Convergence Meeting:
Semantic Interoperability for Clinical Research & Patient Safety in Europe

Summary Report

Paris, January 22, 2013
Organized by:
Sajjad Hussain
Christel Daniel
(INSERM, UMRS 872, eq 20)

Objectives:

Present technical approaches for integrating EHRs data to perform clinical research/patient safety, carried under different European projects—focusing on semantic interoperability issues.

List of Participated Projects:

EHR4CR	EURECA	SALUS	OpenPhacts	Linked2Safety	eTRIKS

Projects Analysis & Comparison Dimensions:

Information Models, Mete-data Repositories	Common Information Model(s) and MDR used for representing/ querying data from EHRs/CDWs How these Information Models and MDR are represented? UML, RDF/XML, others?
Clinical Terminologies /Ontologies	List of standard clinical terminologies used From where these terminologies (or subsets of terminologies) are retrieved? If you have extracted subsets of terminologies, which tools were used?
Terminology Services (Management & Mappings)	List of services supported by the terminology server. Mapping representation/exchange format? Terminology/Mapping versioning management?
Query Language (GUI)	Query building process: GUI (template-based, others) Query representation? Possibility of query expansion using terminology mappings?
EHR Data Quality and Preparation	Quality of source EHR data Any transformations (e.g. ETL, others) were used as data preparation step?
Data Exchange Format	Data exchange formats and protocols used to access EHRs/CDWs/others? Data exchange formats for sharing the query results from EHRs/CDWs/others?
Use-cases Description	List of scenarios/use-cases being targeted
Current Issues	Current issues dealing with semantic interoperability and future steps
Others	Any other important aspects to report
Analysis & comparison summary of the participating projects are shown in Table 2	

1. EHR4CR Project: Semantic Interoperability Approach

1.1. Project Summary

The EHR4CR (Electronic Health Records for Clinical Research) project aims to improve the efficiency and reduce the cost of conducting clinical trials, through better leveraging routinely collected clinical data in EHRs and using it at key points in the trial design and execution life-cycle. The EHR4CR platform will implement four use cases—protocol feasibility testing, patient identification and recruitment for clinical trials, supporting clinical trial execution and adverse event reporting—to be demonstrated by 10 pilots in 5 European countries. The EHR4CR platform will be a loosely coupled service platform, which orchestrates independent services. The EHR4CR architecture designed in WP3 (WP3: Architecture and Integration) will define how the tools and services of WP4 (Semantic interoperability), WP5 (Data Protection, Privacy & Security) and WP6 (end-user Platform Services) will integrate.

Project website: http://www.ehr4cr.eu/

1.2. Information Models, Mete-data Repositories

In EHR4CR we adopted the «A_SupportingClinicalStatementUniversal» model, component of the StudyDesign, proposed by the HL7 Regulated Clinical Research Information Model (RCRIM) Work Group. EHR4CR Information Model contains one central class ClinicalStatement, 4 dimension classes and 82 attributes. The central class ClinicalStatement is specialized into Act classes: i) Procedure, representing information related to clinical procedures; (ii) Observation referring to either a Condition representing the state of the person that is deemed to be "not normal" or a Finding, representing clinical findings about the patient that is observed and evaluated in a quantitative or qualitative way and iii) Substance Administration referring to a Medication prescribed to the patient. The 4 dimensions attached to the central class are: a) Subject, representing the information related to the subject of the clinical statement; b) Encounter, representing the information related to the administrative context of the clinical statement; c) Participation, representing additional information related to the medical context of the clinical statement; d) ClinicalStatementRelationship, representing the relationships between clinical statements.

Structures and Value Sets of Common Data Elements (CDEs) are defined in a Meta Data Repository (MDR) in order to specify additional constraints on the high-level EHR4CR information model in order to represent the fine-grained clinical information included in eligibility criteria constructs. For example, the CDE corresponding to the clinical statement “Systolic Blood Pressure” precisely defines the ”code” of the Clinical Statement (e.g. 271649006 standing for Systolic Blood Pressure in SNOMED CT), the data type of the “value” (e.g. Physical Quantity) and additional constraints of the “value” (e.g the unit (e.g. mmHg) if the data type is Physical Quantity).

1.3. Clinical Terminologies/Ontologies

The current version of the EHR4CR terminology contains various concepts from reference terminologies/ontologies that are uploaded from UMLS (e.g. SNOMED CT, LOINC, ICD-10, ATC codes) and other sources (e.g. PathLex, MedDRA).

1.4. Terminology Services (Management & Mappings)

Terminology Loading & Mappings

Establishing semantic interoperability between local EHRs and the platform for eligibility determination requires semantic matching between data elements describing eligibility criteria and concepts modeling patient data in heterogeneous clinical systems. In EHR4CR, we incorporate relevant modules from reference clinical terminologies (LOINC, SNOMED CT, ICD-10, HL7 vocabulary, PathLex, ATC) by loading their schema models into the EHR4CR terminology server. In addition, local terminologies used in each CDW are also uploaded to the terminology server. We define simple and complex mappings between the EHR4CR reference terminologies and local EHR terminologies. For example, a simple mapping relates one (or many) local code(s) of the plasma glucose measurement to the corresponding code in the EHR4CR pivot terminology (LOINC code 14749-6). A complex mapping shall address issues such as measurement unit conversation (mg/dL to mmol/L) and more complex issues related to the recognition of the context of use of the medical concepts (e.g. specific method or clinical context). The main objective for defining these mappings is to exploit them for extending the user-defined eligibility criteria and to generate more comprehensive and extended queries.

Terminology Services at the Query Workbench

- *Terminology Selection Service* allows users to select the preferred terminology in which the user wants to define the eligibility criteria.
- *Terminology Browsing Service* allows users to browse the preferred terminology and attached value sets to select a list of appropriate concepts to describe the eligibility criteria for querying.

Terminology Services at the Query Endpoint:

- *Query Expansion & Transformation Service* performs query expansion on the user-defined queries by walking through terminology hierarchies for a specific terminology concept to incorporate its narrower concepts (i.e. sub-concepts) into the query set. It invokes Terminology Mapping Services—for mapping between central and local terminology codes. Based on the pre-defined mappings, we transform the defined queries based on the local CDW terminology, which can then be executed across different clinical data warehouses to obtain more comprehensive query results.
- *Result Transformation and Aggregation Service* is designed to translate back the query-results obtained from various CDWs into an integrated result format based on the standardized medical vocabulary representing the initially given eligibility criteria. By using this service, the user can obtain the list of all matched patients from the various CDWs that satisfy the initially given eligibility criteria in one uniform and standard view. It invokes Terminology Mapping Service.

1.5. Query Language & Query Builder (GUI)

A template-based query interface at the User Workbench allows clinical researchers to define eligibility criteria based on the standardize terminologies, data elements and value sets using the EHR4CR Terminology Services. The defined set of eligibility criteria are represented as ECLECTIC queries based on the EHR4CR Eligibility Criteria Model (EC Model). The EC Model, proposed and developed by KCL (Kings College London), represents the defined eligibility criteria and formalize into EHR4CR query language ECLECTIC (Eligibility Criteria Language for European Clinical Trial Investigation and Construction). Based on the EC Model, ECLECTIC allows transformation schemes to transform elementary queries (defined in ECLECTIC) into other query languages—such as OCL (Object Constraint Language), SPARQL and SQL.

1.6. EHR Data Quality and Preparation

In order to investigate EHR data quality and preparation, a survey was designed by WP4 to establish the content, structure, semantics and some operational characteristics of the data sources available to EHR4CR at each hospital site. The structure of the survey is shown diagrammatically in Figure 1.

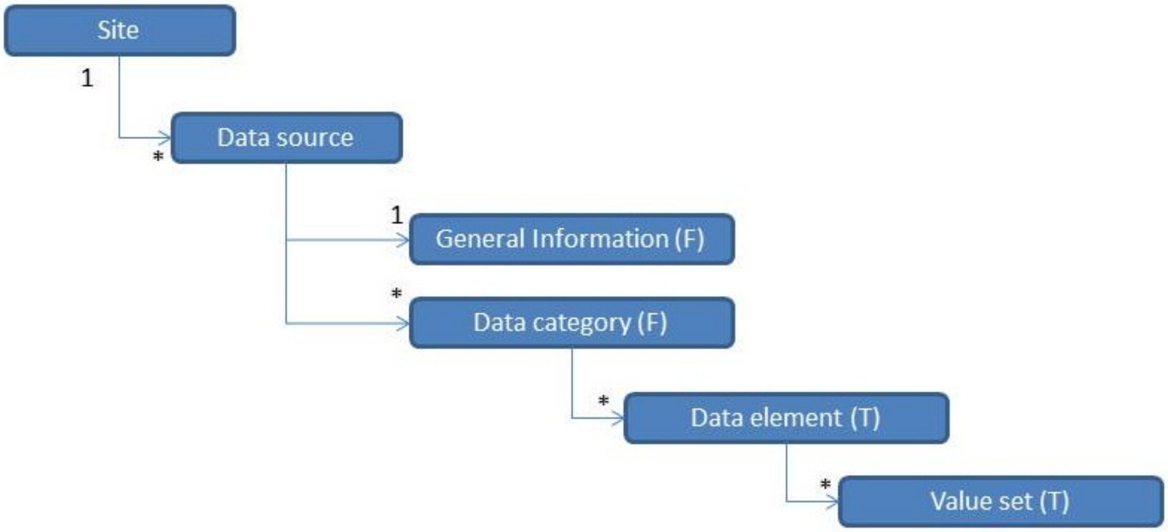

Figure 1. Survey structure and collection methods for hospital sites: F – web forms, T – template files

Each site was surveyed for 9 categories of data: Demography, Diagnosis, Procedure, Laboratory, Anatomic pathology, Medication, Finding, Encounter and Organization. For each category of data (when available) the total number of records and patient counts is requested, along with the first year the category of data was generally available. The granularity of timestamps on the data is also recorded at this level of the survey. Timestamps can be accurate to the year, month, day and second. Finally, for data elements generally found within each category the availability, structure and semantics of the element are requested using two templates: one for the data element itself, and the other for one or more value sets associated with the data element.

1.7. Data Exchange Format

In EHR4CR, all participating hospital site agreed to expose their data in form of Clinical Data Warehouse (CDW) based on two different schema models: (i) i2b2 model and (ii) EHR4CR-CDW model. The EHR4CR-CDW model is based on the EHR4CR Information Model (see Section 1.2). Therefore, the supported data exchange format is UML, mainly based on «A_SupportingClinicalStatementUniversal» model, component of the StudyDesign, proposed by the HL7 Regulated Clinical Research Information Model (RCRIM). In addition, EC Model (see Section 1.5) is the data exchange format between the query workbench and CDWs.

1.8. Use-cases Description

Table 1. EHR4CR use-cases description

	Use cases	Description	Services
WP6	1- Protocol feasibility	Leverage clinical data to design viable trial protocols and estimate recruitment	Distributed queries over heterogeneous EHRs or CDWs
	2-Patient recruitment	Detect patients eligible for trials and better utilize recruitment potential	Distributed queries over heterogeneous EHRs or CDWs Workflow execution
	3-Clinical trial execution	Optimize clinical trial execution Re-use of clinical data to pre-populate eCRFs	Workflow execution Pre-population of forms (distributed queries over heterogeneous EHRs or CDWs)

	4-Pharmacovigilance	Detect adverse events and collect/transmit relevant information	Distributed queries over heterogeneous EHRs or CDWs Workflow execution Pre-population of forms
WP4	Semantic Resources & Terminology Services		
WP5	Access policy, pseudonymisation/de-identification, patient content services		

1.9. Current Issues

In course of developing EHR4CR semantic interoperability framework, we face several challenges: (i) formally defining patient eligibility criteria including temporal constraints, (ii) dealing with heterogeneity between different EHRs, (iii) defining mappings between data elements from eligibility criteria and patient data, and (iv) investigating standard query interfaces for retrieving patient information from heterogeneous EHRs. We also need to continue our efforts at harmonizing the EHR4CR Information Model, common data elements and terminology to other standard-based semantic resources including FHIR, BRIDG, CDISC SHARE (and other meta data repository initiatives such as caDSR, openMDR, eMERGE) and the Ontology of Clinical Research (OCRe).

2. EURECA Project: Semantic Interoperability Approach

2.1. Project Summary

EURECA aims to build an advanced, standards-based and scalable semantic integration environment enabling seamless, secure and consistent bi-directional linking of clinical research and clinical care systems to:

1. Support more effective and efficient execution of clinical research by allowing faster eligible patient identification and enrolment in clinical trials, providing access to the large amounts of patient data, enabling long term follow up of patients, and avoid the current need for multiple data entry in the various clinical care systems.
2. Allow data mining of longitudinal EHR data for early detection of patient safety issues related to therapies and drugs that would not become manifest in a clinical trial either due to limited sample size or to limited trial duration,
3. Allow for faster transfer of new research findings and guidelines to the clinical setting (from bench-to-bedside),
4. Enable healthcare professionals to extract in each patient's case the relevant data out of the overwhelmingly large amounts of heterogeneous patient data and treatment information.

At the core of the project will be achieving semantic interoperability among EHR and clinical trial systems, consistent with existing standards, while managing the various sources of heterogeneity: technology, medical vocabulary, language, etc. This requires the definition of sound information models describing the EHR and the clinical trial systems, and capturing the semantics of the clinical terms by standard terminology systems. The scalability of the solution will be achieved by modularization, identifying core data subsets covering the chosen clinical domains. We demonstrate and validate concepts developed in EURECA by implementing a set of software services and tools that we deploy in the context of pilot demonstrators. EURECA will develop solutions that fulfill the data protection and security needs and the legal, ethical and regulatory requirements related to linking research and EHR data.
EURECA started on February 1, 2012.
Project website: http://eurecaproject.eu/

2.2. Information Models, Meta-data Repositories

The EURECA Common Data Model will be based on HL7 v3. It will contain clinical data from EHR and CT systems. A BRIDG-based database will be used for CT metadata.

2.3. Clinical Terminologies/Ontologies

The EURECA semantic core dataset will consist of a well-defined set of domain concepts that sufficiently describe the semantics of the chosen clinical domain. This core dataset is currently being defined based on an inventory of vocabularies used by the project partners (see Figure below), as well as terms and concepts that occur in the datasets of these partners.

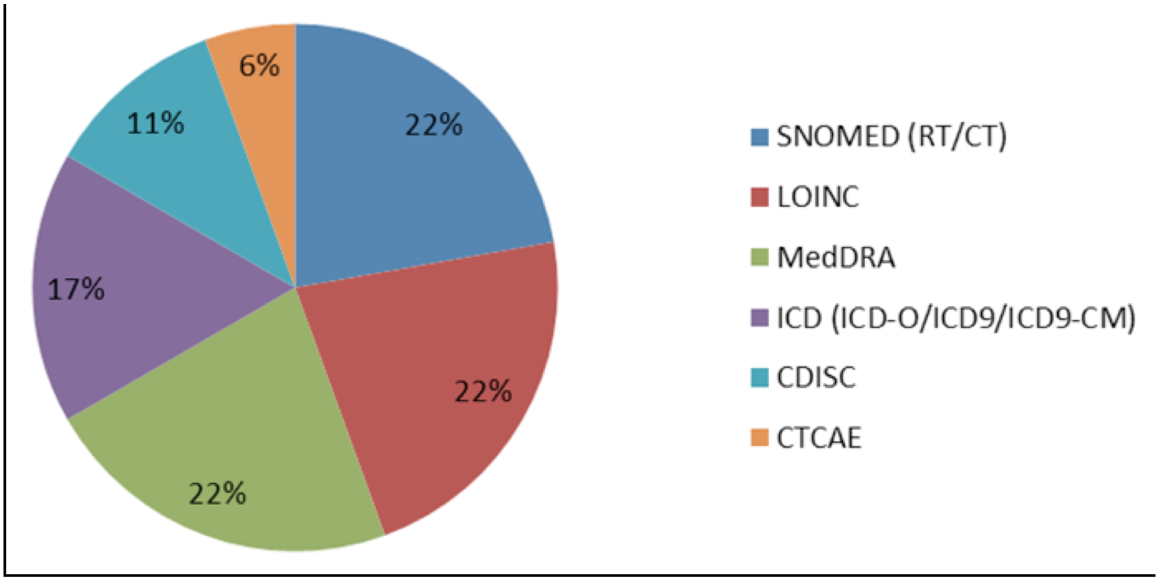

Figure 1 - Use of medical terminologies among interviewed users

2.4. Terminology Services (Management & Mappings)

This semantic data set will be mapped to concepts from relevant existing standardized terminologies or existing mappings will be made available in the EURECA platform. We are currently investigating the use of Bioportal mappings within the project.

2.5. Query Language & Query Builder (GUI)

Semantic layer:

- The Semantic Interoperability Layer provides a query interface, which supports the SPARQL query language.
  Queries can be directed toward the CIM (Common Information Model), i.e. semantically enriched, or to the CDM (Common Data Model), i.e. RIM based data structure.

Query languages (on-top of semantic layer):

- Because of the rich CIM, some EURECA solutions will directly script queries based on an API to the query endpoints of the semantic integration layer.
- EURECA will also provide the “Snaggletooth Query Engine”, which a supports a dynamic query DSL.
- Additionally a GUI will be provided to easily construct queries for the Snaggletooth engine.

2.6. EHR Data Quality and Preparation

EURECA has several clinical participants, each with different information systems: Jules Bordet Institute, MAASTRO Clinic, Breast International Group, German Breast Group, University of Oxford, University of Saarland. The clinical partners use many different clinical information systems and EHR systems, including systems that they have developed themselves. Some Clinical Trials Systems include: Obtima, OpenClinica, Oracle Clinical, Cerner Millennium EPR, and a Computer Aided Theragnostics (CAT) data warehouse developed by MAASTRO Clinic.

Several issues arise in EHR data preparation for sharing with international partners, including data quality, de-identification, informed consent, and support for formats such as HL7 v2. Also, medical terminology is in the local language of each clinical partner (French, Dutch, German, English). Data quality is a significant challenge in the medical environment. Missing or ambiguous data is a common problem faced by clinical researchers, especially those wishing to perform mutli-centric and retrospective studies. Unreleased (from commercial vendors) or changing data schemas also pose a barrier to data collection and retrieval.

Clinical trial eligibility is a use case common to several of the EU projects in the meeting. In the case of newly admitted cancer patients, the oncology clinic must attempt to place eligible patients into a trial as quickly as possible, i.e. before the patient undergoes treatment and usually before the patient's data has been entered into an EHR. Trial physician assistants must often search through a variety of free text to evaluate patient eligibility for trials at an early stage of 'pre-admission' – free text often resulting from OCR (Optical Character Recognition) scans of regional hospital reports and letters.

Some of the current plans within EURECA are to create RDF representations of clinical trial eligibility criteria, use those criteria to scope the extraction and representation of patient data. In the case of non-English patient data, for example Dutch language data, we plan to use the corresponding language labels of concepts and their synonyms from terminologies such as SNOMED. In cases where other vocabularies have been used to annotate data, such as NCI Thesaurus at MAASTRO, we will employ mappings between those other vocabularies and SNOMED.

2.7. Data Exchange Format

Different types of data will be exchanged:

- Clinical data to be loaded into the Common Data Model will be in a HL7 CDA – based format, or in CSV format.
- For semantic information we will use RDF, RDFS, OWL.
- Query results will be returned in RDF (SPAQRL results) or in CSV (SQL results)

2.8. Use-cases Description

Information-related use cases:

- Personal medical information recommender
- Export from an EHR to a PHR system
- Data mining of consultation data
- Giving a contextualized overview of large amounts of data

Investigation-related use cases:

- Support for updates of guidelines
- Training, validation and update of a diagnostic classifier
- Protocol feasibility check

Selection & Recruitment-related use cases:

- Microbiology SAE
- Support for trial recruitment

Reporting-related use cases:

- Reporting episodes of febrile neutropenia
- Cancer registry and tumor bank reporting
- Pre-filling of CRF and AE reports
- Automatic detection and reporting of SAEs/SUSARs

Other use cases:

- Long-term follow-up of patients
- Economic analysis of procedures with respect to outcome and quality of life of an individual patient

2.9. Current Issues

The availability of patient data is difficult due to obvious privacy issues. The consortium, lead by the legal partner Leibniz University Hannover, are working on a data transfer agreement.

3. SALUS Project: Semantic Interoperability Approach

3.1. Project Summary

FP7-287800 SALUS Project, an R&D project co-financed by the European Commission's 7th Framework Programme (FP7), aims to create the necessary semantic and functional interoperability infrastructure to enable secondary use of EHR data in an efficient and effective way for reinforcing the post market safety studies.

The objectives can be summarized as follows:

- Strengthening the spontaneous reporting process by automated ADE detection tools screening EHRs in a hospital so that ADE reporting burden can be overcome within a clinical institute. This can increase data accuracy as it eliminates manual screening of clinical care data for identifying ADEs.
- Enabling ADE reporting by extracting the available information from the EHRs into the individual case safety reports to avoid double data entry. This ensures delivering timely feedback to the regulatory bodies via automatic EHR supported adverse event reporting.
- Strengthening the current signal detection processes in Spontaneous Reporting System (SRS) centers for tracing case reports to their corresponding patient records to allow actual incidence rates to be computed, and to provide additional information on extended parts of the underlying medical history of the patient.
- Enabling real time screening of multiple, distributed, heterogeneous EHRs for early detection of ADE signals. This facilitates proactive safety monitoring as a complementary approach to reactive signal detection based on spontaneous reports.
- Enabling sustainable and scalable EHR re-use facilitating wide scale outcome and effectiveness research, to be able to observe selected cohorts of patients over an extended period of time screening multiple, distributed, heterogeneous EHR systems.

R&D activities will be carried out along the following topics:

- Definition of standard based specifications of messages that will be exchanged through defined transactions between EHR systems and tools supporting post marketing safety analysis (WP4)
- Definition and validation of a core set of common data elements required for post market safety studies from various clinical information systems and EHR systems. Through such a data set, it will be possible to create a common understanding of the data requirements as meaningful fragments necessary to conduct patient safety studies (WP4)
- Development of SALUS semantic resource set as a set of ontologies based on core data sets, and aligning with the available domain ontologies and fragments of terminology systems (WP4). This will be named as SALUS common ontology.
- Development of semantic interfaces on top of existing EHR systems to be able to query them semantically (WP4)
- Development of semantic mediation mechanisms to address the different standard based approaches to represent clinical data (such as different HL7 CDA templates, different CEN/ISO 13606 archetypes) and different terminology systems used by EHR systems and intelligent data analysis tools by making use of SALUS common ontology as a common denominator (WP4)
- Definition and validation of a standardized protocol for both subscription based and query based interaction with EHR systems for secondary use of EHRs (WP5)
- Development of interoperability toolkits to create the standard based individual case safety reports in collaboration with the underlying EHRs and send them to regulatory bodies (WP5)
- Development of interoperability profiles and open source toolsets for ensuring the security and privacy of the clinical information shared among primary care and post market safety studies (WP5)
- Development of applications for the analysis of EHRs for early detection of safety issues during post marketing phase in a proactive manner. These applications will focus on temporal patterns and will be extendable toolkits for ADE detection and exploratory signal studies (WP6)

Project website: http://www.salusproject.eu/

3.2. Information Models, Meta-data Repositories

While collecting the medical summaries from underlying EHR Systems, we have chosen to comply with well-defined EHR interface standards, namely HL7 Clinical Document Architecture Release 2 (CDA) based templates, and ISO/CEN EN 13606 EHRExtract based archetypes and templates. On top of this, we will also allow EHR Systems to open up SPARQL endpoints to expose anonymized medical data sets. For this, we have developed a Data Definition Ontology on top of the ORBIS installation at UKD (University Clinic-Technical University of Dresden).

On the research side, each of post market safety analysis applications and methods may require to retrieve medical data sets in different formats. Based on our initial analysis of the selected use cases, Temporal Pattern Characterization, Temporal Association Screening and Patient History tools prefer to retrieve data in conformance to Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM); while ICSR Reporting tool will produce case safety reports in E2B(R2) specifications along with local models like the ICSR template provided by Italian Medicines Agency (AIFA).

These templates and data models used at the EHR side and by the tools to analyse them at the research side, will constitute the SALUS Content Model Library.

The first version of SALUS Content Model Library (see SALUS Deliverable 4.1.1- R1) covers two content models at the EHRs side, and two content models at the clinical research side (based on the pilot application requirements and our DoW):

- We have defined content model templates as HL7 CDA templates. One of our pilot sites, Lombardy Region will expose medical summaries as HL7 CDA documents compliant with well-known CCD and PCC templates.
- We have defined an archetype library and an EHRExtract template that makes use of these archetypes to represent medical data sets in ISO/CEN EN 13606 format. Although we will not have a physical pilot site that will produce and share medical data sets in EN 13606 format, based on the principles set in SALUS DoW, we have produced these templates and will demonstrate that SALUS System, in particular the semantic interoperability framework, is capable of processing these and can prepare medical data sets that can be consumed by SALUS clinical research applications.
- As several of the clinical research applications that will be developed in SALUS pilots would like to receive medical datasets in OMOP CDM format, OMOP CDM will be used as a content model.
- SALUS ICSR Reporting Tool will support semi-automatic reporting of ADEs to regulatory authorities using the ICH E2B(R2) Electronic Transmission of Individual Case Safety Reports Message Specification. For this reason as one of the target content models, E2B(R2) model will be used.

These constitute our content models; based on the methodology set in SALUS DoW, we are examining the pilot requirements and these content models to identify the required Common Data Elements (CDEs), which will constitute the basis of SALUS semantic resource set as a common dictionary of meaningful fragments to be exchanged between clinical care and research sites. These CDEs will be maintained in a Metadata Registry. In SALUS we are implementing an open source Metadata Registry based on ISO/IEC 11179 Metadata Registry standard. This MDR implementation is a semantic repository itself, the repository is maintained in a triple store, and we have created an ontology of ISO/IEC 11179 meta model. In this MDR, we will not only host the CDEs, but their mappings to the selected content models, which will be the basis of the semantic mediation rules between different content models and SALUS semantic resource set (aka SALUS Common Ontology).

3.3. Clinical Terminologies/Ontologies

First of all, we have developed a set of ontologies for some of the content models we have already developed:

- We built an ontology for HL7 CDA
- We built an ontology for OMOP CDM
- We built an ontology for ORBIS installation at UKD (University Clinic-Technical University of Dresden)

By examining the draft set of CDEs, we have built a draft version of SALUS Common Ontology that will be used as the common denominator for semantic mediation.

As terminologies, we have downloaded and fine-tuned the following terminology system ontologies from BioPortal:

- WHO-ART
- ICD-9-CM
- ICD-10
- MedDRA
- SNOMED CT Clinical Findings sub-hierarchy

On top of this, we have also built an RDF representation of WHO-ATC code system.

3.4. Terminology Services (Management & Mappings)

Terminology reasoning plays a very important role for achieving semantic interoperability in SALUS. We have quite complex and different terminology reasoning requirements in different pilot application scenarios. Terminology systems are also represented as ontologies (N3 format) which forms the majority of the SALUS Semantic Resource Set. Terminology reasoning will not only be carried out by interacting with Terminology Servers through terminology mapping services, as this does not allow us to use the full potential of terminology resources (where semantic relationships between code systems cannot be exploited fully). As described, terminology resources are being included to our semantic resource set as ontologies, and we are building rule based semantic reasoning methods on top of them, especially for analyzing the collected data sets for post market safety analysis studies. When necessary, we are planning to use query expansion, especially while querying subsets of medical summaries of the eligible patients from the underlying EHR systems.

3.5. Query Language & Query Builder (GUI)

We are building Web based graphical interfaces for expressing inclusion/exclusion criteria of the foreground and background populations of the post market safety analysis studies. These interfaces use SALUS common model as the basis to define criteria on top of them. For expressing these queries, we are building a semantic model, based on the formalism introduced in HL7 HQMF. We will share this model when it is ready. We will focus on template based queries, and inside SALUS semantic interoperability framework, these template based queries will be translated to the target query format of the EHR resources. In TUD site (pilot deployment 1), we will use a SPARQL endpoint on top of the local ontology of ORBIS UKD installation, and in LISPA side site (pilot deployment 1), we will use HQMF based population queries.

3.6. EHR Data Quality and Preparation and Data Exchange Format

In TUD site (pilot deployment 1), we will use a SPARQL endpoint on top of the local ontology of ORBIS UKD installation to retrieve the required EHR data. The data collected in local ontology (already in N3 format) will be translated to SALUS common model through semantic mediation tools.

In LISPA site (pilot deployment 2), we will interact with the local data warehouse through extended IHE QED and CM profiles (see SALUS Deliverables 5.1.1 and 5.2.1). The population-based queries will be expressed through HQMF, and the result sets will be shared through HL7 CDA based entry level templates defined in D4.1.1. These then will be translated into a semantic model, and then will be translated to SALUS common model through semantic mediation tools.

3.7. Use-cases Description

The following use cases have been selected by SALUS consortium (Details are available in SALUS D8.1.1)

- Enabling Semi-automatic Notification of Suspected ADEs and Reporting ADEs within a Hospital
  - Enabling Notification of Suspected ADEs
  - Enabling Semi-automatic ADE Reporting
- Supporting Clinical Evaluation of a Potential Signal through Accessing the EHRs
  - Characterizing the cases and contrasting them to a background population
  - Temporal pattern characterization

- Running Exploratory Analysis Studies over EHRs for Signal Detection
    - Temporal association screening on EHRs
    - Manual clinical review of relevant medical history
- Using EHRs as secondary use data sources for Post Marketing safety studies
    - Estimate incidence rates of chronic heart failure (CHF) in diabetic patients with a recent acute coronary syndrome (ACS) event on different diabetic medications

3.8. Current Issues

SALUS Project has initiated contact with IHE Quality, Research and Public Health Domain (QRPH), in order to work on new profile proposals that are of interest to SALUS project interoperability approach. In particular, Gokce B. Laleci (SRDC) will be one of the co-authors of IHE Data Exchange (DEX) Profile together with Landen Bain (CDISC). The aim of IHE DEX Profile is to exploit a metadata registry to annotate both eCRF or ICSR forms and also medical summaries (that may be represented in HL7 CCD) format with Common Data Elements maintained in a metadata registry, so that, interoperability between clinical research and care domains can be achieved on the fly by retrieving extraction specification of a certain data element in one domain from a standard document in another domain. This work is quite parallel with the improvements proposed in an early SALUS publication: "Providing Semantic Interoperability between Clinical Care and Clinical Research Domains", Laleci, G., Yuksel, M., Dogac, A., IEEE Transactions on Information Technology in Biomedicine; hence SALUS project will take active participation in the preparation of this profile.

4. OpenPhacts Project: Semantic Interoperability Approach

4.1. Project Summary

To reduce the barriers to drug discovery in industry, academia and for small businesses, the Open PHACTS consortium is building the Open PHACTS Discovery Platform. This will be freely available, integrating pharmacological data from a variety of information resources and providing tools and services to question this integrated data to support pharmacological research.

Project website: http://www.openphacts.org/

4.2. Information Models, Mete-data Repositories

We use the Vocabulary of Interlinked Datasets (VoID) to describe all our datasets as well as mappings. Please see http://www.openphacts.org/specs/datadesc/ for a complete specification of how this is done. In terms of metadata, management

4.3. Clinical Terminologies/Ontologies

We use a wide variety of ontologies. Please see http://www.openphacts.org/specs/rdfguide/ for a recommended list. These include expected vocabularies such as UMLS, ontologies from bioportal.

4.4. Terminology Services (Management & Mappings)

Please see http://www.openphacts.org/about-ops/201
We use two core services:
Concept Wiki http://ops.conceptwiki.org/wiki/ - for term to identifier mapping
Open Phacts Identity Mappings Service based on Bridge DB – for identifier to identifier mapping.

4.5. Query Language & Query Builder (GUI)

We use the Linked Data API (http://code.google.com/p/linked-data-api/) to present a uniform API for data access. This is backed by both sparql queries and calls to web services. Queries are written after use case and interaction with multiple application builders.

4.6. EHR Data Quality and Preparation

All answers to queries are provided with a complete provenance trace. For chemistry information, we have a structure validation and standardisation platform has been developed to ensure normalisation of chemical structures to rules derived from the FDA structure standardisation guidelines and modified based on input from the EFPIA members.

4.7. Data Exchange Format

Standard serializations of RDF. Turtle is the encouraged serialization format.

4.8. Use-cases Description

Open PHACTS links the large amounts of physicochemical and pharmacological data available in public databases, and provides a means of querying this via the Open PHACTS Explorer. The number of pharmacological questions that could foreseeably be useful to answer is large, and Open PHACTS concentrates on answering the top 20 ranked research questions from a list of 83 proposed by consortium members. These questions can be grouped as Cluster I and Cluster II. The first cluster asks basic pharmacology questions, which are typically asked in the early stages of drug discovery, regarding interactions between a compound or compound group and defined targets. The second cluster asks questions of compound-target interactions, but also extends the query to pathways and diseases. Such questions typically require associated references as they are useful in the lead optimisation phase or for proof of concept studies. Another important concept in the Open PHACTS Discovery Platform is that of data provenance. Allowing the identification of data origins is a vital factor in developing end-user trust.
The list of the top 20 use case questions can be found at:
http://www.openphacts.org/about-ops/200

5. Linked2Safety Project: Semantic Interoperability Approach

5.1. Project Summary

The vision of Linked2Safety is to advance clinical practice and accelerate medical research, to improve the quality of healthcare, benefiting public health, and to enhance patients' safety; by providing pharmaceutical companies, healthcare professionals and patients with an innovative semantic interoperability framework, a sustainable business model, and a scalable technical infrastructure and platform for the efficient, homogenised access to and the effective, viable utilization of the increasing wealth of medical information contained in the EHRs deployed and maintained at regional and/or national level across Europe, dynamically interconnecting distributed patients data to medical research efforts, respecting patients' anonymity, as well as European and national legislation. The Linked2Safety project - with the developed reference architecture, data protection framework, common EHR schema, lightweight semantic model and integrated platform - will facilitate the scalable and standardised semantic interlinking, sharing and reuse of heterogeneous EHR repositories. This in turn will provide healthcare professionals, clinical researchers and pharmaceutical companies' experts with a user-friendly, sophisticated, collaborative decision-making environment. This will allow analysis of all the available data of the subjects, such as genetic, environmental and their medical history during a clinical trial leading to the identification of the phenotype and genotype factors that are associated with specific adverse events and thus early detection of potential patients' safety issues. It will also enable subject selection for clinical trials through the seamless and standardized linking with heterogeneous EHR repositories, providing advice on the best design of clinical studies.

Project website: http://www.linked2safety-project.eu/

5.2. Information Models, Mete-data Repositories

It is expected that Linked2Safety will have multiple outcomes. One of them is the open, generic Linked2Safety Reference Architecture for enabling the reuse of semantically interlinked, interoperable EHR and Electronic Data Capture (EDC) information resources in clinical trials design and execution advancing proactive patient safety and targeted patients selection. The Linked2Safety project covers healthcare standards information model (e.g., OpenEHR, HL7). One of the key tasks in the Linked2Safty project is to ontologise the standard specific information models.

5.3. Clinical Terminologies/Ontologies

The Linked2Safety project built a Semantic EHR (SEHR) ontology, a light-weight and extensible ontology that covers multiple sub-domains of Healthcare and Life Sciences (HCLS) through specialisation of the upper-level Basic Formal Ontology (BFO). The goal of building the Semantic EHR Model is to enable seamless sharing and linking pieces of healthcare, i.e., Electronic Health Records (EHRs) and clinical data/knowledge among the authorised stakeholders. The Semantic EHR Model has a crucial role of sharing consistent knowledge for decision making in medical and clinical research domains. The Semantic EHR Model is a core pillar of the semantically-interlinked Linked2Safety Infrastructure. Further, a common EHR ontology developed from reusing the standard artefacts are aligned with the SEHR.

5.4. Terminology Services (Management & Mappings)

The Linked2Safety project performs mapping/alignment at two levels (1) Instance Level: To respect patient safety, the Linked2Safety project represents clinical data in an anonymised and aggregated multidimensional data-cubes. Therefore, all the data-cubes generated by clinical partners will be semantically interlinked, such as providing typed links between resources in the data-cubes and also to establish links between the datasets internal to the project and the external medical datasets found on the Linked Data Cloud; and (2) Schema Level: Different sets of terminologies originating from diverse and disparate clinical partners are aligned and consolidated the schema level. Some of these clinical terminologies comply with standard medical vocabularies (e.g., SNOMED, ICD-10) and many of them are created locally as per the clinical requirement.

5.5. Query Language & Query Builder (GUI)

The Linked2Safety Platform interface will include a query builder, which will use concepts from the SEHR, and will guide the those who are not SPARQL experts through the process of building a query which federates across the clinical data sets. This is in addition to an envisaged REST API which will also reflect the L2S ontology and allow a developer to program against the Linked2Safety Platform. Additionally, we will develop a less technical interface which will allow the visual exploration of the ontology. This visualisation/navigation will result in the generation of a SPARQL query in the background, again working against the query engine. The only thing an user will require is knowledge and understanding of the model which describes the data – the SEHR.

5.6. EHR Data Quality and Preparation

In Linked2Safety a concept of a "closed-world" room is introduced (a room located within a data provider's premises, featuring the required hardware infrastructure to process EHRs isolated from any kind of network connections). The physical access to this machinery within the room is allowed only to specific personnel of the corresponding data provider and it is off line to the outside world. The data provider's staff will execute a program on the computers located in the "closed-world" room that will aggregate the data generating the data cubes. This program will offer the option to the data provider to limit the way in which the data will be aggregated so that any legal and ethical issues can be addressed. The Linked2Safety consortium has decided to employ a data-cube approach to address the ethical requirements of handling sensitive patient data, namely: respecting patients' anonymity, data ownership and privacy, as well as compliance with legislative, regulatory and ethical requirements. In terms of quality control, raw data are initially loaded and then four quality control tests are performed on them. Initially, the subjects' gender test is performed so that any subjects with erroneous gender values are removed from the dataset. Then the missing data test is performed on the dataset so that any subjects with a missing rate of values above a predefined threshold are removed and then any variables with a missing rate above the predefined threshold are also removed. The output of this test is then used in the Hardy-Weinberg Equilibrium test so that any SNPs that do not conform to it are removed from the dataset. Finally, the output dataset from the previous test is used in the raw data Allele frequency test, where any SNPs with a minor Allele frequency below a predefined threshold are removed. The RDFizer component with the Linked2Safety platform is responsible for converting the aggregated data in the form of data cube, to an RDF version of the same data while using the RDF Data Cube Vocabulary.

5.7. Data Exchange Format

All data and schema used within the project uses Semantic Web technologies (RDF, OWL, SPARQL) as the data exchange standard.

5.8. Use-cases Description

A phase III clinical trial coordinator accessingLinked2Safety can identify the number of subjects that match his/her selection criteria from all the relevant data sources in the Linked2Safety platform. Then he/she can contact the Linked2Safety Governance body for help in accessing the subjects needed to include in the trial. The Governance body will review the research proposal and will decide if, and which of the subjects can be invited to volunteer for the trial. The Governance body will simply act as a mediator between the clinical trial coordinator requesting the access, and the principal investigator(s) of the study(ies). Subjects will only be conducted by the institute they have signed the consent form with. All ethical, legal or other issues will then need to be addressed directly between the phase III clinical trial coordinator requesting the data and the institution(s) that actually holds the data.

5.9. Current Issues

AT the moment project is dealing with the issue of federating of queries over multiple distributed repositories and employing policy based access to the restricted clinical resources.

6. eTRIKS Project: Semantic Interoperability Approach

6.1. Project Summary

eTRIKS is a knowledge management and service infrastructure project aimed at development of a software and hardware system capable of the efficient storage and effective analysis of experimental data from studies in man, in animals and in pre-clinical models, maximising the scientific knowledge that can be extracted from such studies. The project's primary goal is to deliver a knowledge management system for ongoing and future IMI studies that require correlative analysis of both pre-clinical and clinical genome-scale biomarker data (genetics and genomics platforms) in conjunction with medical data from clinical trials. This open-source system will also be available for use outside of projects sponsored by IMI. Our overall aim in "Delivering eTRIKS" is to drive and support the innovation in European Translational Research, with the following clear objectives:

1. Service:
    a. Deploy and host the eTRIKS platform based on the tranSMART technology to provide an integrated service that is fit-for-purpose, secure, easy to access, standardised to support TR KM in IMI (and other translational research) projects in Europe, and that is sustainable in the long-term beyond the project duration.
    b. Provide training, support and consultation activities to all IMI project partners on using the eTRIKS service and platform, with particular emphasis on data security and privacy based on ethical guidelines.
2. Platform:
    a. Develop and maintain the eTRIKS platform built upon tranSMART as a sustainable, interoperable, collaborative, re-usable, open source and scalable TR KM platform adhering to agreed standards and used, and contributed to, by the global research community.
    b. Conduct research & development into effective analytics methods and tools to support TR. Evolve and extend the eTRIKS platform with a rich set of analytical methods and tools for omics, imaging data and text that can leverage cloud-based operations.
3. Content:
    a. Establish eTRIKS as a unique European TR data resource supporting cross-organisation TR studies, including clinical studies and pre-clinical studies, omics data analysis for biomarker discovery and validation, genetics and NGS studies. Incorporate pertinent reference and background data into eTRIKS (eg. molecular pathway data, scientific literature, etc.).
    b. Populate eTRIKS with existing and active data from TR studies and supporting the integration of standardised legacy TR study data.
4. Community:
    a. Promote and lead an active international TR analytics & informatics community, centred around eTRIKS, through active stakeholder engagement and by disseminating tools and expertise worldwide.
    b. Engage in, and influence, international standardisation activities in areas relating to TR informatics.

Project website: http://www.etriks.org/

6.2. Information Models, Meta-data Repositories

eTRIKS does not offer any specific electronic document capture (EDC) solution due to the variability of available platforms used by the supported projects; it would be unreasonable to expect clinical centres with established EDC systems to adopt a new technologies.

The current eTRIKS approach is to understand the data model/structure of the supported project's EDC, convert all data into a common format, tag metadata and organise it into ontologies, namely CDISK/SDTM[1] and i2b2[2].

Data organised in CDISK/SDTM ontology is stored in an ontology repository to enable data querying and ensure data legacy, whilst data organised in i2b2 ontology is loaded into the transMART platform[3]. An example i2b2 data structure is represented in Figure 2.

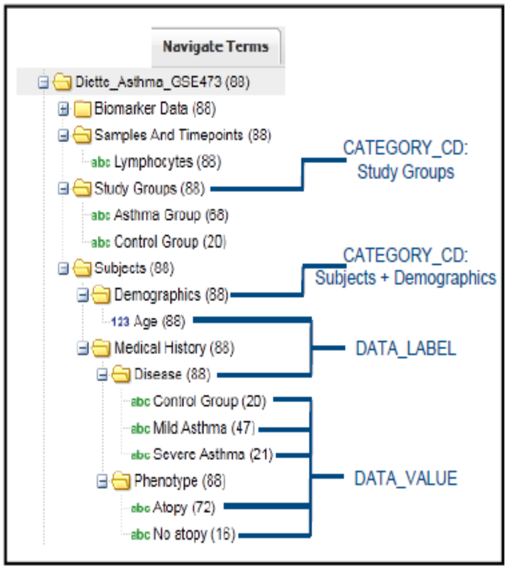

Figure 2: i2b2 datastructure.

6.3. Clinical Terminologies/Ontologies/Standards

- i2b2
- CDISC
  - SDTM: Study Data Tabulation Model
  - ADaM: Analysis Data Model
  - SEND: Standard for Exchange of Nonclinical Data
  - CDASH: Clinical Data Acquisition Standards Harmonization
  - PR: Protocol Representation
  - TDM: Study/Trial Design Model
  - ODM: Operational Data Model
  - LAB: Laboratory Data Model
  - BRIDG Model
- UMLS (Vocabularies include: CPT®, ICD-10-CM, LOINC®, MeSH®, RxNorm, and SNOMED CT®)
- CTSA
- NCBO Human related Phenotype ontology
- Omics ontology

[1] http://www.cdisc.org/sdtm

[2] https://www.i2b2.org/

[3] http://www.transmartproject.org/

6.4. EHR Data Quality and Preparation

eTRIKS comprises of two specific work packages, focused on data standards (WP3) and data curation (WP4). WP3 role is to study the ontology and data capture procedure used by the supported project and build a project specific STDM and i2b2 ontology. WP4 then proceeds by converting the data into common formats and mapping it to the relevant fields, denoted in the ontologies built by WP3, tags metadata and performes QC. Once processed and check the data is then stored in the ontology repository and loaded into transMART (Figure 3).

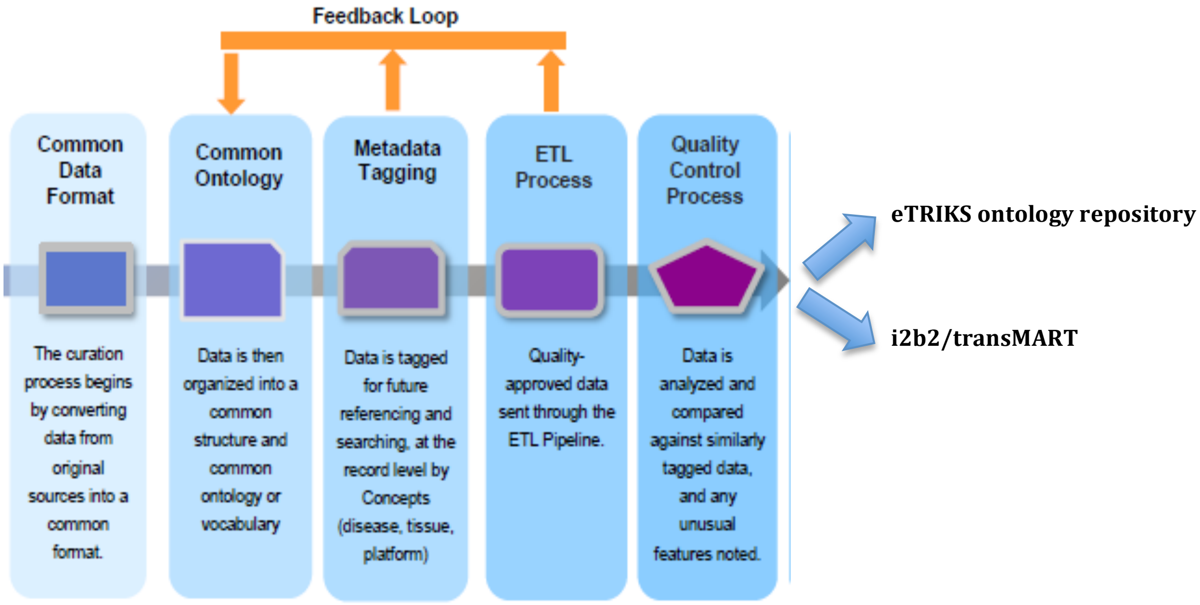
Feedback Loop
Common Data Format
Common Ontology
Metadata Tagging
ETL Process
Quality Control Process
The curation process begins by converting data from original sources into a common format.
Data is then organized into a common structure and common ontology or vocabulary
Data is tagged for future referencing and searching, at the record level by Concepts (disease, tissue, platform)
Quality-approved data sent through the ETL Pipeline.
Data is analyzed and compared against similarly tagged data, and any unusual features noted.
eTRIKS ontology repository
i2b2/transMART

Figure 3: eTRIKS clinical data preparation

6.5. Data Exchange Format

The BRIDG model is being considered to facilitate data exchange.

6.6. Use-cases Description

U-BIOPRED (Unbiased BIOmarkers in PREDiction of respiratory disease outcomes) an IMI research project to understand more about severe asthma.

The projects main objectives are:

1. Reaching international consensus on diagnostic criteria
2. Creating adult/pediatric cohorts and biobanks
3. Creating novel biology 'handprints' by combining molecular, histological, clinical and patient-reported data
4. Validating such 'handprints' in relation to exacerbations and disease progression
5. Refining the 'handprints' by using preclinical and human exacerbation models
6. Predicting efficacy of gold-standard and novel interventions
7. Refining the diagnostic criteria and phenotypes
8. Establishing a platform for exchange, education and dissemination

To facilitate U-BIOPRED clinical research several platforms/solutions have been implemented.

- **U-BIOPRED ontology**, a respiratory disease tailored ontology.

- **Clinical Data de-identification protocols.**
- Clinical **data transformation protocols** to organise data captured using the Actide eCRF (Nubilaria)[4], into the U-BIOPRED, as well as i2b2.
- **Web-based knowledge management portal**, facilitating collaboration through hypothesis building, analysis method development, analysis result saving and reporting and searching.
- **Cloud-based Transmart Instance**, hosting all data being produced as well as relevant legacy datasets.

6.7. Current Issues

- Building relevant data models for a variety of supported projects / biomedical research fields.
- Development of custom ETL procedures for legacy data for certain projects / studies, where data is poorly structured and inadequately documented.

[4] http://www.nubilaria.com/prodotti-e-soluzioni/medical/clinical-trials-studi-clinici/actide/

Table 2. Projects Analysis & Comparison Summary

Projects	Information Models (MDRs)	Clinical Terminologies/Ontologies	Terminology Services (Management & Mappings)	Query Language & Query Builder (GUI)	EHR Data Quality and Preparation	Data Exchange Format	Current Issues
EHR4CR	HL7 v3 models «StudyDesign»* CDA , CDISC IHE Profiles	LOINC, ATC, ICD, SNOMED-CT, PathLex MedDRA	Mapping Browsing Searching Expansion Management	Eclectic/ OCL/ SPARQL Template-based	EHR→ETL → CDWs	HL7 v3 models i2b2 model EC model	Dealing with clinical data structures templates, data elements
EURECA	Common Data Model based on HL7 v3, BRIDG, and EURECA Core dataset	Definition of a '*Core dataset*' with relevant concepts from ontologies used by partners	-Reasoning -Mapping (probably Bioportal mappings)	-SPARQL endpoint -Snaggletooth Query Engine -GUIS	Probably structured as well as unstructured data/plain text	RDF(S), OWL, HL7 CDA, CSV.	Availability of patient data.
SALUS	CDA, OMOP ODM, CDISC ISO EN 13606, ICH E2B(R2) IHE Profiles	SNOMED-CT MedDRA WHO-ART ICD-10 ICD-9-CM -SALUS Common ontology -HL7 CDA Ontology -OMOP CDM Ontology -UKD Data definition ontology -CDE Ontology (MDR ISO 11179)	Reasoning Convergence rules	SPARQL Rule-based calculations (Temporal Constraints), HQMF Queries	-SPARQL endpoints on EHRs -Extended IHE QED and CM Interfaces on top of EHRs	-RDF -CDA entry level templates	Interfacing with DEX, IHE profiles other standard information models
OpenPhacts	RDF with Dataset descriptions using VOI	BioPortal sources UMLS, ChEBI, etc VOID, QUDT PROV	Mappings Reasoning Provenance Curation	Application-specific GUIs Using Restful Services/API	Provenance is tracked through out	RDF in standard serializations (turtle)	Data updates in particular with respect to changing ontologies
Linked2Saftey	HL7/openEHR	DSM-4, SNOMED, LOINC,	Aligning global (BFO, DSM-4, SNOMED) and local terminologies	SPARQL	RDF Data-Cubes, SPARQL endpoints on EHRs.	RDF/OWL	Data-cube building SIG recruitment Legal ethical framework and exploitation strategy
eTRIKS	Common Data format → i2b2, CDISK/SDTM	i2b2, CDISK, UMLS, CTSA, NCBO, Omics Standards	eTRIKS ontology management service Storage		Any data → ETL → i2b2, CDISC/SDTM	Considering BRIDG	Variable data models, poorly structured legacy data

*«A_SupportingClinicalStatementUniversal»