-up mapping, which covers generic metadata and life science-specific schemas (Gonzalez-Beltran 2015, Rocca-Serra 2015). Phase 2 covers the adjustment of the DATS model as needed, the definition of a core as well as an extended model and the definition of best practice guidelines regarding the interoperability with the DataMed prototype. This phase is expected to be completed by mid 2016.

3.13 DKTK CCP-IT

The German Cancer Consortium (DKTK, see: https://www.dkfz.de/en/dktk/about_us/about_us.html) aims to combine clinical data of all partner sites for networked biospecimen retrieval and data sharing. A clinical communication platform (CCP) has been established as the central interface between clinical research and basic research. In this context DKTK CCP-IT is focused on the provision of the relevant networking, data sharing and data harmonisation technologies. For this purpose the DKTK CCP-IT research group has proposed the decentral search – a modification of federated search services that exploits distributed, heterogeneous, highly sensitive datasets from heterogeneous systems for overarching research questions (Lablans 2015). In order to create the technical and semantic interoperability required to search the heterogeneous databases, the decentral search relies on a set of software components called bridgeheads. Using ETL processes, relevant data are regularly extracted from the primary system, transformed into a common format and loaded into each partner’s bridgehead. Semantic interoperability between the bridgeheads is achieved by means of a central metadata repository (Kadioglu 2013) that can import and reuse data elements from other ISO 11179-based implementations and is build upon the OSSE MDR (see 3.19).

The metadata repository provides attributes according to the ISO 11179 representative layer via a RESTful web interface. However, within DKTK it is still up to the consortium to define the required data elements and it is the task of the data owner to map each attribute’s key to a data element in the metadata repository. The DKTK CCP-IT team describes that the latter process may require the harmonisation of different data models, which is a complex issue that may involve more than one-to-one mappings and simple data transformations (for tools to support this compare e.g. Mate 2015).

3.14 OSSE Metadata Repository

The OSSE project (Open-Source-Registersystem für Seltene Erkrankungen in der EU/ Open-Source Registry System for Rare Diseases in the EU), which was funded by the German Federal Ministry of Health, aims to provide reusable software for RD registries (Muscholl 2014). The project has developed various open-source software components, one of them being a metadata repository, the OSSE MDR (https://www.osse-register.de/en/). The main purpose of the MDR is to aid with harmonising data elements. It is designed to operate in harmony with other OSSE components, e.g. in combination with the OSSE form repository, which is used to define and store EDC forms. While it also has interfaces to other ressources, such as the below mentioned BMBF MDR and the National Cancer Institute's caDSR allowing to reuse and interlink data elements, it can also be used as a stand-alone system (https://bitbucket.org/medinfo_mainz/samply.edc.osse/wiki/Home). The OSSE MDR was also used as a basis for the DKTK CCP-IT MDR (see 3.13) and for the development of the OnkoWiki system (www.tumor-wiki.de), a system to develop and refine data elements for standardised cancer patient documentation in German clinical cancer registries.

3.15 TMF/BMBF German National Metadata Repository Prototype

Initiated by the TMF (http://www.tmf-ev.de/) and funded by the German Federal Ministry of Education and Research, a prototype for a national metadata repository (MDR) has been developed that aims to support the reuse of data elements for the planning and support of databases and documentation processes for prospective clinical and epidemiologic studies and registries (https://mdr.imise.uni-leipzig.de/). The system has been built based on the below-mentioned ISO/IEC 11179 standard (Ngouongo 2013). This shared data model allows the exchange of content with other international repositories by providing methods for the naming, identification, classification and representation of data elements. Grouping of data elements and the possibility to assign data elements to a clinical study context enables the efficient structuring of content. The system is capable of referencing common medical classifications and terminologies, such as ICD-10, OPS, MedDRA, TNM, LOINC, SNOMED CT, CDISC CDASH and UCUM (https://mdr.imise.uni-leipzig.de/MDR-Broschuere.pdf).

3.16 The DPKK Ontology-Based Data Harmonization Pilot

The DPKK (Deutsches Prostatakarzinom Konsortium e. V.) is a German cross-institutional research network consisting of more than 70 urologists, pathologists and scientific researchers to fight prostate cancer. Similar to the CPCTR’s efforts in the U.S. (Patel 2005), one of their goals is to establish a shared database of tissue specimen, containing annotation data from the patients’ medical history, surgery and pathology (Mate 2011). As a basis for this in 2011, a common dataset has been defined by DPKK experts in Erlangen and Münster, comprising 26 medical concepts (e.g. pTNM) with 154 atomic enumerable values (e.g. pN=0) and 12 medical concepts with non-enumerable values (e.g. the PSA value). In this context an ontology based system for the mapping of EMR data to the common data elements has been developed and further refined in the following years (Mate 2011, Mate 2015). In this ontology-based approach, instead of defining ETL procedures at the database level, ontologies were used to organise and describe the medical concepts of both the source system and the target system. Instead of using unique, specifically developed SQL statements or ETL jobs, declarative transformation rules were defined within ontologies. This research illustrates how these constructs can then be used to automatically generate SQL code to perform the desired ETL procedures. It demonstrates how a suitable level of abstraction may not only aid the interpretation of clinical data, but can also foster the reutilisation of methods for unlocking it.

3.17 DebugIT

The DebugIT EU project aims to analyse and align antibiotics treatment schemes in order to counteract the increasing antibiotics resistances (Schober 2010). For this purpose, DebugIT compiles a global database of EHR data from several European hospitals. The mapping takes place in several steps from the local data to representations with a core ontology for epidemiological research. After completing the required levels of data integration, the data is globally accessible and clinical questions can be processed. Those questions are expressed with the DebugIT core ontology. Then "a query generation service translates the clinical questions to corresponding SPARQL queries expressed with the source ontology, which are then executed on SPARQL endpoints at each site" (Sun 2015). The mapping of the source data to the core ontology is conducted by a conversion service using the EYE reasoner. This approach allows the hospitals to keep their EHR data locally, while the outcomes of clinical questions can be viewed in aggregated form on a central dashboard (Sun 2015). DebugIT (Assele Kama 2012) uses D2RQ to generate a Data Definition Ontology (DDO) to achieve database bindings.

3.18 SALUS

SALUS is a STREP-funded EU project (Daniel 2014) that aims to help to post-market safety of medication by integrating and using EHR data. This includes the automatic detection and reporting of adverse drug events (ADE) and real-time screening of multiple, distributed, heterogeneous EHRs for ADE detection (Erturkmen 2012). According to (Erturkmen 2011), the project aims to achieve this by implementing a semantic interoperability solution for EHR data, by implementing mechanisms for security and privacy, and by "providing a novel exploratory analysis framework for open-ended temporal pattern discovery for safety studies" (Erturkmen 2011). SALUS apparently aimed to build upon existing standards whenever possible. According to (Erturkmen 2011), "SALUS functional interoperability profiles will be based on the available initiatives for achieving syntactic interoperability for reuse of EHRs for clinical trial execution, such as the IHE Retrieve Form for Data-Capture (RFD) [...], Clinical Research Data Capture (CRD) [...], Drug Safety Content Profile (DSC) [...] Profiles, HL7 Clinical Research Filtered Query Service Function Model (CRFQ SFM) [...], and where suitable by proposing the necessary extensions for enabling such a standard based interoperability architecture for post-market surveillance".

As outlined by these authors (Erturkmen 2011), the project faced the requirement to identify patient cohorts through a machine processable mechanism, that can automatically process eligibility criteria. In their later paper (Erturkmen 2012), they presented a semantic approach that facilitates semantic web technology and is used for the bridging between various clinical research standards (CDISC, MedDRA, WHODD) and clinical care standards (HL7, SNOMED CT, LOINC, ICD-10). This could be achieved by transforming the BRIDG DAM model into a RDF representation to act as a central information model. Similarly, standards, such as HL7 CDA, RxNORM, MedDRA, SNOMED CT, ICD-9 and ICD-10 were integrated as RDF into this model, therefore allowing transformation of data across these.

Later research (Declerck 2015) indicates that the project analysed many further standards (e.g. HL7/ASTM CCD and IHE PCC templates, HL7 Health Quality Measures Format (HQMF) and many more) and outlines that other common source data models (IHE PCC/HL7 CDD, ORBIS data model, ISO/CEN EN 13606, HL7 HQMF) and target data models (OMOP CDM, ICH E2B(R2)) were mapped to a common information model, called SALUS CIM. The project also uses SKOS for ontological modelling, e.g. for ICD-10, SNOMED CT, MedDRA, ICD-9-CM and various other ontologies to map between those. Via terminology reasoning, additional information can be inferred (e.g. transitivity of hierarchical relations) (Declerck 2015).

3.19 LexEVS

The LexEVS terminology server is a metadata repository software that was developed within the caBIG project and is now being used by several projects (eg, ePCRN, NCI thesaurus browser, BioPortal). Internally, LexEVS unificates terminologies by relying on the LexGrid format. Since version 6, LexEVS implements the HL7 common terminology services 2 (CTS 2) service functional model (SFM) (Ethier 2013). According to its website (https://wiki.nci.nih.gov/display/LexEVS/LexEVS), "LexEVS provides a common terminology model and open access to a wide range of terminologies, terminology value sets, and cross-terminology mappings needed by NCI and its partners."

4 General Research on Matching and Mapping

Besides major medical or infrastructure projects (as described above), several publications provide detailed descriptions of possible methods for semantic data harmonisation. The following can and shall not provide a complete overview, but aims on illustrating some of the aspects which are typically tackled in the matching/mapping process.

A common challenge is to map data elements from the local systems to a harmonised information model. The process of finding correspondences between the two is called "matching" (Bernstein 2011, Cruz 2009, Engmann 2007, Aumüller 2005). Rahm (Rahm 2001) classifies and describes various approaches in the field of automated schema matching, which address the challenge of finding correspondences between multiple relational databases. These can be found throughout in the scientific literature (e.g. in Bernstein 2011, Massmann 2011, Krogh 2011, Zhao 2007, Massmann 2006, Bernstein 2006, Shvaiko 2005, Pottinger 2003, Do 2002, Rahm 2001, Madhavan 2001, Milo 1998). Many of these approaches are also used for ontology-based implementations (Slabbekoom 2012, Hussain 2012, Giunchiglia 2012, Jean-Mary 2009, Euzenat 2007, Aleksovski 2006, Noy 2004, Mork 2004), where one tries to find correspondences between multiple ontologies or their concepts.

Automating the matching process is critical in the medical domain, because this laborous task is difficult to be handled manually. While structured and coded data elements, which are coded according to standardised terminologies and are typically used for billing, do not require matching (this process can be handled automatically during ETL), structured but uncoded data elements, which are used in assessment forms of electronic medical records (EMRs) or laboratory information systems can benefit from such approaches. Matching algorithms assist the user in mapping local data elements to standard classification. This has e.g. been demonstrated with mappings from legacy laboratory codes to LOINC (Zunner 2013, Lee 2013, Lau 2000). More advanced matchers reuse ontological knowledge. For example Fung (2007) demonstrated how the UMLS can be reused as a knowledge base to improve matching results.

5 Relevant ontologies, standards and metadata models in data harmonisation

In the following we briefly mention some nomenclatures, classification systems, terminologies, information models and meta models which might be relevant to consider and keep in mind when developing harmonisation tools. Due to the large amount of such various types of ontologies and information models being applied in many countries all over the world, it is impossible to mention all of them in this context. The below list is only meant to be exemplary and comprise our subjective selection of the "most important" and "most used" ontologies and information models.

- **SNOMED CT (Systematized Nomenclature of Medicine - Clinical Terms):** SNOMED CT is a terminology system that is owned, maintained and licensed by the International Health Terminology Standards Development Organisation (IHTSDO). It can be used to codify clinical EHR data to facilitate the electronic exchange of health information. In contrast to classification systems like ICD, SNOMED CT provides a higher degree of granularity and allows to define individual clinical concepts and their relationships. Moreover, while ICD generates information output for secondary data purposes, SNOMED CT enables information input for primary data purposes. However, it is possible to map SNOMED CT to other coding systems in order to achieve semantic interoperability (https://www.nlm.nih.gov/healthit/snomedct/; http://www.ihtsdo.org/snomed-ct). For example, SNOMED CT has been used in (El-Fadly 2011) as a "pivot terminology" to facilitate mappings (Ouagne 2013).

- **ICD (International Statistical Classification of Diseases and Related Health Problems):** "ICD is the foundation for the identification of health trends and statistics globally, and the international standard for reporting diseases and health conditions. It is the diagnostic classification standard for all clinical and research purposes" (http://www.who.int/classifications/icd/en/). It combines data of diseases, disorders, injuries and other related health conditions. It is structured hierarchically and is designed to enable (international) comparability in the collection, processing, classification and presentation of healthcare information and its statistical output. Thus healthcare facilities, regions, and even countries are able to share and compare health data among each other (http://www.who.int/classifications/icd/en/).

- **HL7 RIM (Reference Information Model):** HL7 RIM is, in addition to the Abstract Data Types Model and the Vocabulary Model one of the core models that are fundamental for all information models within HL7 (http://www.hl7.org/implement/standards/product_brief.cfm?product_id=77). It "is a static model of health-care information that broadly and abstractly covers all aspects of a health-care organization's clinical and administrative information" (Eggebraaten 2007). The constrained models derived from these "serve as documents, data for services, and messages" (http://www.hl7.org/implement/standards/product_brief.cfm?product_id=77). And "although it was not intended for the purpose of database design, RIM provides an integrated model for health-care data," it seems to be "a suitable basis for a data model used in a data warehouse architecture" (Eggebraaten 2007).

- **HL7 QRDA (Quality Reporting Document Architecture):** In order to evaluate health care quality, the HL7 QRDA standard was developed. By applying HL7 QRDA quality measure data can be communicated in an appropriate structure to the organisations that will then analyse and interpret the reported information. Further benefits is the compatibility with semi-automated reporting (e.g. information from manual chart reviews and abstraction), and the reuuse of templates from the HL7 Implementation Guide for CDA® Release 2 (http://www.hl7.org/implement/standards/product_brief.cfm?product_id=35).

- **openEHR EHR Information Model:** The openEHR project provides an EHR Information Model, "which is a model of an interoperable EHR in the ISO RM/ODP information viewpoint. [It] defines a logical EHR information architecture rather than just an architecture for communication of EHR extracts or documents between EHR systems" (Beale 2008a). Meaning, that it defines the "containment and context semantics of the (...) major (...) components of [an] EHR" (Beale 2008b).

- **ISO/IEC 11179:** ISO/IEC 11179 is an international standard that defines the method of representing metadata for organisations or computer systems in a metadata registry. Such a registry allows to manage data elements in a semantically precise structure. The data element, "a unit of data for which definition, identification, representation, classification and permissible values are specified by means of a set of attributes" (Ngouongo 2013) is therefore a key concept in an ISO/IEC 11179 metadata registry (Ngouongo 2013).

6 Conclusion

Data integration and data sharing within various types of research networks is currently an issue of many projects all over the world and by far not limited to biobanking networks. The above presented list can thus only illustrate a selective and subjective view of the Common Service IT WP 8 project team. Some of the projects described above have only recently started and do currently only focus on the provision of data sharing platforms without comfortable querying interfaces to support data retrieval (e.g. the EGA, see 3.11). Further, within some projects, retrieval and querying of data sets is yet only supported on an aggregated level, where complete data sets may be retrieved based on their metadata description, but querying for single patient- or sample-related data records is not yet possible (e.g. RD-Connect, see 3.10). For some projects the challenges and brief concepts of data harmonisation are described in respective publications, but more detailed tool descriptions or even freely available and reusable open-source software modules do not seem to be available (e.g. EHR4CR, DebugIT and SALUS, compare 3.5, 3.17, 3.18).

Several projects have described the usage of a common data model (often in combination with a predefined set of vocabularies) as a major prerequisite for data harmonisation and often such common data models are comprehensively described on the projects' websites. The OMOP CDM (see 3.1), the PCORnet CDM (see 3.4) (both stemming from earlier work leading to the Sentinel Distributed Database and Common Data Model; https://www.sentinelsystem.org/sentinel/data/distributed-database-common-data-model) and SALUS CIM (3.18) are only exemplary presented in this document, while many more such common data models exist. While those CDM are typically based on relational database schema with concepts represented as database tables, the i2b2/SHRINE data model is based on the entity-attribute-value concept, thus providing a more broader design approach. As described above in many projects transformation and mapping tools have been developed which support the transfer of data between those different common data models (even though in many cases such a transfer/mapping may not be pursued without at least partial Information loss).

In other projects, such as OSSE (3.14) and the DKTK CCP-IT development (3.13) a central metadata repository is established as a core component for data harmonisation, even though automated support for data element mapping is still missing. Such tools however have been realised and implemented for example in the MIABIS Connect (3.6) and MOLGENIS projects (3.7). Tools which seems to provide very efficient support in this context are MOLGENIS BiobankConnect (3.7.2) and SORTA (3.7.3).

Considering such previous developments, which partly are available as open-source tools, we recommend to not develop all such harmonisation tools from scratch, but consider reusing at least some of such tools (especially those, where the development teams are partners in BBMRI-ERIC CS-IT and where good documentation is available) and to evaluate those in more detail in terms of their applicability and integrability into a comprehensive toolset and harmonisation pipeline. For this further evaluation the requirements identified and described in parallel within D8.1 (BBMRI-ERIC CS-IT's Requirements Specification document; Eklund 2016) also need to be considered. The most important tools to consider for this purpose are the OSSE/DKTK CCP-IT MDR, MOLGENIS BiobankConnect and SORTA, as well as the ontology-based mapping library developed by Mate in the DPKK pilot project (Mate 2015).

References

Adamusiak T, Parkinson H, Muilu J, Roos E, van der Velde KJ, Thorisson GA, Byrne M, Pang C, Gollapudi S, Ferretti V, Hillege H, Brookes AJ, Swertz MA. Observ-OM and Observ-TAB: Universal syntax solutions for the integration, search, and exchange of phenotype and genotype information. Hum Mutat 2012;33(5):867–873.

Aleksovski Z, Klein M, ten Kate W, van Harmelen F. Matching Unstructured Vocabularies Using a Background Ontology. Proceedings of the 15th International Conference on Knowledge Engineering and Knowledge Management. 2006;182-197.

Assele Kama A, Primadhanty A, Choquet R, Teodoro D, Enders F, Duclos C, Jaulent MC. Data Definition Ontology for clinical data integration and querying. In: Mantas J et al. (Eds.). Quality of Life through Quality of Information. IOS Press, 2012. doi:10.3233/978-1-61499-101-4-38.

Aumüller D, Do HH, Massmann S, Rahm E. Schema and Ontology Matching with COMA++. Proceeding. SIGMOD '05 Proceedings of the 2005 ACM SIGMOD international conference on Management of data. 2005;906-8. doi: 10.1145/1066157.1066283.

Bache R, Miles S, Taweel A. An adaptable architecture for patient cohort identification from diverse data sources. J Am Med Inform Assoc. 2013;20(e2):e327-33. doi: 10.1136/amiajnl-2013-001858.

Beale T, Heard S, Kalra D, Lloyd D (Editors). The openEHR Reference Model.EHR Information Model. 2008a. © Copyright openEHR Foundation 2001-2008. All rights reserved. www.openEHR.org

Beale T, Heard S. openEHR Architecture. Architecture Overview. 2008b.© Copyright openEHR Foundation 2001-2008. All rights reserved. www.openEHR.org.

Bernstein PA, Melnik S, Churchill JE. Incremental Schema Matching. Proceeding. VLDB '06 Proceedings of the 32nd international conference on Very large data bases. 2006;1167-1170.

Bernstein PA, Madhavan J, Rahm E. Generic Schema Matching, Ten Years Later. Proceedings of the VLDB Endowment. 2011;4(11):695-701.

Collins FS, Hudson KL, Briggs JP, Lauer MS. PCORnet: turning a dream into reality. J Am Med Inform Assoc. 2014 Jul-Aug;21(4):576-7. doi: 10.1136/amiajnl-2014-002864. Epub 2014 May 12. PubMed PMID: 24821744; PubMed Central PMCID: PMC4078299.

Cruz IF, Antonelli FP, Stroe C. AgreementMaker: Efficient Matching for Large Real-World Schemas and Ontologies. Proceedings of the VLDB Endowment. 2009; 2(2):1586-9. doi: 10.14778/1687553.1687598.

Daniel C, Sinaci A, Ouagne D, Sadon E, Declerck G, Kalra D, Charlet J, Forsberg K, Bain L, Mead C, Hussain S, Laleci Erturkmen GB. Standard-based EHR-enabled applications for clinical research and patient safety: CDISC - IHE QRPH - EHR4CR & SALUS collaboration. AMIA Summits on Translational Science Proceedings. 2014;2014:19-25. PMCID: PMC4419753.

Declerck G, Hussain S, Daniel C, Yuksel M, Laleci GB, Twagirumukiza T, Jaulent MC. Bridging Data Models and Terminologies to Support Adverse Drug Event Reporting Using EHR Data. Methods Inf Med. 2015;54:24-31. http://dx.doi.org/10.3414/ME13-02-0025.

De Moor G, Sundgren M, Kalra D, Schmidt A, Dugas M, Claerhout B, Karakoyun T, Ohmann C, Lastic PY, Ammour N, Kush R, Dupont D, Cuggia M, Daniel C, Thienpont G, Coorevits P. Using electronic health records for clinical research: the case of the EHR4CR project. J Biomed Inform. 2015 Feb;53:162-73. doi: 10.1016/j.jbi.2014.10.006. Epub 2014 Oct 18. PubMed PMID: 25463966.

Do HH, Rahm E. COMA - A system for flexible combination of schema matching approaches. Proceedings of the 28th VLDB Conference. 2002.

Doiron D, Burton P, Marcon Y, Gaye A, Wolffenbuttel BHR, Perola M, Stolk RP, Foco L, Minelli C, Waldenberger M, Holle R, Kvaløy K, Hillege HL, Tassé AM, Ferretti V, Fortier I. Data harmonization and federated analysis of population-based studies: the BioSHaRE project. Emerg Themes Epidemiol. 2013;10:12. Published online 2013 November 21. doi:10.1186/1742-7622-10-12 PMCID: PMC4175511.

Doods J, Botteri F, Dugas M, Fritz F. A European inventory of common electronic health record data elements for clinical trial feasibility. Trials. 2014;15:18. doi:10.1186/1745-6215-15-18.

Doods J, Bache R, McGilchrist M, Daniel C, Dugas M, Fritz F. Piloting the EHR4CR Feasibility Platform across Europe. Methods Inf Med. 2014;53(4):264-8. doi:10.3414/ME13-01-0134.

Eggebraaten TJ, Tenner JW, Dubbels JC. A health-care data model based on the HL7 reference information model. *IBM Syst. J.* 46, 1 (January 2007), 5-18. DOI=http://dx.doi.org/10.1147/sj.461.0005.

Eklund N, Mate S, Schüttler C, Miettinen T, Knuuttila J, Prokosch HU, Silander K. D8.1 Requirements specification for data harmonisation and terminology mapping tools. 2016. Zenodo. https://doi.org/10.5281/zenodo.164398.

El Fadly A, ,Rance B, Lucas N, Mead C, Chatellier G, Lastic PY, Jaulent MC, Daniel C. Integrating clinical research with the Healthcare Enterprise: From the RE-USE project to the EHR4CR platform. Journal of Biomedical Informatics. 2011;44:S94 - 102.

Engmann D, Massmann S. Instance Matching with COMA++. Contribution to the BTW Workshop. 2007.

Erturkmen GB, Dogac A, Yuksel M, Hussain S, Declerck G, Daniel C, Sun H, Depraetere K, Colaert D, Devlies J, Krahn T, Thakrar B, Freriks G, Bergvall T, Sinaci AA. Building the Semantic Interoperability Architecture Enabling Sustainable Proactive Post Market Safety Studies. 2011.

Erturkmen GB, Dogac A, Yuksel M. SALUS: Scalable, Standard based Interoperability Framework for Sustainable Proactive Post Market Safety Studies. In: Mantas J et al. (Eds.). Quality of Life through Quality of Information. 2012.

Ethier JF, Dameron O, Curcin V, et al. A unified structural/terminological interoperability framework based on LexEVS: application to TRANSFoRm. J Am Med Inform Assoc. 2013;20(5):986-94.

European Commission. GEN2PHEN Report Summary. Final Report Summary – GEN2PHEN (Genotype-To-Phenotype Databases: A Holistic Solution). 2014.

Euzenat J, Shvaiko P. Ontology Matching. Springer-Verlag New York. 2007. ISBN: 3540496114.

FitzHenry F, Resnic FS, Robbins SL, Denton J, Nookala L, Meeker D, Ohno-Machado L, Matheny ME. Creating a Common Data Model for Comparative Effectiveness with the Observational Medical Outcomes Partnership. Appl Clin Inform. 2015 Aug 26;6(3):536-47. doi: 10.4338/ACI-2014-12-CR-0121. eCollection 2015. PubMed PMID: 26448797; PubMed Central PMCID: PMC4586341.

Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. J Am Med Inform Assoc. 2014 Jul-Aug;21(4):578-82. doi: 10.1136/amiajnl-2014-002747. Epub 2014 May 12. PubMed PMID: 24821743; PubMed Central PMCID: PMC4078292.

Fritz F, Tilahun B, Dugas M. The European initiative EHR4CR - Lessons learned for EHR implementations in Africa. Journal of Health Informatics in Africa. 2014;2(2). doi: 10.12856/JHIA-2014-v2-i2-102.

Fung KW, Bodenreider O, Aronson AR, Hole WT, Srinivasan S. Combining Lexical and Semantic Methods of Inter-terminology Mapping Using UMLS. In: Kuhn K et al. (Eds.). MEDINFO 2007. IOS Press, 2007.

Gaye A, Marcon Y, et al. DataSHIELD: taking the analysis to the data, not the data to the analysis. Int J Epidemiol. 2014 December; 43(6):1929–1944. Published online 2014 September 27. doi:10.1093/ije/dyu188. PMCID: PMC4276062.

Giunchiglia F, Autayeu A, Pane J. S-Match: an open source framework for matching lightweight ontologies. Semantic Web. 2012;3(3):1-9. doi: 10.3233/SW-2011-0036.

Gonzalez-Beltran A, Rocca-Serra P, Sansone SA et al. Standard Operating Procedure in: NIH BD2K bioCADDIE WG3 DataMed Data Discovery Index – DATS Metadata Specification v1.1.

Hamilton CM,Strader LC, Pratt J, Maiese D, Hendershot T, Kwok R, Hammond J, Huggins W, Jackman D, Pan H, Nettles D, Beaty T, Farrer L, Kraft R, Marazita M, Ordocas J, Pato C, Spitz M, Wagener D, Williams M, Junkins H, Harlan W, Ramos E, Haines. The PhenX Toolkit: Get the Most From Your Measures. American Journal of Epidemiology. 2011;174(3), 253-60. doi: 10.1093/aje/kwr193.

Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, Suchard MA, Park RW, Wong IC, Rijnbeek PR, van der Lei J, Pratt N, Norén GN, Li YC, Stang PE, Madigan D, Ryan PB. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform. 2015;216:574-8. PubMed PMID: 26262116.

Hussain S, Roo JD, Jaulent MC. Proof-based Ontology Matching: Finding semantic similarities between ancestor graph structures. 2012 IEEE Sixth International Conference on Semantic Computing (ICSC). doi: 10.1109/ICSC.2012.23.

Jean-Mary YR, Shironoshita EP, Kabuka MR. Ontology matching with semantic verification. Web Semant. 2009;7(3):235-251. doi: 10.1016/j.websem.2009.04.001.

Kadioglu D. Institutionsübergreifende Nutzung Verteilter Metadata Repositories [Master Thesis]. [Dortmund]: Fachhochschule Dortmund; 2013.

Kalra D, Stroetmann V, Sundgren M, Dupont D, Schlünder I, Thienpont G, Coorevits P and De Moor G. (2016) The European Institute for Innovation through Health Data, Learn Health Sys, doi: 10.1002/lrh2.10008.

Klann JG, Abend A, Raghavan VA, Mandl KD, Murphy SN. Data interchange using i2b2. J Am Med Inform Assoc. 2016 Feb 5. pii: ocv188. doi: 10.1093/jamia/ocv188. [Epub ahead of print] PubMed PMID: 26911824.

Kohane IS, Churchill SE, Murphy SN. A translational engine at the national scale: informatics for integrating biology and the bedside. J Am Med Inform Assoc. 2012 Mar-Apr;19(2):181-5. doi: 10.1136/amiajnl-2011-000492. Epub 2011 Nov10. PubMed PMID: 22081225; PubMed Central PMCID: PMC3277623.

Kovalevskaya N. DNAdigest interviews MIABIS Connect. March 30, 2016.

Krogh B, Weisberg A, Bested M. DBLint: A Tool for Automated Analysis of Database Design. Master Thesis. 2011.

Lablans M, Kadioglu D, Muscholl M, Ückert F. Exploiting Distributed, Heterogeneous and Sensitive Data Stocks while Maintaining the Owner's Data Sovereignty. Methods Inf Med. 2015b; 54(4):346-52. doi: 10.3414/ME14-01-013. Epub 2015 Jul 21. PubMed PMID: 26196653.

Lappalainen I, Almeida-King J, Kumanduri V, Senf A, Spalding JD, Ur-Rehman S, Saunders G, Kandasamy J, Caccamo M, Leinonen R, Vaughan B, Laurent T, Rowland F, Marin-Garcia P, Barker J, Jokinen P, Torres AC, de Argila JR, Llobet OM, Medina I, Puy MS, Alberich M, de la Torre S, Navarro A, Paschall J, Flicek P. The European Genome-phenome Archive of human data consented for biomedical research. Nat Genet. 2015 Jul;47(7):692-5. doi: 10.1038/ng.3312. PubMed PMID: 26111507.

Lastic PY. Integrating Clinical Research Services into Hospital Information Systems: The IHE-CDISC Perspective illustrated through the RE-USE and EHR4CR projects. Presentation, 2011.

Lau LM, Johnson K, Monson K, Lam SH, Huff SM. A Method for the Automated Mapping of Laboratory Results to LOINC.Proc AMIA Symp. 2000; 472-476. PMCID: PMC2244065.

Lee LH, Groß A, Hartung M, Liou DM, Rahm E. A multi-part matching strategy for mapping LOINC with laboratory terminologies. J Am Med Inform Assoc. 2014;21(5):792-800; doi: 10.1136/amiajnl-2013-002139.

Madhavan J, Bernstein PA, Rahm E. Generic Schema Matching with Cupid. Proceeding. VLDB'01 Proceedings of the 27th International Conference on Very Large Data Bases. 2001; 49-58. ISBN: 1-55860-804-4.

Makadia R, Ryan PB. Transforming the Premier Perspective Hospital Database into the Observational Medical Outcomes Partnership (OMOP) Common Data Model. EGEMS (Wash DC). 2014 Nov 11;2(1):1110. doi: 10.13063/2327-9214.1110. eCollection 2014. PubMed PMID: 25848597; PubMed Central PMCID: PMC4371500.

Mandl KD, Kohane IS, McFadden D, Weber GM, Natter M, Mandel J, Schneeweiss S, Weiler S, Klann JG, Bickel J, Adams WG, Ge Y, Zhou X, Perkins J, Marsolo K, Bernstam E, Showalter J, Quarshie A, Ofili E, Hripcsak G, Murphy SN. Scalable Collaborative Infrastructure for a Learning Healthcare System (SCILHS): architecture. J Am Med Inform Assoc. 2014 Jul-Aug;21(4):615-20. doi: 10.1136/amiajnl-2014-002727. Epub 2014 May 12. PubMed PMID: 24821734; PubMed Central PMCID: PMC4078286.

Massmann S, Engmann D, Rahm E. COMA++: results for the ontology alignment contest OAEI 2006. Proceeding. OM'06 Proceedings of the 1st International Conference on Ontology Matching. 2006;225:107-114.

Massmann S, Raunich S, Aumüller D, Arnold P, Rahm E. Evolution of the COMA match system. Proceeding. OM'11 Proceedings of the 6th International Conference on Ontology Matching. 2011;814:49-60.

Matcho A, Ryan P, Fife D, Reich C. Fidelity assessment of a clinical practice research datalink conversion to the OMOP common data model. Drug Saf. 2014 Nov;37(11):945-59. doi: 10.1007/s40264-014-0214-3. PubMed PMID: 25187016; PubMed Central PMCID: PMC4206771.

Mate S, Bürkle T, Köpcke F, Breil B, Wullich B, Dugas M, Prokosch HU, Ganslandt T**.** Populating the i2b2 database with heterogeneous EMR data: a semantic network approach. Stud Health Technol Inform. 2011;169:502-6. PubMed PMID: 21893800.

Mate S, Köpcke F, Toddenroth D, Martin M, Prokosch HU, Bürkle T, Ganslandt T**.** Ontology-based data integration between clinical and research systems. PLoS One. 2015 Jan 14;10(1):e0116656. doi: 10.1371/journal.pone.0116656. eCollection 2015. Erratum in: PLoS One. 2015;10(3):e0122172. PubMed PMID: 25588043; PubMed Central PMCID: PMC4294641.

McCowan C, Thomson E, Szmigielski CA, Kalra D, Sullivan FM, Prokosch HU, Dugas M, Ford I. Using Electronic Health Records to Support Clinical Trials: A Report on Stakeholder Engagement for EHR4CR. Biomed Res Int. 2015;2015:707891. doi: 10.1155/2015/707891. Epub 2015 Oct 11. PubMed PMID: 26539523; PubMed Central PMCID: PMC4619877.

McMurry AJ, Gilbert CA, Reis BY, Chueh HC, Kohane IS, Mandl KD. A Self-scaling, Distributed Information Architecture for Public Health, Research, and Clinical Care. J Am Med Inform Assoc. 2007;14:527-533. doi:10.1197/jamia.M2371.

McMurry AJ, Murphy SN, MacFadden D, Weber G, Simons WW, Orechia J, Bickel J, Wattanasin N, Gilbert C, Trevvett P, Churchill S, Kohane IS. SHRINE: enabling nationally scalable multi-site disease studies. PLoS One. 2013;8(3):e55811. doi:10.1371/journal.pone.0055811. Epub 2013 Mar 7. PubMed PMID: 23533569; PubMed Central PMCID: PMC3591385.

McMurry J, Parkinson H, Gormanns P, Muilu J, Sariyar M, Swertz M, Hendriksen D, Kelpin F, Jetten J, Pang Chao. Mapping and registry of ESFRI BMS standards (eSTR). Deliverable D3.2 of the BioMedBridge project. 2014.

Merino R, Amatya S, Jimenez R, Villaveces J, Swertz M, Roos M, Litton JE, Holub P, Brunak S, Sarkans U. A prototype federated query interface for information on biosamples, and linking of biosamples and disease terminology to genome. Deliverable D10.3 of the BioMedBridges project. 2015.

Milo T, Zohar S. Using Schema Matching to Simplify Heterogeneous Data Translation. Proceeding. VLDB '98 Proceedings of the 24th International Conference on Very Large Data Bases. 1998. 122-133. ISBN: 1-55860-566-5.

Mo H, Jiang G, Pacheco JA, Kiefer R, Rasmussen LV, Pathak J, Denny JC, Thompson WK. A Decompositional Approach to Executing Quality Data Model Algorithms on the i2b2 Platform. AMIA Jt Summits Transl Sci Proc. 2016 Jul 20;2016:167-75. PubMed PMID: 27570665; PubMed Central PMCID: PMC5001760.

Mork P, Bernstein PA. Adapting a Generic Match Algorithm to Align Ontologies of Human Anatomy. ICDE '04 Proceedings of the 20th International Conference on Data Engineering. 2004; 787. ISBN: 0-7695-2065-0.

Murphy S, Churchill S, Bry L, Chueh H, Weiss S, Lazarus R, Zeng Q, Dubey A, Gainer V, Mendis M, Glaser J, Kohane I. Instrumenting the health care enterprise for discovery research in the genomic era. Genome Res. 2009 Sep;19(9):1675-81. doi: 10.1101/gr.094615.109. Epub 2009 Jul 14. PubMed PMID: 19602638; PubMed Central PMCID: PMC2752136.

Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, Kohane I. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010 Mar-Apr;17(2):124-30. doi:10.1136/jamia.2009.000893. PubMed PMID: 20190053; PubMed Central PMCID: PMC3000779.

Muscholl M, Lablans M, Wagner TOF, Ückert F**.** OSSE – open source registry software solution. Orphanet Journal of Rare Diseases 2014 9(Suppl 1):O9 DOI: 10.1186/1750-1172-9-S1-O9.

Ngouongo SMN, Löbe M, Stausberg J. The ISO/IEC 11179 norm for metadata registries: Does it cover healthcare standards in empirical research? Journal of Biomedical Informatics. 2013; 46: 318-327.

Norlin L, Fransson MN, Eriksson M, Merino-Martinez R, Anderberg M, Kurtovic S, Litton JE. A Minimum Data Set for Sharing Biobank Samples, Information, and Data: MIABIS. Biopreserv Biobank. 2012 Aug;10(4):343-8. doi: 10.1089/bio.2012.0003. PubMed PMID: 24849882.

Noy NF. Semantic Integration: A Survey Of Ontology-Based Approaches. SIGMOD Record. 2004;33(4): 65-70.

Ohno-Machado L, Alter G, Fore I, Martone MA, Sansone SA, Xu H. bioCADDIE White Paper. June 3rd, 2015. https://figshare.com/articles/bioCADDIE_white_paper_Data_Discovery_Index/1362572.

Ouagne D, Hussain S, Sadou E, Jaulent MC, Daniel C. The Electronic Healthcare Record for Clinical Research (EHR4CR) information model and terminology. In: Mantas J et al. (Eds.). Quality of Life through Quality of Information. IOS Press, 2013. 534-38. doi: 10.3233/978-1-61499-101-4-534.

Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc. 2012 Jan-Feb;19(1):54-60. doi: 10.1136/amiajnl-2011-000376. Epub 2011 Oct 28. PubMed PMID: 22037893; PubMed Central PMCID: PMC3240764.

Patel AA, et al. The development of common data elements for a multi-institute prostate cancer tissue bank: The Cooperative Prostate Cancer Tissue Resource (CPCTR) experience. *BMC Cancer* 5 (2005).

Pang C, Hendriksen D, Dijkstra M, van der Velde KJ, Kuiper J, Hillege HL, Swertz MA. BiobankConnect: software to rapidly connect data elements for pooled analysis across biobanks using ontological and lexical indexing. J Am Med Inform Assoc. 2015a Jan;22(1):65-75. doi:

10.1136/amiajnl-2013-002577. Epub 2014 Oct 31. PubMed PMID: 25361575; PubMed Central PMCID: PMC4433361.

Pang C, Sollie A, Sijtsma A, Hendriksen D, Charbon B, de Haan M, de Boer W, Kelpin F, Jetten J, van der Velde JK, Smidt N, Sijmons R, Hillege H, Swertz M. SORTA: a system for ontology-based re-coding and technical annotation of biomedical phenotype data. Database. 2015b, 1 - 13. doi: 10.1093/database/bav089.

Pathak J, Pan H, Wang J, Kashyap S, Schad PA, Hamilton CM, Masys DR, Chut CG. Evaluating Phenotypic Data Elements for Genetics and Epidemiological Research: Experiences from the eMERGE and PhenX Network Projects. AMIA Jt Summits Transl Sci Proc. 2011; 2011: 41–45.

PCORnet PPRN Consortium, Daugherty SE, Wahba S, Fleurence R. Patient-powered research networks: building capacity for conducting patient-centered clinical outcomes research. J Am Med Inform Assoc. 2014 Jul-Aug;21(4):583-6. doi: 10.1136/amiajnl-2014-002758. Epub 2014 May 12. PubMed PMID: 24821741; PubMed Central PMCID: PMC4078295.

Pottinger RA, Bernstein PA. Merging Models Based on Given Correspondences. Proceedings of the 29th VLDB Conference. 2003.

Rahm E, Bernstein PA. A survey of approaches to automatic schema matching. The VLDB Journal. 2001;10:334-350. doi: 10.1007/s007780100057.

Rocca-Serra P, Vardigan M et al. Use Cases and Derived Metadata in: NIH BD2K bioCADDIE WG3 DataMed Data Discovery Index – DATS Metadata Specification v1.1.

Rijnbeek PR. Converting to a common data model: what is lost in translation? Commentary on "fidelity assessment of a clinical practice research datalink conversion to the OMOP common data model". Drug Saf. 2014 Nov;37(11):893-6. doi: 10.1007/s40264-014-0221-4. Erratum in: Drug Saf. 2014 Dec;37(12):1073. PubMed PMID: 25187018.

Sansone SA. bioCaddie Informational Poster. March 30th, 2015. https://biocaddie.org/sites/default/files/d7/presentation/430/biocaddie-poster_2_1.pdf.

Schober D, Boeker M, Bullenkamp J, Huszka C, Depraetere K, Teodoro D, Nadah N, Choquet R, Daniel C, Schulz S. The DebugIT Core Ontology: semantic integration of antibiotics re-sistance patterns. In: Safran C et al. (Eds.). MEDINFO 2010. IOS Press, 2010. doi:10.3233/978-1-60750-588-4-1060.

Schuemie MJ, Gini R, Coloma PM, Straatman H, Herings RM, Pedersen L, Innocenti F, Mazzaglia G, Picelli G, van der Lei J, Sturkenboom MC. Replication of the OMOP experiment in Europe: evaluating methods for risk identification in electronic health record databases. Drug Saf. 2013 Oct;36 Suppl 1:S159-69. doi: 10.1007/s40264-013-0109-8. PubMed PMID: 24166232.

Shvaiko P, Euzenat J. A Survey of Schema-based Matching Approaches. In: Spaccapietra S. Journal on Data Semantics IV. 2005; 146-171. doi: 10.1007/11603412_5.

Slabbekoorn K, Hollink L, Houben GJ. Domain-Aware Ontology Matching. Proceedings of the 11th Intl. Semantic Web Conference (ISWC), 2012.

Sun H, Depraetere K, De Roo J, Mels G, De Vloed B, Twagirumukiza, Colaert D. Semantic processing of EHR data for clinical research. Journal of Biomedical Informatics. 2015;58:247-259. http://dx.doi.org/10.1016/j.jbi.2015.10.009.

Thompson R, Johnston L, Taruscio D, Monaco L, Béroud C, Gut IG, Hansson MG, 't Hoen PB, Patrinos GP, Dawkins H, Ensini M, Zatloukal K, Koubi D, Heslop E, Paschall JE, Posada M, Robinson PN, Bushby K, Lochmüller H. RD-Connect: an integrated platform connecting databases, registries, biobanks and clinical bioinformatics for rare disease research. J Gen Intern Med. 2014 Aug;29 Suppl 3:S780-7. doi: 10.1007/s11606-014-2908-8. Review. PubMed PMID: 25029978; PubMed Central PMCID: PMC4124112.

Voss EA, Makadia R, Matcho A, Ma Q, Knoll C, Schuemie M, DeFalco FJ, Londhe A, Zhu V, Ryan PB. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. J Am Med Inform Assoc. 2015 May;22(3):553-64. doi: 10.1093/jamia/ocu023. Epub 2015 Feb 10. PubMed PMID: 25670757; PubMed Central PMCID: PMC4457111.

Weber GM. Federated queries of clinical data repositories: Scaling to a national network. J Biomed Inform. 2015 Jun;55:231-6. doi: 10.1016/j.jbi.2015.04.012. Epub 2015 May 6. PubMed PMID: 25957825; PubMed Central PMCID: PMC4464929.

Weber GM, Murphy SN, McMurry AJ, Macfadden D, Nigrin DJ, Churchill S, Kohane IS. The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. J Am Med Inform Assoc. 2009 Sep-Oct;16(5):624-30. doi: 10.1197/jamia.M3191. Epub 2009 Jun 30. PubMed PMID: 19567788; PubMed Central PMCID: PMC2744712.

Zhao H, Ram S. Combining schema and instance information for integrating heterogeneous data sources. Data & Knowledge Engineering. 2006;61(2):281-303. doi: 10.1016/j.datak.2006.06.004.

Zhou X, Murugesan S, Bhullar H, Liu Q, Cai B, Wentworth C, Bate A. An evaluation of the THIN database in the OMOP Common Data Model for active drug safety surveillance. Drug Saf. 2013 Feb;36(2):119-34. doi: 10.1007/s40264-012-0009-3. PubMed PMID: 23329543.

Zunner C, Bürkle T, Prokosch HU, Ganslandt T. Mapping local laboratory interface terms to LOINC at a German university hospital using RELMA V.5: a semi-automated approach. J Am Med Inform Assoc. 2012;20(2): 293-7. doi: 10.1136/amiajnl-2012-001063.