



ICT-2011-288048

EURECA

**Enabling information re-Use by linking clinical
Research and CAre**

IP

Contract Nr: 288048

**Deliverable: D1.3 Report on state of the art on relevant
knowledge and data sources and on reusable tools**

Due date of deliverable: (02-28-2013)
Actual submission date: (MM-DD-YYYY)

Start date of Project: 01 February 2012

Duration: 42 months

Responsible WP: UdS

Revision: <outline, **draft**, proposed, accepted>

Project co-funded by the European Commission within the Seventh Framework Programme (2007-2013)		
Dissemination level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Service	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (excluding the Commission Services)	

0 DOCUMENT INFO

0.1 Author

Author	Company	E-mail
M. Scott Marshall	MAASTRO	scott.marshall@maastro.nl
Andre Dekker	MAASTRO	andre.dekker@maastro.nl
Cyril Krykwinski	IJB	cyril.krykwinski@bordet.be
Kristof De Schepper	Custodix	kristof.deschepper@custodix.com
Cédric Vansuyt	Custodix	cedric@custodix.com
Raúl Alonso	UPM	ralonso@infomed.dia.fi.upm.es
Juan Manuel Moratilla	UPM	jmmoratilla@infomed.dia.fi.upm.es
Haridimos Kondylakis	FORTH	kondylak@ics.forth.gr
Lefteris Koumakis	FORTH	koumakis@ics.forth.gr
Jasper Van Leeuwen	Philips	jasper.van.leeuwen@philips.com
Ronald Siebes	Stoneroos	ronald.siebes@stoneroos.nl
Holger Stenzhorn	UdS	holger.stenzhorn@uniklinikum-saarland.de
Sheng Yu	UOXF	sheng.yu@oncology.ox.ac.uk
Francesca Buffa	UOXF	francesca.buffa@oncology.ox.ac.uk
Laura Hollink	VUA	l.hollink@vu.nl
Zhisheng Huang	VUA	huang.zhisheng.nl@gmail.com

0.2 Documents history

Document version #	Date	Change
V0.1	09.01.2013	Starting version, template (based on D1.2)
V0.2	16.01.2013	Definition of ToC
V0.3	15.02.2013	First complete draft
V0.4	01.03.2013	Integrated version (send to WP members)
V0.5	04.03.2013	Updated version (send PCP)
V0.6	05.03.2013	Updated version (send to project internal reviewers)
Sign off		Signed off version (for approval to PMT members)
V1.0		Approved Version to be submitted to EU

0.3 Document data

Keywords	
Editor Address data	Name: M. Scott Marshall Partner: MAASTRO Address: Dr. Tanslaan 12, 6229 ET Maastricht Phone: +31 (0)88 44 55 666 E-mail: scott.marshall@maastro.nl
Delivery date	

0.4 Distribution list

Date	Issue	E-mailer

Table of Contents

0	DOCUMENT INFO	2
0.1	Author	2
0.2	Documents history	2
0.3	Document data	2
0.4	Distribution list	3
1	INTRODUCTION	6
2	CLINICAL SOURCES, TOOLS AND SERVICES (ALREADY USED BY CLINICAL PARTNERS)	8
2.1	Clinical trial systems	8
2.1.1	OBTIMA - ONTOLOGY-BASED TRIAL MANAGEMENT APPLICATION	8
2.1.1.1	CRF Creator	8
2.1.1.2	CRF Repository	9
2.1.1.3	Master Protocol Creator	9
2.1.1.4	Biobank Connector	9
2.1.1.5	Data Security	9
2.1.2	COMPUTER AIDED THERAGNOSTICS (CAT) RESEARCH PORTAL	9
2.1.3	ORACLE CLINICAL	10
2.1.4	OPENCLINICA	10
2.2	Electronic Health (Patient) Record systems	11
2.2.1	EHR AND EURECA'S CLINICAL PARTNERS	11
2.2.2	CERNER MILLENNIUM EPR IN OXFORD	12
2.2.3	MIRTH CONNECT, AN HL7 MESSAGING ENGINE AND INTEGRATION TOOL	14
3	REUSABLE TOOLS	15
3.1	Previous and running EU funded projects	15
3.1.1	RELEVANT TOOLS FROM INTEGRATE	15
3.1.1.1	INTEGRATE Semantic Interoperability Layer	15
3.1.1.2	Trial metadata repository of INTEGRATE	16
3.1.2	SEMANTICCT	17
3.1.3	SEMANTIC PLATFORM LARKC (LARGE KNOWLEDGE COLLIDER PROJECT)	20
3.1.4	SWI-PROLOG	20
3.1.5	LINKED LIFE DATA (LLD)	21
3.2	EHR4CR Convergence Meeting, Paris, January, 2013	24
-	W3C HCLS Interest Group as meeting platform for similar projects	24
3.2.1	EHR4CR INNOVATIVE MEDICINES INITIATIVE	24
3.2.2	SALUS	25

3.2.3	OPEN PHACTS INNOVATIVE MEDICINES INITIATIVE	26
3.2.4	LINKED2SAFETY.....	26
3.2.5	ETRIKS INNOVATIVE MEDICINES INITIATIVE.....	27
3.3	Other related projects	27
3.3.1	W3C HEALTH CARE AND LIFE SCIENCES INTEREST GROUP	27
3.3.2	PENTAHO DATA INTEGRATION (KETTLE).....	28
3.3.3	EHEALTHMONITOR	29
3.3.4	D2R SERVER.....	30
3.3.5	DR EYE	31
3.3.6	SMART & INDIVOX.....	32
3.3.7	THE SPECIALIST NLP TOOLS AND MEDICAL TEXT PROCESSING TOOLKIT	33
3.4	Security en Privacy Enhancing Tools.....	34
3.4.1	PIMS	34
3.4.2	CATS.....	35
3.4.3	SHIBBOLETH.....	36
3.4.4	SECURITY TOKEN SERVICE.....	36
3.4.5	A XACML ENGINE	37
3.4.6	LDAP.....	37
4	EXTERNAL DATA SOURCES AND KNOWLEDGE BASES	39
4.1	Sources of Patient Data	39
4.1.1	CYPRESS	39
4.1.2	SYNAPSE	39
4.1.3	CLINICAL AVATARS.....	40
4.1.4	ADVANCED PATIENT DATA GENERATOR (APDG).....	40
4.2	ClinicalTrials.gov	42
4.2.1	DATA IN THE REGISTRY	43
4.2.2	LINKEDCT.....	44
4.2.3	CLINICALTRIALSREGISTER.EU	44
4.3	BioPortal SPARQL endpoint.....	44
4.4	Trial feasibility data sources	46
4.4.1	LITERATURE	46
4.4.2	CANCER REGISTRIES.....	47
4.4.3	OTHER RELEVANT RESOURCES	47
5	CONCLUSION	48

1 Introduction

Previous deliverables *D1.1 “User needs and specifications for the EURECA environment and software services”* and *D1.2 “Definition of relevant user scenarios based on input from users”* have defined user needs and scenarios, along with a corresponding set of technical use cases. This deliverable catalogues existing data and knowledge resources and tools that can be used to implement applications for EURECA use cases and otherwise inform the design of user interfaces, user interaction and data exchange. The resources briefly described here will include such resources as data, or software APIs and libraries, as well as sources of expertise about key processes, standards and technologies, such as the expertise of a main actor in a scenario (e.g. Trial Physician Assistant who evaluates patient eligibility for clinical trials). Where possible, these resources will be related to the use cases. Of course, several types of resources are discussed in other deliverables and will therefore receive less attention here. For example, vocabularies and ontologies that are already being considered for the CDM (Common Data Model) have been described in previous deliverables, such as *D4.1 “Requirements analysis and selection of the initial clinical scenarios for core datasets”*, or will be documented in upcoming deliverables, such as *D9.2 “Canonical models of EHRs and CT systems”*, which will provide more details about how some EHR and CT systems are used in the clinical environment.

EURECA’s goal of connecting clinical care with clinical research will effectively require information systems from the health care and research domains to interoperate, i.e. for information to be exchanged and integrated across those domains. For this reason, we are describing several systems that EURECA clinical partners currently use for either clinical trial management or electronic health record management. If our clinical partners can find practical ways to connect systems that already form an integral part of their internal processes, it could ensure a seamless deployment of EURECA with non-invasive, incremental changes that cause minimal disruption to existing systems and processes. Such a connection involves not only conversion between import and export formats but also identifying the ways to reconcile data between those systems and the EURECA framework. Initial efforts focus reconciliation of data from clinic trial and EHRs with the EURECA CDM. The reconciliation is currently done with mapping scripts that convert serialized data documents. This process could eventually be accomplished by interacting directly with data and knowledge bases through SPARQL endpoints and database connectors.

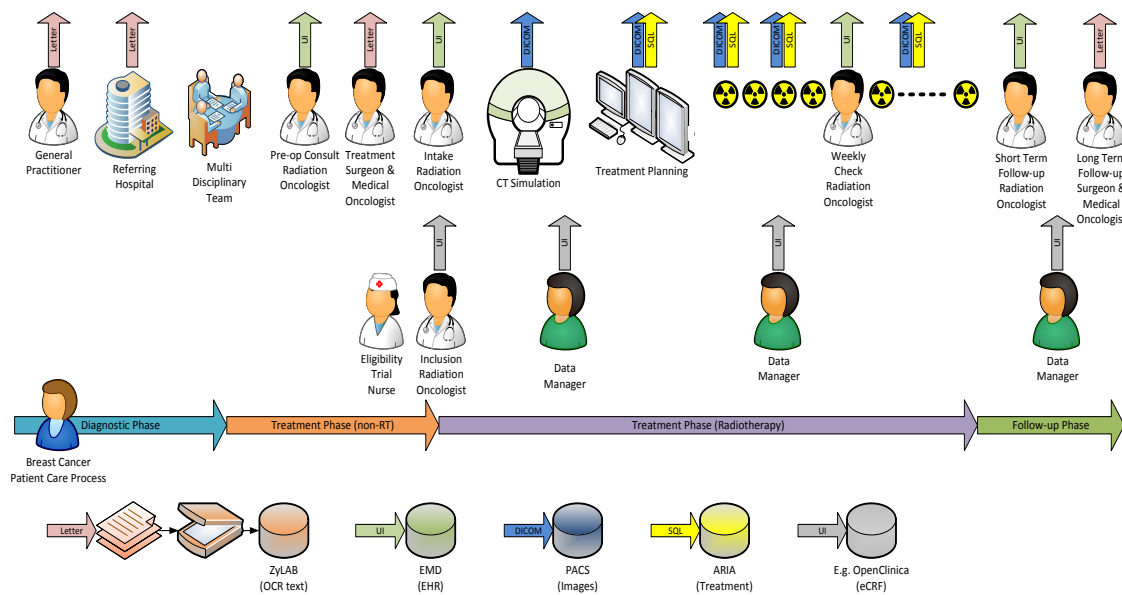


Figure 1-1: MAASTRO care process and data sources. Top: Radiotherapy care process including data generation (arrows). Top-Middle: Additional data generation if patient is included in a clinical trial. Bottom-Middle: Chronological treatment phases (not to scale). Bottom: Clinical data sources to be used in EURECA.

One of the critical challenges to EURECA is the acquisition of data attributes specifically required for some use cases, some of which might not be available – neither in structured data sources nor in unstructured data sources such as the natural language (non-coded) documents and fields typical of medical information systems. When data is available, it is often in unstructured free text documents and the challenge becomes how to reliably identify the presence of such information and extract it in a specific context. When it cannot be found, the challenge is how to proceed with missing data. This challenge has been encountered during the first year of EURECA as technology partners request specific types of data from clinical partners. Although there has been much interest in national standards for EHRs that would presumably create enough uniformity to make data exchange feasible, such standards have not yet been adopted or implemented in most countries, i.e. the semantic content and format of EHR data has not yet been standardized. Clinical partners in EURECA must therefore rely on referring hospitals to provide initial data about the patient and collect the rest themselves. For example, for the MAASTRO Clinic, initial data arrives by post from referring regional hospitals in the form of (paper) printed documents, such as referral letters. Those documents must then be scanned and converted from an image back to digital text using Optical Character Recognition (OCR). For this reason, text mining, information extraction, and natural language processing (NLP) tools play an important role and a few of the tools already in use by EURECA partners will be mentioned here. Some partners are developing customized tools from this genre for application to, for example, English language specifications of clinical trial eligibility criteria as found at ClinicalTrials.gov. When extracting patient information relevant to a use case, clinical partners must detect that the information is available and extract it from documents in a local language such as Dutch, German, or French. The challenge of such multi-language activities is heightened by a lack of labels in the Dutch, German, or French languages for popular medical coding systems such as SNOMED CT. If available with other language labels, such codes could serve to automatically translate between data sets and their descriptions. Such multi-language support will eventually be crucial to adoption in each participating country because physicians document patient data in the local language.

2 Clinical sources, tools and services (already used by clinical partners)

EURECA has several clinical participants, each with different information systems and languages: Jules Bordet Institute, MAASTRO Clinic, German Breast Group, University of Oxford, and 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: Optima, OpenClinica, Oracle Clinical, Cerner Millennium EPR, and a Computer Aided Theragnostics (CAT) data warehouse developed by MAASTRO Clinic.

A number of clinical trial and electronic patient record (EPR) systems are already in use or planned for use by EURECA clinical partners. Some systems, such as Optima have been initially developed during previous EU projects and are being extended in EURECA, while others such as OpenClinica, have been developed as open source by a consortium. We will also describe a few commercial systems such as Oracle Clinical and Cerner Millennium EPR, which have already been deployed by certain clinical partners.

2.1 Clinical trial systems

2.1.1 ObTiMA - Ontology-based Trial Management Application

The ObTiMA¹ application is an ontology-based clinical trial management system that has been initially developed as a proof-of-concept within the ACGT (Advancing Clinico-Genomic Trials on Cancer) EU-project to highlight the various possibilities and advantages of an ontology-based management of clinical trial data.

In the EU-projects p-medicine and EURECA, the major goal for ObTiMA is now to develop the system further to reach an industry-level application readily usable in real clinical trials. An initial production-ready version of ObTiMA was finalized at the beginning of 2013 in order to employ it within the new trial "Improving Population Outcomes for Renal Tumours of Childhood (IMPORT)". Before the roll-out of the system planned for middle of 2013, it is currently being tested at the Saarland University Hospital by various end-user types, e.g. trial chairman, study nurse and data manager.

The design and development of ObTiMA follows a modular pattern with a set of core modules handling the foundational tasks of patient, user and (trial) administrative data management as well as providing all necessary security-related functionality and the possibility of integrating additional modules offering further functionality.

Some of the above mentioned core modules are described in the following subsections.

CRF Creator

With the help of the CRF Creator, a trial chairman can design case report forms, i.e. questionnaires to collect patient data, in electronic form. Recent work on this module focused e.g. on making the creation of complex questions and groups of such questions possible with special consideration for its ease of use.

¹ <http://obtima.org/>

CRF Repository

The CRF Repository is a centralized storage place – accessible directly from within the CRF Creator – where trial chairmen can share their own CRFs or retrieve CRFs previously stored by themselves or other trial managers to e.g. allow reuse of existing CRFs instead of recreating them from scratch.

Master Protocol Creator

This creator provides trial chairmen with a template-based, straightforward interface guiding them through the preparation of the Master Protocol which has to encompass – besides the actual trial description – all additional legal and ethical information and documentation pertaining to the trials execution.

Biobank Connector

The Biobank Connector, developed by Fraunhofer IBMT, makes it possible to manage biomaterial specimen and related data directly from within ObTiMA and link this to the clinical (treatment) data of the patients. It therefore becomes possible to access patient data stemming from different sources through one common access point.

Data Security

All patient data entered into ObTiMA is pseudonymised and encrypted on-the-fly using the industry-proven technology called Privacy Enhanced Storage Framework (PESF) developed and integrated by Custodix allowing for a data handling which conforms to all relevant legal regulations.

2.1.2 Computer Aided Theragnostics (CAT) Research Portal

Together with Siemens Knowledge Solutions, MAASTRO developed a Computer Aided Theragnostics (CAT) research portal, which extracts medical data from the connected systems via a synchronization manager (sync manager) and stores the data centrally in a data warehouse. The operational, patient-centric structure of the data sources is converted into a disease-centric structure suitable for research. In the MAASTRO radiotherapy department, the sync manager extracts data from various sources:

1. the electronic medical record (EMR), which is a database with both structured and unstructured data,
2. the RT Picture Archiving Communication System (PACS), consisting of diagnostic imaging and treatment DICOM RT data, e.g. treatment plans, predicted dose matrices, delineations, and digitally reconstructed radiographs for setup verification,
3. the Record and Verify system (R&V) containing the actual delivered treatment parameters.

The CAT Research Portal User Interface offers four core user functionalities: a Patient browser, Query builder, eCRF module, Sandbox upload, and XML export.

2.1.3 Oracle Clinical

Oracle Clinical² is a clinical data management system (CDMS) developed and designed by Oracle to provide data management and electronic data capture (EDC), as well as data entry and data validation functionalities to the clinical trial process, from clinical study protocol description to management. Oracle Clinical is currently used by the Jules Bordet Institute for the management of most of its larger studies.

The major functions supported by Oracle Clinical are:

- Clinical study protocol definition and management
- Definition of metadata collected during a clinical study
- Security and administration of the access of study data for the users
- Creation of data entry system
- Creation of data management system to clean and reconcile data
- Validation procedures
- Data loading and extracting
- Thesaurus management system for coding medical terms (optional)
- Laboratory reference range management system

2.1.4 OpenClinica

OpenClinica³ is open-source web-based CDMS software for EDC built by OpenClinica LLC⁴, and widely used in clinical research to manage the data of a clinical trial. It is available as free Community Edition or commercial Enterprise Edition (with support and validation support), and helps with processing data from source through validation checks, analysis, reporting, storage and coding adverse events and medications, using MedDRA (Medical Dictionary for Regulatory Activities) and WHOART (WHO Adverse Event Reactions Terminology).

OpenClinica runs on any platform, and allows the management of diverse clinical studies through a unified interface, clinical data entry and validation, data extraction, study oversight, auditing, and reporting. It is implemented in Java as a web application with PostgreSQL as database backend (previously also support for Oracle as a backend). An OpenClinica instance can only use a single database and lays an emphasis on compliance with Good Clinical Practice data processing regulations (such as 21 CFR 11) and industry standards for data exchange (CDISC ODM).

OpenClinica can be used to administer several studies. Setting up a study in OpenClinica includes the following steps:

- entering metadata for the study
- setting up the eCRF(s) (via Microsoft Excel format tables)
- creating event definitions
- creating subject group classes

² <http://www.oracle.com/us/products/applications/health-sciences/e-clinical/clinical/index.html>

³ <https://openclinica.com/>

⁴ <https://openclinica.com/about-openclinica>

- defining the validation check rules
- assigning (existing or new) users
- adding the study's sites with metadata
- registering study subjects
- scheduling study events for the subjects

There is only a single interface for all user roles, but users can be assigned different roles on a per study or per site basis. Data for the individual study events can be entered into the system in three ways:

1. manual entry via the web interface (with an optional double data entry),
2. import from a file (which can be scheduled to take place recurrently),
3. data for a whole study or a specially defined dataset can be imported from several formats (e.g., tab-delimited, HTML, XLS, SPSS, SQL (for use with the OpenClinica DataMart⁵), CDISC ODM).

User documentation⁶ is in his current version 3.1. User manual is available through a Wiki⁷.

OpenClinica is currently used by Jules Bordet Institute for smaller academic clinical trials and by MAASTRO for eCRF. MAASTRO clinic has been using OpenClinica for electronic data capture in clinical trials since 2010. Currently, seven studies are operational or in the stage of study setup using OpenClinica. Furthermore, efforts are being undertaken to import clinical data from the electronic medical file of a patient. MAASTRO clinic is using OpenClinica version 3.0.4, but an update to 3.1 is expected to occur this year.

2.2 Electronic Health (Patient) Record systems

2.2.1 EHR and EURECA's clinical partners

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 and v3. Also, medical terminology is in the local language of each clinical partner (French, Dutch, German and 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 multi-centric and retrospective studies. Unreleased, undocumented, or changing data schemas, sometimes from external vendors and services, 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. At MAASTRO, a dedicated trial physician assistant must search through a variety of free text to evaluate patient eligibility for trials at an early stage of 'pre-admission' – free text

⁵ <https://docs.openclinica.com/3.1/openclinica-user-guide/data-mart-openclinica-enterprise-edition-only>

⁶ <https://docs.openclinica.com/3.1>

⁷ http://en.wikibooks.org/wiki/OpenClinica_User_Manual

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, and to use those criteria to establish the scope of extraction and representation of patient data. We will also employ query expansion of eligibility criteria using terminologies and ontologies in order to enhance information extraction from EHR 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 (noting that there are still only a few thousand Dutch labels that are part of the SNOMED release). In cases where other vocabularies have already been used to annotate data, such as NCI Thesaurus at MAASTRO, we will employ mappings between those other vocabularies and the EURECA vocabulary. This EURECA vocabulary called Core Dataset, will be compound of subsets from standard vocabularies – such as SNOMED-CT, HGNC or LOINC – enough to describe data from all scenarios in EURECA.

2.2.2 Cerner Millennium EPR in Oxford

The adoption of Cerner Millennium EPR is being rolled out in the Oxford University Hospital (OUH) NHS Trust as an effort to employ a primary EPR system for patient management and administration. The EPR system has been introduced to replace and link existing legacy clinical systems such as the pathology system and PACS. At the time of writing, the EPR system is up and running in the three main hospitals at Oxford, responsible for their patient management, maternity and accident & emergency (A&E) departments. The rationale is to link each individual information system and combine different clinical data sources into a single, interoperable, clinical data platform.

There are many ways to integrate with the Cerner Millennium EPR. The Cross-Enterprise Document Sharing (XDS) framework provides integration profiles that can be used to connect different clinical systems. The EPR has extensions that support this functionality. Furthermore, the EPR system sends out HL7 standard based messages that the receiving system could interpret in a meaning way. More details about inter-connecting systems by using a HL7 messaging engine will be discussed in the next section.

During the deployment and preparation of the EPR system, a data warehouse will be in place to be connected with Cerner Millennium to house clinical data that are required for clinical research. The EPR also aims to replace the current information system called "Case Note" that aggregates each department's clinical data, which forms an intranet where diagnostic results are shared between clinical professionals. Table 2-1 shows strategic impact of the adoption of Cerner Millennium to the current existing clinical systems.

System	Main purpose	Strategy after EPR
ORBIT	Main database for producing contract data sets from EPR.	To be further developed as the core data warehouse for the trust to underpin patient level costing and performance-reporting.
EPDS	Real-time copy of patient administration system (PAS)	Maintained post EPR – some data inconsistency exists and, over time, its use will diminish as systems integrate using the Mirth Integration Engines.
InfoFlex	Used for 18 week Cancer wait monitoring	Retained for use on Cancer 2 week wait but this may be incorporated into EPR later in the year
Powerchart - Maternity	Used for reporting on maternity	This has replaced the old OxMAT system and will be further developed to improve the Trust's CNST rating
Critical Care datasets	Additional dataset is current entered into PAS – a special form will be developed for this in Millennium for go live	The Millennium Critical Care system is being evaluated by Neuro and Neonatal Critical Care and the licensed version can produce both the Icnarc dataset and the Critical Care dataset as an integrated solution.
Varian Chemo Prescribing and radiotherapy systems	Used for Chemo prescribing and Radiotherapy scheduling	Will be maintained until Cerner can provide similar functionality
Radiology and lab systems	These systems will link to Order communications and feed into EPR	Radiology systems need to be merged and will probably require a local instance to do this; Radiology may be replaced with Cerner at a future date; Laboratory Medicine will be evaluating options to upgrade their systems over the next 2 years.

Table 2-1: Clinical systems that are intended to be replaced by Cerner Millennium EPR⁸

⁸ <http://www.ouh.nhs.uk/about/trust-board/2012/july/documents/TB201264ii-IMT-2017.pdf>

2.2.3 Mirth Connect, an HL7 messaging engine and integration tool

As introduced above, with Cerner Millennium EPR being in the centre of hospital IT systems, the whole electronic ecosystem requires the capability of handling real time events and messages. Mirth Connect⁹, an open source HL7 integration engine, has been employed to link the EPR system with other clinical systems in the OUH NHS trust. The integration engine uses channels to handle different HL7 messaging sources. Mirth Connect is highly configurable to define rules to route, filter and transform HL7 messages. The platform has been built using the Java programming language and it has a very active community for developers and users.

Figure 2-1 shows how the integration engine is configured to connect the EPR system with other operational clinical systems and the data warehouse. The performance of the platform is under assessment and a set of pilot studies are carried out to test how the system could handle heavy workload.

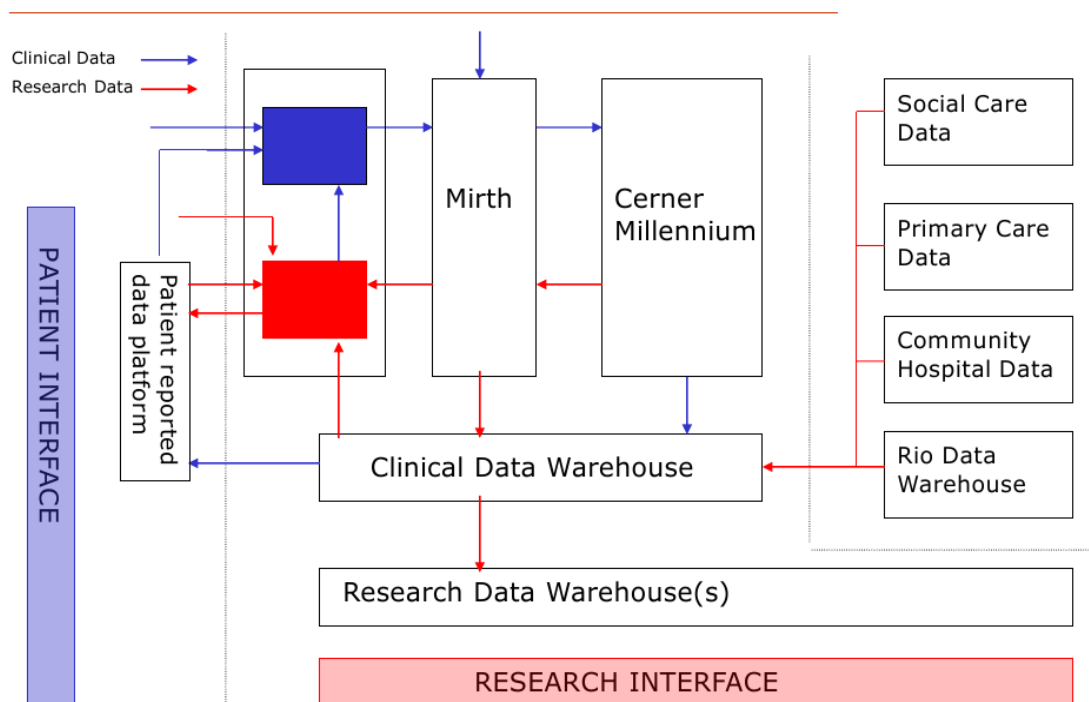


Figure 2-1 - Mirth Connect connects with Cerner Millennium EPR system to help populating the clinical data warehouse.

⁹ www.mirthcorp.com/products/mirth-connect

3 Reusable tools

3.1 Previous and running EU funded projects

3.1.1 Relevant tools from INTEGRATE

INTEGRATE Semantic Interoperability Layer

The INTEGRATE Semantic Interoperability Layer is aimed to enable the sharing of breast cancer clinical trials data. A global view of the structure of this layer can be seen in Figure 3-1.

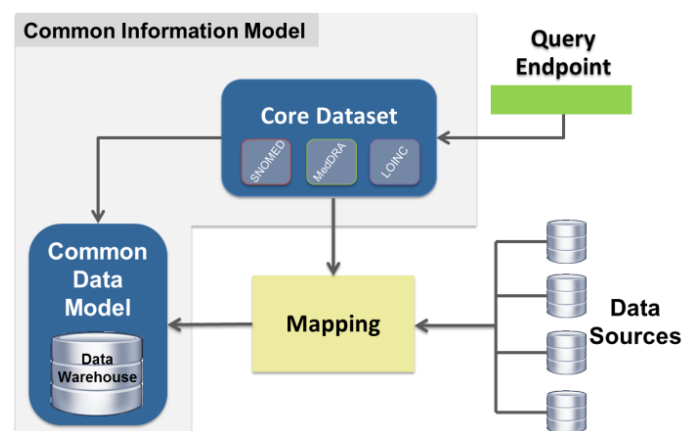


Figure 3-1: INTEGRATE Semantic Interoperability Layer

The main components of this structure are the Common Information Model compound by the Common Data Model (CDM) and the Core Dataset. The CDM is the system where the information should be stored following a common structure. Thus, the CDM offers a homogenous access for the different data extracted from the different clinical trials. Attempting to homogenize the data as much as possible, the Core Dataset is the 'lingua franca' used in the platform (compound of different well known vocabularies such as SNOMED, MedDRA, HGNC and LOINC). The Core Dataset normalises the concepts that are going to be used in the platform and the relationship between them. An endpoint is offered to the users, through it, the information can be asked and retrieved using different kinds of reasoning depending on the queried information.

Additionally, a mapping (Extract, Transform and Load (ETL)) process is needed to populate the CDM with the information that comes from the different clinical trials.

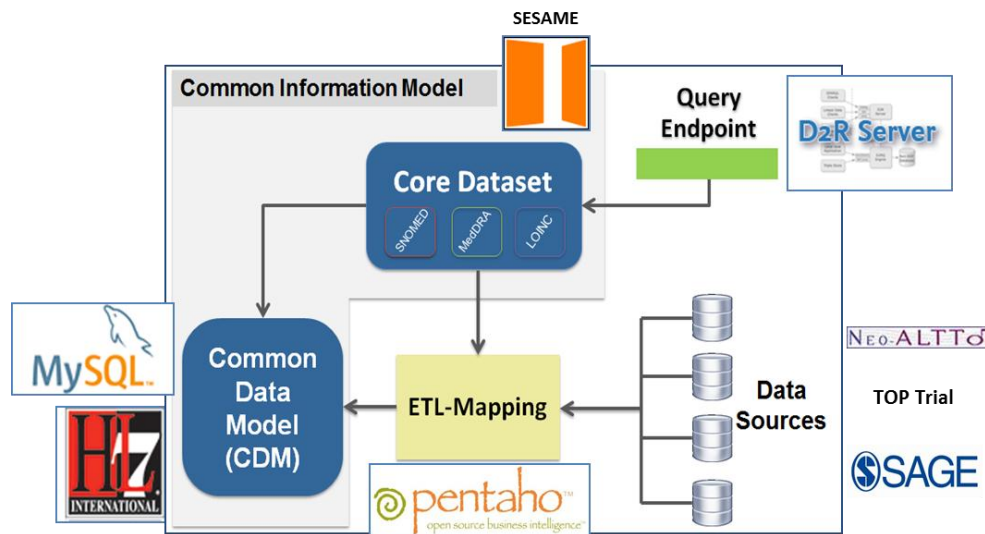


Figure 3-2: INTEGRATE Semantic Interoperability Layer tools usage

A general view of tools used in the INTEGRATE Semantic Interoperability Layer is depicted in Figure 3-2. This figure shows the different technologies used in each part of the layer. In detail, in the CDM the structure of the uniform model has been designed based on the HL7 Reference Information Model (RIM) to create a database under the MySQL relational database management system. In order to extract the information from the different clinical trials, transform it and load into the CDM, the Pentaho Data Integration (Kettle) has been utilised. Kettle will be described in section 3.3.2 Other tools used are: a D2R Server which provides a SPARQL end-point for querying the information stored in the CDM; and a Sesame server that is used to store the Core Dataset.

Currently the INTEGRATE interoperability layer is being enhanced for adaptation to EURECA requirements. For example, a free-text search engine is being implemented, and Mirth Connect is being studied for replacing Pentaho Kettle.

Trial metadata repository of INTEGRATE

The INTEGRATE consortium has been developing a demonstrator to assist a clinician in finding relevant trials for a patient. A trial metadata repository has been developed as part of this demonstrator (and is currently evolving as the INTEGRATE project progresses).

In the workflow, the clinician selects a patient from the patient (work) list, selects the trials he would like to consider for enrolment, and can dive deeper into specific trials to determine eligibility. The demonstrator supports automated evaluation of trial eligibility criteria (using patient data) where possible.

A trial metadata repository is being developed in this context. In the current approach, the repository tries to leverage the Biomedical Research Integrated Domain Group (BRIDG) Model¹⁰ initiative. The BRIDG model is a joint effort from several important stakeholders in the clinical care and research field, namely the Clinical Data Interchange Standards Consortium (CDISC), the HL7 Regulated Clinical Research Information Management

¹⁰ <http://www.bridgmodel.org/>

Technical Committee (RCRIM) Work Group, the US National Cancer Institute (NCI), and the US Food and Drug Administration (FDA).

The goal of the BRIDG Model (see footnote [10]) is *“to produce a shared view of the dynamic and static semantics for the domain of protocol-driven research and its associated regulatory artifacts. This domain of interest is further defined as: Protocol-driven research and its associated regulatory artifacts: i.e. the data, organization, resources, rules, and processes involved in the formal assessment of the utility, impact, or other pharmacological, physiological, or psychological effects of a drug, procedure, process, or device on a human, animal, or other subject or substance plus all associated regulatory artifacts required for or derived from this effort, including data specifically associated with post-marketing adverse event reporting.”*

The BRIDG model consists of a domain analysis model developed in UML¹¹ (mainly class diagrams) and is intended to capture the shared view of the domain (as defined above) and convey that understanding to parties outside of the initiative.

The approach taken in the INTEGRATE project is to leverage BRIDG for the development of the trial metadata repository with as aim to improve future interoperability of the repository and to gain the knowledge of the different perspectives ultimately encoded in the domain model by the various organizations outside of the INTEGRATE consortium.

In INTEGRATE, we have analysed the requirements from our stakeholders for the screening scenario and have identified the relevant parts in the BRIDG model. The subset of classes and class attributes of BRIDG that are relevant are taken and extended with INTEGRATE application specific constructs, resulting in the information model underlying the trial metadata repository. The information model is subsequently exposed by webservices. The information model contains (amongst others) constructs to express inclusion and exclusion criteria and their relation to trials, including INTEGRATE specific content to allow the actual matching/verification of a trial criterion with a patient's data.

It is foreseen that it might be possible to extend the trial metadata repository to aid, for instance, the trial recruitment scenario.

3.1.2 SemanticCT

SemanticCT¹² is a semantically enabled system for clinical trials. SemanticCT has been semantically integrated with various data in clinical trials, which include various trial documents with semantically annotated eligibility criteria and large amount of patient data with structured EHR and clinical medical records. Well-known medical terminologies and ontologies, such as SNOMED, LOINC, etc., have been used for the semantic interoperability.

SemanticCT is built on the top of LarkC¹³ (Large Knowledge Collider), a platform for scalable semantic data processing. With the built-in reasoning support for large-scale RDF/OWL data of LarkC, SemanticCT is able to provide various reasoning and data

¹¹ <http://www.uml.org/>

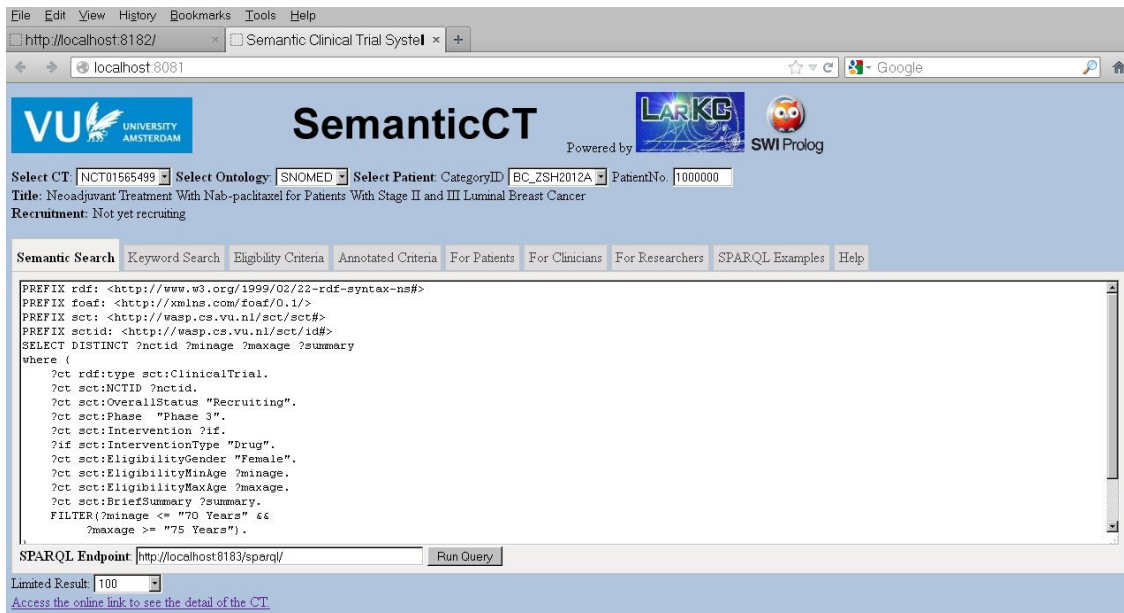
¹² <http://wasp.cs.vu.nl/sct>

¹³ <http://www.larkc.eu>

processing services for clinical trials, which include faster identification of eligible patients for recruitment service and efficient identification of eligible trials for patients.

SemanticCT supports for a rule-based reasoning over the formalization of eligibility criteria. That rule-based reasoning is developed based on SWI-Prolog¹⁴, a logic programming language.

Semantic Querying through the SemanticCT SPARQL endpoint



```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX sct: <http://wasp.cs.vu.nl/sct/sct#>
PREFIX sctid: <http://wasp.cs.vu.nl/sct/id#>
SELECT DISTINCT ?nctid ?minage ?maxage ?summary
where {
  ?ct rdf:type sct:ClinicalTrial.
  ?ct sct:NCTID ?nctid.
  ?ct sct:OverallStatus "Recruiting".
  ?ct sct:Phase "Phase 3".
  ?ct sct:Intervention ?if.
  ?if sct:InterventionType "Drug".
  ?ct sct:EligibilityGender "Female".
  ?ct sct:EligibilityMinAge ?minage.
  ?ct sct:EligibilityMaxAge ?maxage.
  ?ct sct:BriefSummary ?summary.
  FILTER(?minage <= "70 Years" ^^
    ?maxage >= "75 Years").
}
```

Figure 3-3: A SPARQL query can be posted from the interface of SemanticCT for semantic search over semantic data which have been loaded into the SemanticCT system.

¹⁴ <http://www.swi-prolog.org/>

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX sct: <http://wasp.cs.vu.nl/sct/sct#>
PREFIX sctid: <http://wasp.cs.vu.nl/sct/id#>
SELECT DISTINCT ?nctid ?minage ?maxage ?summary
where {
  ?ct rdf:type sct:ClinicalTrial.
  ?ct sct:NCTID ?nctid.
  ?ct sct:OverallStatus "Recruiting".
  ?ct sct:Phase "Phase 3".
  ?ct sct:Intervention ?if.
  ?if sct:InterventionType "Drug".
  ?ct sct:EligibilityGender "Female".
      ?ct sct:EligibilityMinAge ?minage.
  ?ct sct:EligibilityMaxAge ?maxage.
  ?ct sct:BriefSummary ?summary.
  FILTER(?minage <= "70 Years" &&
    ?maxage >= "75 Years").
}
ORDER BY ?nctid
LIMIT 100
```

Figure 3-4: SPARQL query which searches for all recruiting phase 3 trials for female patients with age between 70 and 75

The semantic data of 4665 clinical trials of breast cancer have been integrated in SemanticCT. Thus, we can use SemanticCT to find the core data of SNOMED CT, which have been used in the annotations of the clinical trials of breast cancer by the SPARQL query in Figure 3-5.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX sct: <http://wasp.cs.vu.nl/sct/sct#>
PREFIX sctid: <http://wasp.cs.vu.nl/sct/id#>
PREFIX snomed: <http://www.ihtsdo.org/>
SELECT DISTINCT ?cid ?label
where {
  ?cta sct:AnnotationBean ?ab.
  ?ab sct:OntologyID '46116'.
  ?ab sct:ConceptID ?cid.
  ?sc sct:id ?cid.
  ?sc rdf:type sct:SNOMEDConcept.
  ?sc rdfs:label ?label.
  FILTER(?cid > '10000000' &&
    ?cid < '20000000'
  ).
}
ORDER BY ?cid
```

Figure 3-5: SPARQL query which lists SNOMED concepts which have been used for the annotations of clinical trials of breast cancer.

3.1.3 Semantic Platform LarkC (Large Knowledge Collider project)

There have been several well-developed triple stores which can be used to serve as a semantic platform to build SPARQL endpoints for the services of querying over large-scale semantic data. Well-known triple stores are OWLIM¹⁵ and Virtuoso¹⁶. Those triple stores usually support for basic RDFS reasoning over semantic data.

LarkC is a platform for scalable semantic data processing. OWLIM is used to be the basic data layer of LarkC. LarkC fulfils the needs in sectors that are dependent on massive heterogeneous information sources such as telecommunication services, biomedical research, and drug-discovery. The platform has a pluggable architecture in which it is possible to exploit techniques and heuristics from diverse areas such as databases, machine learning, cognitive science, the Semantic Web, and others. LarkC provides a number of pluggable components: retrieval, abstraction, selection, reasoning and deciding. In LarkC, massive, distributed and necessarily incomplete reasoning is performed over web-scale knowledge sources.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix larkc: <http://larkc.eu/schema#> .

# Define the plug-in
_:plugin a
<urn:eu.larkc.plugin.reason.GenericReasoner.GenericReasonerPlugin> .

# Define a path to set the input and output of the workflow
_:path a larkc:Path .
_:path larkc:hasInput _:plugin .
_:path larkc:hasOutput _:plugin .

# Connect an endpoint to the path
<urn:eu.larkc.endpoint.sparql.ep1> a <urn:eu.larkc.endpoint.sparql> .
<urn:eu.larkc.endpoint.sparql.ep1> larkc:links _:path .
```

Figure 3-6: LarkC workflow of generic reasoner plugin for SemanticCT

In SemanticCT, a management component manages SPARQL endpoints which are built as SemanticCT workflows in LarkC. A generic reasoning plug-in in LarkC provides the basic reasoning service over large-scale semantic data, like RDF/RDFS/OWL data for SemanticCT. Figure 3-6 shows a workflow specification which calls the generic reasoner plugin for SemanticCT.

3.1.4 SWI-Prolog

In SemanticCT, the rule-based formalization is developed based on the logic programming language Prolog. The SWI-Prolog (<http://www.swi-prolog.org/>) has been

¹⁵ <http://www.ontotext.com/owlim>
¹⁶ <http://virtuoso.openlinksw.com>

selected to be the basic language for the rule-based formalization of eligibility criteria, because of the following features of SWI-Prolog:

i) Semantic Web Support

SWI-Prolog has been facilitated with powerful libraries for semantic data processing and services. It provides a basic tool for communication with SPARQL endpoints and other REST based web servers. Furthermore, SWI-Prolog also supports the basic reasoning and storage of semantic data. Thus, the SWI-Prolog has the advantage of the support for semantic data processing;

ii) Powerful Processing Facilities

SWI-Prolog provides various libraries for data processing, which includes not only the tools for text processing and database-like storage management, but also workflow processing and distributed/parallel processing.

Figure 3-7 shows an example of the formalization of inclusion criteria for the clinical trial 'NCT00002720' as a rule in SWI-Prolog. The rule states that the inclusion criteria are: patients of stage I, invasive breast cancer, oestrogen receptor positive, progesterone receptor positive or negative, the age between 65 and 80, and the menopausal status is postmenopausal.

```
meetInclusionCriteria(_PatientID, PatientData, CT,
_NotYetCheckedItems) :-
    CT = 'nct00002720',
    breast_cancer_stage(PatientData, '1'),
    invasive_breast_cancer(PatientData),
    er_positive(PatientData),
    known_pr_status(PatientData),
    age_between(PatientData, 65, 80),
    postmenopausal(PatientData).
```

Figure 3-7: SWI-Prolog Rule which formalises the inclusion criteria of the trial NCT00002720

3.1.5 Linked Life Data (LLD)

Linked Life Data¹⁷ (LLD) is a semantic data integration platform for the biomedical domain. LLD enables search and exploration across RDF statements from various sources including UniProt, PubMed, EntrezGene and many others. LLD can perform complex SPARQL queries and retrieve more than one billion RDF resources.

The current version of Linked Life Data (version 1.1), loaded on July 20, 2012, consists of 8,740,201,002 triples, which cover 2,068,072,570 entities. The following Table 3-1 shows the semantic data set of LLD:

Data source	Named graph under http://linkedlifedata.com/resource/	Load date	Statements	Instances type
-------------	--	-----------	------------	----------------

¹⁷ <http://linkedlifedata.com/>

Data source	Named graph under http://linkedlifedata.com/resource/	Load date	Statements	Instances type
BioGRID	biogrid	2012.07.20	14,327,672	biopax-2:entity
CellMap	cellmap	2012.07.20	154,863	biopax-2:biochemicalReaction
ChEBI	chebi	2012.07.20	323,220	skos:Concept
DailyMed	dailymed	2012.07.20	162,972	dailymed:drugs
DiseaseOntology	diseaseontology	2012.07.20	90,652	skos:Concept
Diseasome	diseasome	2012.07.20	72,445	diseasome:diseases
DrugBank	drugbank	2012.07.20	517,023	drugbank:drugs
Freebase	freebase	2012.07.20	705,161,223	
GeneOntology	geneontology	2012.07.20	364,947	skos:Concept
HapMap	hapmap	2012.07.20	22,462,178	-
HPRD	hprd	2012.07.20	1,972,499	biopax-2:entity
HumanCYC	humancyc	2012.07.20	332,828	biopax-2:entity
HumanPhenotypeOntology	phenotype	2012.07.20	62,240	skos:Concept
IMID	imid	2012.07.20	117,675	biopax-2:entity
IntAct	intact	2012.07.20	2,845,521	biopax-2:entity
LHGDN	lhgdn	2012.07.20	316,021	-
LinkedCT	linkedct	2012.07.20	9,804,652	linkedct:condition
MINT	mint	2012.07.20	17,249,403	biopax-2:entity
NCBIGene	entrezgene	2012.07.20	186,904,730	entrezgene:Gene
NCI Nature	nci-nature	2012.07.20	914,442	biopax-2:entity
PubMed	pubmed	2012.07.20	1,454,405,726	pubmed:Citation
Reactome	reactome	2012.07.20	1,082,499	biopax-2:entity
SIDER	sider	2012.07.20	101,542	sider:drugs
SymptomOntology	symptom	2012.07.20	5,210	skos:Concept
UMLS	umls	2012.07.20	129,803,921	skos:Concept

Data source	Named graph under http://linkedlifedata.com/resource/	Load date	Statements	Instances type
UniProt	uniprot	2012.07.20	3,177,871,239	uniprot:Protein

Table 3-1: The data sources of Linked Life Data

Figure 3-8 shows the interface of Linked Life Data, from which users can make SPARQL queries for semantic search over the data sources above or search by ordinary keywords.



Figure 3-8: Interface of Linked Life Data

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX          drugbank:          <http://www4.wiwiss.fu-
berlin.de/drugbank/resource/drugbank/>
PREFIX          disease-instance:  <http://www4.wiwiss.fu-
berlin.de/disease/resource/disease/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX linkedct: <http://data.linkedct.org/resource/linkedct/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX          relationontology: <http://linkedlifedata.com/resource/relationontology/>

SELECT ?drugname ?exclusion ?trial
WHERE {
    ?drug rdf:type drugbank:drugs .
    ?drug drugbank:possibleDiseaseTarget ?disease .
    ?disease disease-instance:class <http://www4.wiwiss.fu-
berlin.de/disease/resource/diseaseClass/Respiratory> .
    ?drug skos:exactMatch ?intervention .
    ?intervention rdf:type linkedct:intervention .
    ?trial linkedct:intervention ?intervention .
    ?trial relationontology:hasExclusionCriteria ?exclusionCriteria .
    ?drug rdfs:label ?drugname .
    ?disease rdfs:label ?diseasename .
    ?intervention linkedct:description ?desc .
    ?trial linkedct:brief_title ?trialTitle .
    ?exclusionCriteria skos:prefLabel ?exclusion .
}
```

Figure 3-9: What are the exclusion criteria for studies that use a drug intervention to treat respiratory diseases?

3.2 EHR4CR Convergence Meeting, Paris, January, 2013

W3C HCLS Interest Group as meeting platform for similar projects

In January 2013, the EHR4CR Innovative Medicines Initiative¹⁸ hosted a meeting¹⁹ in Paris and invited several EU projects with similar goals, including EURECA, SALUS, OpenPhacts, Linked2Safety, and eTRIKS. In the following sections, we attempt to highlight areas of overlap between EURECA and the other projects at the meeting based on the meeting summary report²⁰.

The World Wide Web Consortium's (W3C²¹) Health Care and Life Sciences Interest Group (described in section 3.3.1) has offered to host the slides, summary report, as well as future activities in support of cross-project collaboration in the form of teleconferences and meetings, as part of the Clinical Observations Interoperability task force.

3.2.1 EHR4CR Innovative Medicines Initiative

¹⁸ <http://www.ehr4cr.eu/>

¹⁹ More information on

<http://www.w3.org/wiki/HCLS/ClinicalObservationsInteroperability/Convergence>

²⁰ <http://www.w3.org/wiki/images/3/32/Convergence-Meeting-2013-Summary.pdf>

²¹ <http://www.w3.org/>

Start Date: 01/03/2011

Duration: 48 months

EHR4CR and EURECA have much in common. EHR4CR started about 1 year earlier than EURECA. The EHR4CR platform plans to implement four use cases:

1. protocol feasibility testing
2. patient identification and recruitment for clinical trials
3. supporting clinical trial execution
4. adverse event reporting

These are to be demonstrated by 10 pilots in 5 European countries. Also similar to EURECA, the EHR4CR platform will be a loosely coupled service platform, which orchestrates independent services, with attempts to harmonize with FHIR, BRIDG, and CDISC SHARE. Both EHR4CR and EURECA are working on formal representations of eligibility criteria, with Semantic Web representations as well, which appears to hold large potential for collaboration and harmonization between the projects. EHR4CR is developing a query language ECLECTIC (Eligibility Criteria Language for European Clinical Trial Investigation and Construction), which can be used to transform elementary queries into other query languages such as OCL, SPARQL, and SQL. The ECLECTIC query language could serve as a starting point for collaboration.

3.2.2 SALUS

Start Date: 01/02/2012

Duration: 36 months

SALUS has built a number of resources that would be useful to integrate into EURECA. SALUS built ontologies for HL7 CDA and OMOP CDM, as well as a SALUS Common Ontology for semantic mediation. SALUS has downloaded and fine-tuned the terminologies WHO-ART, ICD-9-CM, ICD-10, MedDRA, and the SNOMED CT Clinical Findings sub-hierarchy. It has also built an RDF representation of WHO-ATC code system.

SALUS is building Web-based graphical interfaces for expressing inclusion/exclusion criteria of the foreground and background populations of the post market safety analysis studies. Those interfaces use the SALUS common model and the semantic model of the queries (to be shared when ready) is based on the formalism introduced in HL7 HQMF.

SALUS employs EHR interface standards, namely HL7 Clinical Document Architecture Release 2 (CDA) based templates, and ISO/CEN EN 13606 EHR Extract based archetypes and templates and has developed a Data Definition Ontology on top of the ORBIS installation at UKD (University Clinic-Technical University of Dresden). The TUD site will build a SPARQL endpoint on top of the local ontology of the ORBIS UKD installation.

SALUS and CDISC will supply one of the co-authors of IHE Data Exchange (DEX) Profile. 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. An early SALUS publication discusses similar ideas: “Providing Semantic Interoperability between Clinical Care and Clinical Research Domains”, Laleci, G., Yuksel, M., Dogac, A., IEEE Transactions on Information Technology in Biomedicine.

3.2.3 Open PHACTS Innovative Medicines Initiative

Start Date: 01/03/2011

Duration: 36 months

The Open PHACTS consortium is building the Open PHACTS Discovery Platform for drug discovery, which will freely provide tools and services to support pharmacological research. A number of technological applications from Open PHACTS should prove to be useful guides in EURECA, especially with implementation experience creating a large scale triplestore (knowledge base), Semantic Web APIs, and SPARQL Services. For example, Open PHACTS uses the Vocabulary of Interlinked Datasets (VOID) to describe all its datasets, as well as mappings for metadata management²². Open PHACTS is also committed to providing data provenance in RDF. The project has also made a guide to RDF²³ and uses the Linked Data API²⁴. The Concept Wiki²⁵ is used for term to identifier mapping and the Open Phacts Identity Mappings Service based on Bridge DB is used for identifier to identifier mapping. For its Use Case, Open PHACTS concentrates on answering the top 20 ranked research questions²⁶ from a list of 83 proposed by consortium members.

3.2.4 Linked2Safety

Start Date: 01/10/2011

Duration: 36 months

The Linked2Safety project will facilitate the scalable and standardised semantic interlinking, sharing and reuse of heterogeneous EHR repositories and provide healthcare professionals, clinical researchers and pharmaceutical companies' experts with a user-friendly, sophisticated, collaborative decision-making environment. This will enable 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.

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). Linked2Safety represents clinical data in anonymised and aggregated multidimensional data-cubes. Linked2Safety performs mapping at the Instance Level and the Schema

²² <http://www.openphacts.org/specs/datadesc/>

²³ <http://www.openphacts.org/specs/rdfguide/>

²⁴ <http://code.google.com/p/linked-data-api/>

²⁵ <http://ops.conceptwiki.org/wiki/>

²⁶ <http://www.openphacts.org/about-ops/200>

Level. The Linked2Safety Platform interface will include a query builder that makes use of the SEHR to guide non-SPARQL experts in the query building process and a tool that assists in the visual exploration of ontologies.

3.2.5 eTRIKS Innovative Medicines Initiative

Start Date: 01/10/2012
Duration: 60 months

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.

The current eTRIKS approach is to understand the data model/structure of the supported project's electronic document capture (EDC), convert all data into a common format, tag metadata and organise it into ontologies, namely CDISK/SDTM²⁷ and i2b2²⁸. 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 platform²⁹.

3.3 Other related projects

3.3.1 W3C Health Care and Life Sciences Interest Group

The HCLS IG³⁰ was originally chartered in 2005 to advocate developing and applying Semantic Web technologies across healthcare, life sciences, clinical research and the continuum of translational medicine. In recent years, the HCLS IG grew to about 100 participants and a mailing list of ~600. The current 2011 charter continues to focus on the use of Semantic Web technologies to realize specific use cases which themselves have a specific clinical, research or business values. Its current activities and task forces can be found on the wiki page³¹. It is specifically within the scope of the HCLS IG to:

- Create Linked Data and guidelines to help others create Linked Data.
- Create vocabularies and vocabulary bridges.
- Build demonstrations and test suites.
- Assist other groups to create data and tools within the scope of this interest group.
- Advise industry on the relevance and maturity of tools.

²⁷ <http://www.cdisc.org/sdtm>

²⁸ <https://www.i2b2.org/>

²⁹ <http://www.transmartproject.org/>

³⁰ <http://www.w3.org/blog/hcls/>

³¹ <http://www.w3.org/wiki/HCLSIG>

Over the course of the previous charter, the HCLS IG developed a set of data and demonstrators enabling life science and health care practitioners to consume and reason over domain data. Here are some highlights up until 2011:

- Linked Open Drug Data (LODD) and winning the iTriplification Challenge 2009
- Light-weight CDISC – HL7 bridge for patient eligibility studies
- SPARQL Federation and RDF to SQL through SWObjects³²
- Semantic integration of LODD, patient data via the Translational Medicine Ontology (TMO)
- Non-proliferation of ontologies and interoperability with other standards and conventions
- HCLS Knowledge Base

HCLS IG has assembled a more complete list of deliverables³³ and a list of tools³⁴ that have been used by HCLS IG members. Recently, a W3C Note³⁵ has been produced as a guide to RDF in the HCLS domain which describes Semantic Web principles relevant to EURECA. The HCLS IG has accumulated a wide variety and depth of experience in areas directly relevant to the EURECA project and is directly involved in the latest developments within HL7 such as FHIR, so it should prove to be a valuable source of knowledge and expertise.

3.3.2 Pentaho Data Integration (Kettle)

Pentaho Data Integration (PDI), also known as Kettle, is an open source Extract, Transform and Load (ETL) tool. It is part of the Pentaho Business Analytics suite, which offers solutions for data mining, data integration, knowledge discovery, analysis and visualization. Kettle can be run as a server for real time data or integrated with another program such as WEKA, a popular suite of machine learning software written in Java.

Kettle enables users to build ETL processes graphically through an intuitive interface with several options. An ETL process is built as a data flow, where the different operations are added from drag and drop menu. This graphical design interface of Kettle is called Spoon. After a transformation (Spoon uses this name referring to the ETL process) is built, Kettle stores it in XML format.

Kettle provides a Java API that enables the building of ETL processes directly coded in Java or integrated with another program with the use of the transformation built with Spoon.

This tool could be used in the EURECA project to extract the information coming from different clinical trials and transform it to be stored following the structure of a Common Data Model created for the project. Figure 3-10 shows an example of transformation using Kettle.

³² <http://en.wikipedia.org/wiki/SWObjects>

³³ <http://www.w3.org/wiki/HCLSIG/Products>

³⁴ <http://www.w3.org/wiki/HCLSIG/Tools>

³⁵ <http://www.w3.org/2001/sw/hcls/notes/hcls-rdf-guide/>

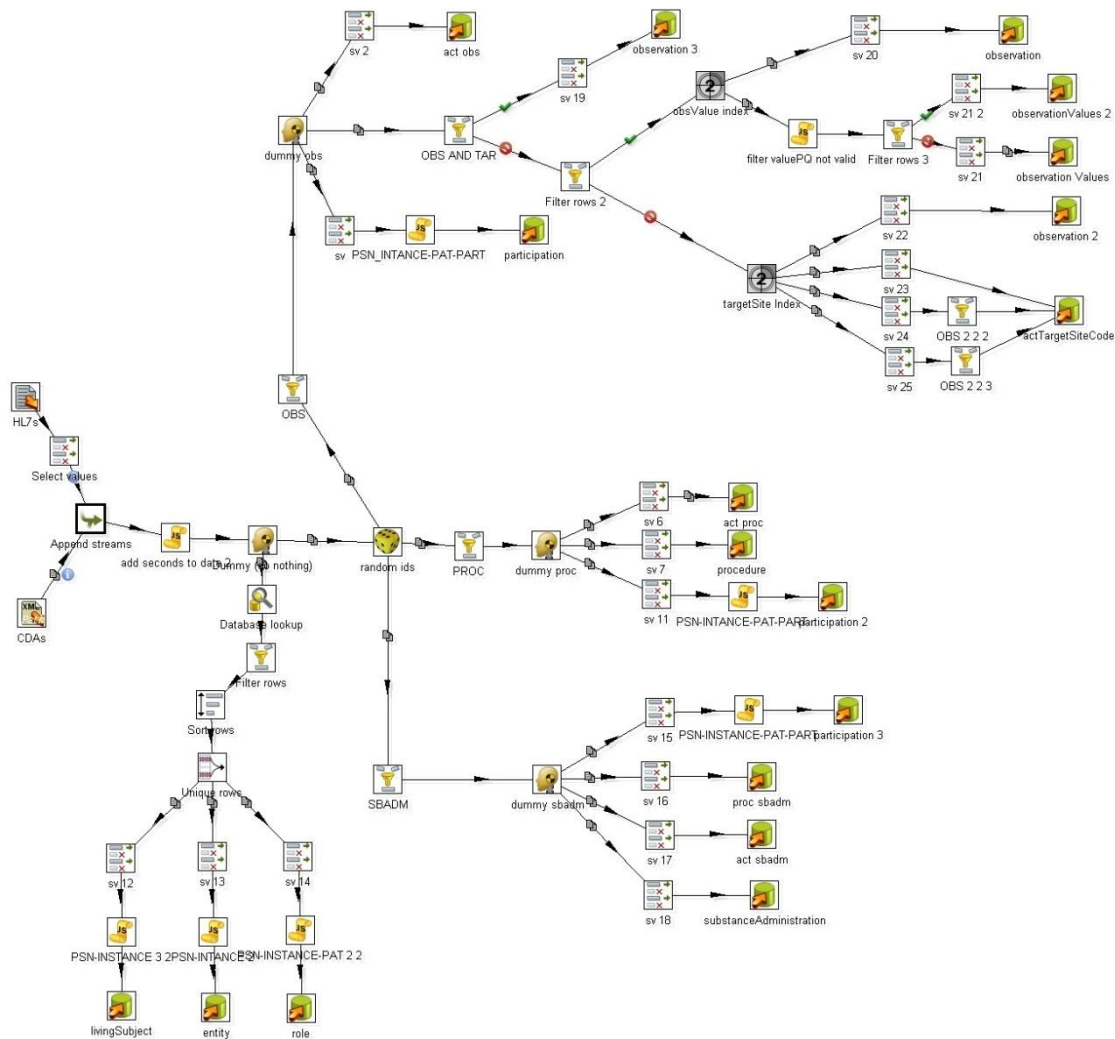


Figure 3-10: INTEGRATE ETL process

3.3.3 eHealthMonitor

The eHealthMonitor³⁶ project provides a platform that generates a Personal eHealth Knowledge Space (PeKS) as an aggregation of all relevant sources (e.g., EHR and PHR) relevant for the provision of individualized personal eHealth Services. The PeKS will be used and validated by end users in two hospital-based scenarios – covering dementia and cardio-vascular domain as well as one prevention-based scenario in the health insurance domain.

eHealthMonitor will develop an adaptive platform architecture for individualized personal electronic healthcare services. This serves as a basis for personal eHealth services that support cooperation and decision making of the involved participants (patients, clinicians, social services) through web, mobile and remote access channels:

³⁶ <http://www.ehealthmonitor.eu/>

- a) Medical decision support services support medical professionals during diagnosis of health risks and diseases.
- b) Personal information services provide risk factor and treatment related information as well as recommendations for action to individual users as part of prevention strategies or actual treatments.
- c) Environmental and lifestyle risk factor monitoring services provide means to monitor risk factors affecting the health status.
- d) Physiological and bio-chemical data monitoring services analyse data from sensors, imaging, and laboratory findings, considering online data from wearable sensors as well as existing data in Electronic Health records (EHR) and Personal Health Records (PHR).

eHealthMonitor's vision is to significantly increase the individualization of personal eHealth services and thereby the quality and patients' acceptance of electronic healthcare services for treatment and prevention. So within EURECA tools & scenarios developed in eHealthMonitor promoting this individualization could be exploited and reused.

3.3.4 D2R Server

D2R server is a tool for publishing relational databases as Semantic Web resources. Since data on the Semantic Web is modelled and represented using RDF, D2R server allows relational data to be browsed as RDF resources and queried through SPARQL endpoints. To achieve that, customizable mappings called *D2RQ mappings*, link relational tables to RDF resources. Then, when SPARQL queries are issued, query rewriting techniques are used to rewrite them into SQL queries via the mapping. This on-the-fly translation allows publishing of RDF from large live databases and eliminates the need for replicating the data into a dedicated RDF triple store.

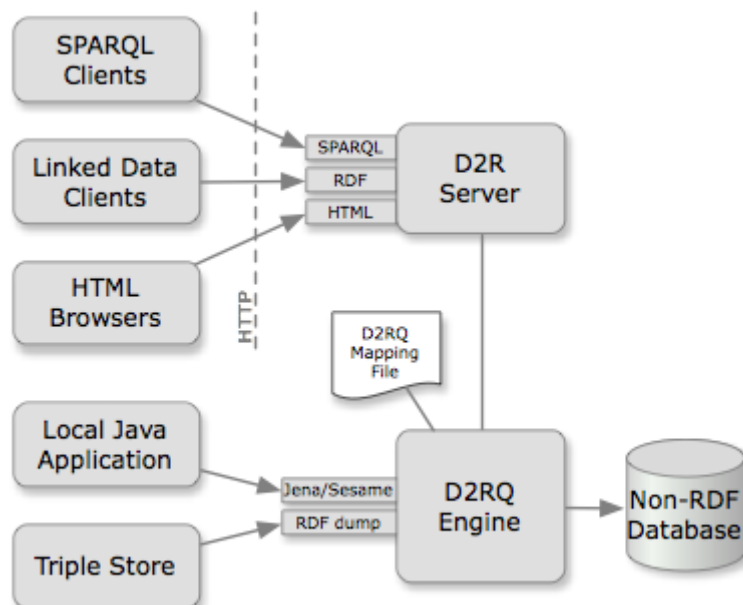


Figure 3-11: D2R Architecture

Finally the entire platform is open source and published under the Apache license. Moreover it is widely used in semantic web related research projects and has a strong community supporting it.

3.3.5 Dr Eye

Dr Eye³⁷ is an open access, flexible and easy to use platform, for intuitive annotation and segmentation of tumour regions. Its clinically driven development followed an open modular architecture focusing on plug-in components. DrEye's main advantage is that the user can quickly and accurately delineate complex areas in medical images in contrast with other platforms that do not facilitate the delineation of areas with complicated shapes. Additionally, multiple labels can be set to allow the user to annotate and manage many different areas of interest in each selected slide. The close collaboration with clinicians in designing the platform has ensured that it can be effectively used in the clinical setting.

Another reported feature that adds value to the platform is that it allows computational "in-silico" models of cancer growth and simulation of therapy response to be easily plugged in, in order to provide a future integrated platform for modelling assisted therapy decision making. Currently, DrEye's development team is working towards incorporating such models in the platform and the new, stable version will also be available. In this context, the platform could also serve as a validation environment where the simulation predictions could be compared with the actual therapy outcome in order to achieve a global optimization of the modelling modules.

DrEye platform is based on the .NET framework architecture and can be used in any Windows-based computer. The graphical interface is based on well-known Microsoft Office applications to ensure a user-friendly environment.

DrEye is regularly maintained according to feedback received by a number of regular users from different clinical settings. Its functionality is expanding according to clinical needs that arise from existing and new users. DrEye has been funded within ContraCancrum (completed), TUMOR and p-medicine and there are several plans to sustain the platform both from EC projects and self -funding mechanisms of FORTH for at least the next five years.

In its current version, DrEye is proprietary software that it is available for use at no cost (it is free).

Features List

1. Support for multiple users (with roles and access management).
2. View a single DICOM image or a whole series of DICOM images.
3. Tabbed interface, which allows for multiple series to be opened at once.
4. Configuration of DICOM Level and DICOM Width for a selected image or for the whole series.
5. Intuitive navigation/viewing.
6. Support for multiple annotations per DICOM image that feature: Label, Colour, Types, Opacity, Support for Annotation management (merge, sort, ...) and batch editing (rename all, ...)

³⁷ http://biomodeling.ics.forth.gr/?page_id=8

7. Powerful annotation tools: Pen, Eraser, Rectangular Marque, Elliptical Marquee, Boolean operations among ROIs, Magic Wand, Active contours using Greedy algorithm, Active contours using Snakes algorithm, Semi-automatic selection of outer boundaries, and more...
8. Metrics (Ruler, Surface estimation and Volume estimation of a selected ROI)
9. Histogram generation for multiple ROIs
10. More features/functionalties can be added with 3rd party plugins that can be embedded in the platform seamlessly. SDK and guides are available.
11. Import/Export in various common formats (comma separated csv files, excel files, text files, xml, ...)
12. Embedded Viewer for DICOM tags
13. 3D visualization of a selected series and of its annotations.

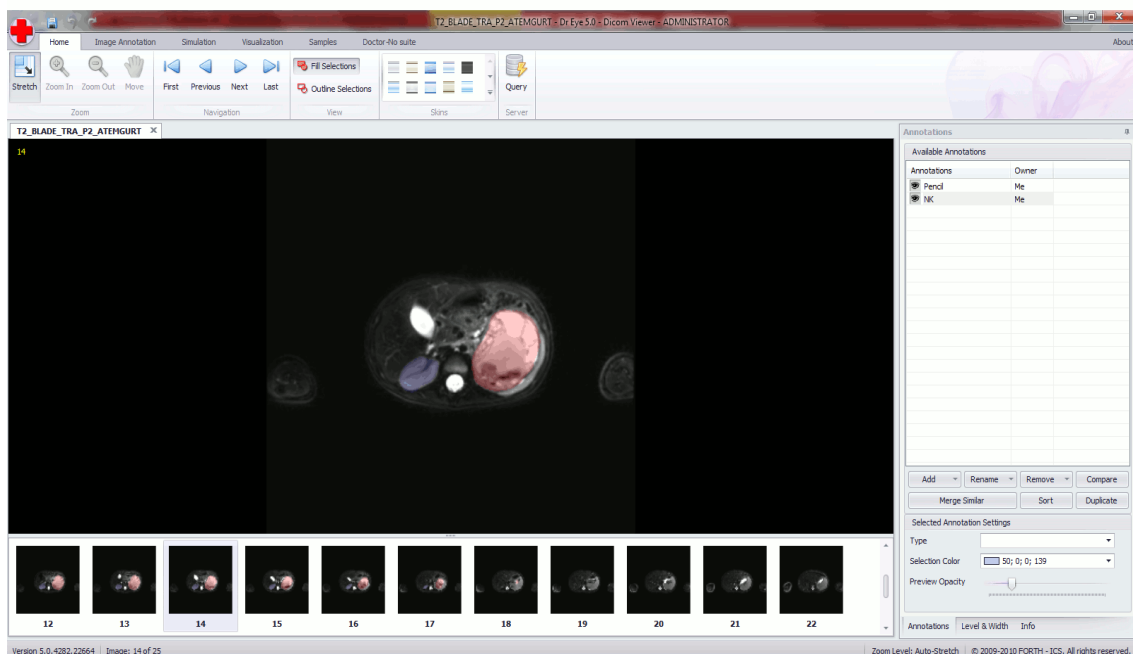


Figure 3-12: The main window of the Doctor Eye platform with a MRI data set and annotations

3.3.6 SMART & IndivoX

SMART³⁸ (Substitutable Medical Apps and Reusable Technologies) is a project funded by the Office of the National Coordinator for Health Information Technology (ONC) through the Strategic Health IT Advanced Research Projects (SHARP) program. The major deliverable of this project is a platform architecture that achieves two goals:

³⁸ <http://smartplatforms.org>

- a) A user interface that allows substitutability for medical apps
- b) A set of scalable services that enable efficient data capture, storage, retrieval and analysis by respecting institutional autonomy and patient privacy.

The SMART team has completed the first phase of the project (i) defining an app programming interface, (ii) developing containers, and (iii) producing a set of charter apps that showcase the system capabilities.

One of those containers implementing the SMART API is IndivoX³⁹. IndivoX is a personal health platform, enabling users to own and manage a complete digital copy of their health and wellness information. Moreover, IndivoX allows the ready integration of diverse sources of medical data under a patient's control through the use of standards-based communication protocols and APIs for connecting PCHRs to existing and future health information systems. Furthermore, the system allows the easy sharing of patient information among institutions, doctors and carenets securely and with the consent of the user.

Finally, IndivoX is open source and web-based. It is extensible via a standard API and has a strong community supporting it.

3.3.7 The SPECIALIST NLP Tools and Medical text processing toolkit

The SPECIALIST NLP tools⁴⁰ are a set of libraries and programs written in the Java programming language to provide lexical tools for medical text processing. Developed by the Lister Hill National Center for Biomedical Communications, the toolkit contains a comprehensive list of medical lexicon, libraries and programs for parsing and processing medical text. Although not a complete application itself, the toolkit is flexible for achieving different NLP tasks and programming needs. One of the advantages of the SPECIALIST NLP tools over other sophisticated systems such as MedLee is that the tools consist of many components which can be used independently. The flexibility ensures that the different programming needs can be met.

The tools contain a large and comprehensive medical lexicon and a set of text processing utilities. The following table summarises the functionalities of each component.

³⁹ <http://indivohealth.org/>

⁴⁰ <http://lexsrv3.nlm.nih.gov/Specialist/Home/index.html>

The SPECIALIST lexicon	A large syntactic lexicon of biomedical and general English, designed to provide the information needed for the SPECIALIST NLP System.
LexAccess	A tool provides access to information from the SPECIALIST LEXICON.
Lexical tools	A tool set designed to manage lexical variation, indexing, and normalization, etc. in biomedical text.
Text tools	A nested set of JAVA objects designed to help users analyse free text documents into words, terms, phrases, sentences and sections.
GSpell	Includes two programs GSpell a spelling suggestion tool and BagOwordsPlus a phrase retrieval tool.
POS tagger	The Tagger is a Part of Speech (POS) tagger.

Table 3-2: SPECIALIST NLP Tools

Sophisticated applications can be built by combining these tools with other technology such as the open source full text search engine Apache Lucene⁴¹.

3.4 Security and Privacy Enhancing Tools

Through project partner Custodix, EURECA gains deployment experience and expertise with a comprehensive suite of tools and technologies including PIMS, CATS, Shibboleth, Security Token Service, XACML, and LDAP.

3.4.1 PIMS

PIMS offers a central service for tracking patient identifiers originating from different data sources. Feeding it with personal identifying information coming from different sources, allows PIMS to issue pseudonyms to different domains. A pseudonym is a cryptographically strong identifier, created randomly as a GUID. A given person can be uniquely identified within an administrative domain by assigning domain specific pseudonyms, which are completely isolated. This means pseudonyms are not derivable from one another. The re-identification module allows a user to translate an issued pseudonym back to its original personal identifying information. This is accomplished by keeping all information in a secured database.

PIMS is able to link person records that slightly differ but actually match the same person with each other. This process is called indexation. It matches records, based on personal identifying information, and adds them to an index kept in an index tree, the Master Patient Index (MPI). By indexing records the same domain specific pseudonym can be issued for multiple matching person records. Background processes will scan the secured database for newly added person records and add them to the MPI. PIMS

⁴¹ A demo that has been built using the SPECIALIST NLP tools is available at <http://147.252.66.54/Concilio/ConcilioDemo.html>.

incorporates a probabilistic matching engine based on the well-known Fellegi-Sunter⁴² algorithm. Using fuzzy matching techniques (Jaro⁴³-Winkler⁴⁴) and the calculation of relative occurrences on record fields, a weight is assigned to comparison of two records. Based on that weight a match/non-match decision is made. The matching engine is fully configurable to reduce the number of false positives and false negatives to a minimum. PIMS (Personal Information Management System) can be used in the EURECA platform as identity manager; guarding and linking the different EURECA domain ids (care, research and trial support; see deliverable 7.1) from a patient.

3.4.2 CATS

CATS (Custodix Anonymisation Tool Service) is a service, developed by Custodix, responsible for the de-identification or anonymisation of (clinical) data files. Based on a predefined set of transformation rules, called a privacy profile, CATS will process an input file and deliver it to the next component in chain (e.g. a database on a research platform). CATS supports multiple privacy profiles. Before processing a file, it will select the correct privacy profile based on the detected file type, and given content. Important transformations are:

- Pseudonymisation: Based on person identifying information a pseudonym is added. CATS can easily be integrated with PIMS.
- De-identification: Person identifying information is cleared from the input.
- Encryption: Sensitive data can be encrypted with configurable public key.
- String replacement: Based on regular expression, string values can be replaced

Because of the modular nature of the CATS platform, the set of already supported standard data formats can be extended with new formats. The CATS service can be invoked by using one of the provided interfaces:

- A web interface: An end-user can upload files for transformation through a web front end.
- A web service interface: CATS is equipped with a web service layer (SOAP, WSDL), secured with SAML tokens.
- CATS client tool: A user can launch a client side CATS tool (Java web start) to process files locally before uploading to CATS.

In EURECA, CATS can be used as de-identifying tool in order to comply with the legal requirements defined for the privacy framework in the project development phase. As stated in deliverable 7.1: *“Once the data should enter the EURECA project this pseudonymised data set is being de-identified. The de-identification mechanism has to be carried out using a state of the art pseudonymisation tool”*. Each data file that enters this domain is transformed by the CATS service before it can be stored in the data

⁴² Fellegi, Ivan, Sunter, Alan (December 1969), "A Theory for Record Linkage". Journal of the American Statistical Association 64 (328): pp. 1183–1210

⁴³ Jaro, M. A. (1989), "Advances in record linkage methodology as applied to the 1985 census of Tampa Florida". Journal of the American Statistical Society 84 (406): 414–20

⁴⁴ Winkler, W. E. (1990), "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage". Proceedings of the Section on Survey Research Methods (American Statistical Association): 354–359.

warehouse. Using this approach, the research data warehouses contain only anonymous data, meeting the legal requirements for EURECA.

3.4.3 Shibboleth

The Shibboleth⁴⁵ System is a standards-based, open source software package for web Single Sign-On (SSO) across or within organizational boundaries. The primary function of the Shibboleth system is to support identity federation between multiple sites using the SAML⁴⁶ protocol standard. Shibboleth's added value lies in support for privacy, business process improvement via user attributes, extensive policy controls, and large-scale federation support via metadata. Hence Shibboleth accommodates richer and more complex metadata distributed by a federated operator. It has more refined capabilities for managing trust implicit in larger communities. It allows users and enterprises to manage attribute releases, reflecting the greater number and variety of participants.

The implementation part of Shibboleth offers an implementation of three main components meeting the SSO profile and protocol requirements:

- The Identity Provider (IdP) is an entity that authenticates principals and produces assertions of authentication and attribute information.
- A Service Provider (SP) is an entity that gives access to resources.
- Next to these two components there is also an optional Discovery Service component. This component can keep track (in case of multiple IdPs) of the IdP that was selected by a user, using a browser cookie.

Shibboleth is developed by Internet 2⁴⁷, a networking consortium containing people of different domains (communities, industry and government). The main objective of Internet 2 is to develop and maintain a leading-edge network.

In the EURECA platform, Shibboleth will be the central authentication component. When a user wants to use a service of the EURECA platform that is protected by access control, he is redirected to the Shibboleth identity provider where he can authenticate him/herself. This IdP invokes and sends a security token that is validated by the services. Next to the security token, the IdP can also send additional authorisation attributes in the responses to the services.

3.4.4 Security Token Service

A secure token service (STS) is a web service implementation of the WS-Trust⁴⁸ specification for issuing security tokens. These tokens can be used for authentication in other services. This allows for authentication to be centralized, making authentication easier and more secure.

In EURECA an STS will fill the gap of missing authentication functionality in Shibboleth. The current version of Shibboleth is focused on web browser applications, meaning that there is limited support for web service standards like the WS-* specifications.

⁴⁵ Shibboleth, <http://shibboleth.net/> [8 February 2013]

⁴⁶ OASIS SAML, <https://www.oasis-open.org/committees/download.php/27819/sstc-saml-tech-overview-2.0-cd-02.pdf> [8 February 2013]

⁴⁷ Internet2, <http://www.internet2.edu/> [8 February 2013]

⁴⁸ WS-Trust, 2007, "WS-Trust 1.3", available from: <http://docs.oasis-open.org/ws-sx/ws-trust/200512/ws-trust-1.3-os.pdf> [1 February 2013]

Several third-party implementations of the STS WS-Trust specification are available like Metro STS⁴⁹ and CXF STS⁵⁰.

3.4.5 A XACML Engine

A XACML Decision Engine provides an authorisation service by making XACML access control decisions for incoming XACML access requests. These decisions are the result of evaluating user defined XACML policies with the incoming requests.

The engine is an implementation of the OASIS XACML⁵¹ de-facto standard, meaning it provides complete support for all the mandatory features of XACML. Specifically, there is support for parsing policy and request/response files, decision making for incoming requests using the policies and determining applicability of policies. The engine will also support some specific features like role based access control profile of XACML.

The XACML Decision Engine will be the central component for authorisation in the EURECA platform. All access control requests are evaluated by this engine using the policies that contain the EURECA access rules determined in the security model.

The standard XACML functionality of the engine can be provided by a third-party implementation like Sun⁵² or JBOSS⁵³ XACML engine. Next to this standard functionality, the engine will probably need extensions in order to provide support for missing functionality.

3.4.6 LDAP

LDAP is an application protocol that can be used to access and maintain distributed directory information services over an internet protocol network. The core of the protocol is defined by the Internet Engineering Task Force (IETF) in RFC4510⁵⁴. The directory information services contain information that is organized in a hierarchical directory structure. This information can be queried and filtered so that only the required information is returned. LDAP directory is a "write once, read many times" service.

LDAP is a good solution to use as user credential store as part of the user management services in EURECA (mainly due to the presence of the flexible password policies, which take a high implementation effort when building from scratch). The most important aspect of LDAP is the possibility to have fine-grained control over the use of passwords. This means that users with a higher degree of access can be forced to have more secure passwords. The passwords itself are managed with password policies. LDAP is a good solution to use as user credential store as part of the user management services in EURECA. The flexible password policies, which take a high implementation effort when building from scratch). For this the EURECA platform will include an already existing implementation of the LDAP protocol like OpenDS⁵⁵ or OpenLDAP⁵⁶.

⁴⁹ Metro STS, available at: <http://metro.java.net/> [8 February 2013]

⁵⁰ CXF STS, available at: <http://cxf.apache.org/> [8 February 2013]

⁵¹ OASIS XACML, http://docs.oasis-open.org/xacml/2.0/access_control-xacml-2.0-core-spec-os.pdf [8 February 2013]

⁵² Sun XACML, <http://sunxacml.sourceforge.net/> [8 February 2013]

⁵³ JBoss, <http://www.jboss.org/> [8 February 2013]

⁵⁴ IETF, LDAP Technical Specification Road Map, 2006, <http://tools.ietf.org/html/rfc4510> [8 February 2013]

⁵⁵ OpenDS, Open Source Java LDAP Directory Service, <http://www.opensds.org/> [8 February 2013]

⁵⁶ OpenLDAP, <http://www.openldap.org/> [8 February 2013]



To manage all of the EURECA principals, LDAP (Lightweight Directory Access Protocol) can be used. As principal management is needed in EURECA as well, this component can be re-used for the EURECA platform.

4 External Data Sources and Knowledge Bases

4.1 Sources of Patient Data

The development and investigation of medical applications requires a variety of patient data from electronic health records (EHR) or clinical records. However, in practice, access to patient data is heavily regulated to avoid unauthorized access, due to ethical concerns about patient privacy. Thus, many researchers and developers find it difficult to acquire patient data required to test and validate their research and tools. The disclosure of patient data for peer review and experimental reproducibility, even in de-identified form, is problematic when publishing results in scientific journals. For this reason, test data that is already publicly available can be useful for both prototyping and publishing.

The cost of acquiring and extracting specific types of patient data for the testing and validation of clinical trial and EHR software can be prohibitive. Certain types of data are even impossible to extract because, for example, not all data needed to establish a particular type of eligibility are normally collected. A specialized data generator can address the problems of availability and extractability, without posing the legal and ethical challenges of patient data. We will briefly describe a few potential sources of patient data, including the Advanced Patient Data Generator (APDG) from EURECA partner Vrije Universiteit in Amsterdam.

4.1.1 Cypress

Cypress⁵⁷ is an open source clinical quality measure testing tool that automates the validation of clinical quality measure calculations. Cypress is the rigorous and repeatable testing tool of Electronic Health Records (EHRs) and EHR modules in calculating Meaningful Use (MU) Stage 2 Clinical Quality Measures (CQMs). The U.S. Office of the National Coordinator for Health IT (ONC) sponsored the development of Cypress by MITRE and it now serves as the official testing tool used by the Authorized Testing Labs in the 2014 EHR Certification program. As part of the cypress project, 215 sample test patients⁵⁸ are available in the HL7 CDA format (XML). As part of a hackathon day at the Semantic Web Applications and Tools for the Life Sciences (SWAT4LS) Workshop in Paris 2012, the test patient data was converted to RDF and several SPARQL queries were able to identify patients using some criteria⁵⁹.

4.1.2 Synapse

Synapse⁶⁰ is an approach to sharing data, models, and analysis developed by Sage Bionetworks. Synapse is a collaborative compute space that allows scientists to share and analyse data together. Data is clearly labelled with terms of use and accessible via a Web Client, as well as via the R Client. Synapse consists of a web portal integrated with

⁵⁷ <http://projectcypress.org/>

⁵⁸ <https://github.com/projectcypress/test-deck>

⁵⁹ http://www.w3.org/wiki/HCLS/SWAT4LS2012/Hackathon/TrialProtocols#RDF_and_SPARQL_against_CDA_Patient_Documents

⁶⁰ This section is based on text from the web pages of <https://synapse.prod.sagebase.org>.

the R/Bioconductor statistical package (Synapse will be integrated with additional tools in the future). The web portal is organized around the concept of a Project which is an environment where you can interact, share data, and analysis methods with a specific group of users or broadly across open collaborations. Projects provide an organizational structure to interact with data, code and analyses, and to track data provenance. A project can be created by anyone with a Synapse account and can be shared among all Synapse users or restricted to a specific team. You have access to any Public data and data in Private projects that have been shared with you. Of note among public data are projects such as the Synapse Commons Repository (SCR) (syn150935) and the metaGenomics project (syn275039). The SCR provides access to raw data and phenotypic information for publicly available genomic data sets, such as GEO, ArrayExpress, and TCGA.

4.1.3 Clinical Avatars

The Laboratory for Personalized Medicine⁶¹ (LPM) at Harvard Medical School has developed a methodology for creating virtual representations of people for the purpose of conducting personalized medicine simulations. We call these virtual representations "Clinical Avatars"⁶². Avatars can be configured so that their statistical distribution matches the requirements of a particular population. Pre-set and example configurations are provided, and advanced users can use our application to create avatar configurations based on tabular data that they upload. These avatars can then be directly used in simulations. The application can export both the conditional probability tables (CPT) as XML and generated patient avatars as tsv (tab-separated values), which can be read in a spreadsheet program.

Initially, LPM has used clinical avatars to better understand complicated genomic-based drug dosing regiments. We selected the drug warfarin as our first use-case example.

Current Limitations

1. The population size is restricted to 5,000 avatars
2. Clinical trial simulation is 20 avatars, limited due to the computational complexity

4.1.4 Advanced Patient Data Generator (APDG)

The Advanced Patient Data Generator provides a knowledge-based approach to synthesizing large scale patient data. The basic rationale for this synthesis is to make the generated patient data as realistic as possible, by using domain knowledge to control the patient data generation. That domain knowledge is collected from biomedical publications such as those listed in PubMed, medical textbooks, and medical web sites.

That knowledge is formalized in the Patient Data Definition Language (PDDL) for patient data generation. We have used the APDG system to generate large scale data for breast cancer patients as test data for the SemanticCT system, a semantically-enabled system for clinical trials. This data has been mapped⁶³ and loaded into the EURECA Common Data Model (CDM).

⁶¹ <http://lpm.hms.harvard.edu/>

⁶² This section is based on text from the web page at <http://clinicalavatars.org>.

⁶³ http://atlas.ics.forth.gr/EURECA/wiki/index.php/T4.2_-_Mapping_formalism_and_mappings_between_the_core_data_set_and_EHR_and_CT_models#APDG_mappings_to_EURECA_CDM

PDDL is an XML-based language designed to define the general format of patient data and its relevant domain knowledge to control the procedure of patient data generation. PDDL uses the general structure 'Session-Archetype-Slot' for patient data generation. In PDDL, each entity (i.e., session, archetype, or slot) has a value property to define the entity name. An archetype is allowed to contain other (non-recursive) archetypes or slots. Slots are termination tabs which are used to state the reference models to define possible values of slots, like these:

```
<Session value="BasicData">
  <Archetype concept = "Patient">
    <Slot value="PatientID" type="string"/>
    <Slot value="Gender"/>
    <Slot value="BirthYear">
  </Session>
```

The data distribution is defined inside the DataRange with the special tab 'distribution'. The distribution value takes a real number between 0 and 100, like this:

```
<Slot value="Gender">
  <DataRange>
    <enumeration value="female"/>
    <enumeration value="male"/>
    <Distributions type="enumeration">
      <Distribution item="female" pfrom="0" pto="100"/>
      <Distribution item="male" pfrom="0" pto="0"/>
    </Distributions>
  </DataRange>
</Slot>
```

The distribution can be stated by its distribution type (e.g., uniform random or normal distribution) on an enumeration set, like those in the following example:

```
<Slot value="DiagnosisMonth" type="month">
  <DataRange>
    <Distributions type="enumeration">
      <Distribution disttype="uniformrandom"
        set="1,2,3,4,5,6,7,8,9,10,11,12"/>
    </Distributions>
  </DataRange>
</Slot>
```

or by stating a data range (with a type) over the distribution, like this:

```
<Slot value="BirthYear">
  <DataRange>
    <Distributions type="year" variable="$birthyear">
      <Distribution disttype="uniform" datatype="minmax(int)"
        data="1913,2013"/>
    </Distributions>
  </DataRange>
</Slot>
```

The condition statements are used to state the conditions which depend on some variables which have been defined in the previous distributions slots, like this:

```
<Slot value="MenopausalStatus">
  <DataRange>
    <Distributions type="enumeration"
variable="$menopausalstatus">
    <Distribution item="premenopausal" pfrom="0" pto="100"
condition="$birthyear >1970"/>
    <Distribution item="perimenopausal" pfrom="0" pto="80"
condition="$birthyear =<1970 AND
$birthyear >=1950"/>
    <Distribution item="postmenopausal" pfrom="80" pto="100"
condition="$birthyear =< 1970 AND
$birthyear >= 1950"/>
    <Distribution item="postmenopausal" pfrom="0" pto="100"
condition="$birthyear < 1950"/>
    </Distributions>
  </DataRange>
</Slot>
```

The tab 'ConceptMapping' is designed to map the PDDL entity into their corresponding concepts in ontologies. For example, the following statement states that the slot 'gender' has the concept in the ontology SNOMED with a concept id '263495000'.

```
<Slot value="Gender">
  <ConceptMapping ontology="snomed" conceptid="263495000"/>
  <DataRange>
    <enumeration value="female"/>
    <enumeration value="male"/>
  </Distributions>
  </DataRange>
</Slot>
```

In an attempt to create more realistic data, we are planning to create a special distribution type for customized distributions that will be based on value distributions from actual patient data at a given EURECA partner. For example, by collecting statistics about the Menopausal Status, Currently Pregnant, Currently Nursing, Hispathology, HER2, ER, PR, Stage, Tumour Size, Lymph Nodes, and Distant Metastases at the MAASTRO Clinic, we can create a specialized profile for APDG that reflects the types of patients that we encounter at the MAASTRO Clinic. Such a customized distribution might also be built from a Cancer Registry, provided that EURECA members are permitted access. We also want to create applications that can deal with missing data.

4.2 ClinicalTrials.gov

ClinicalTrials.gov⁶⁴ is a registry and results database of publicly and privately supported clinical studies of human participants conducted around the world. The U.S. National Institutes of Health (NIH), Department of Health and Human Services, through its

⁶⁴ <http://ClinicalTrials.gov>

National Library of Medicine (NLM), has developed ClinicalTrials.gov to provide patients, their family members, health care professionals, researchers, and the public with easy access to information on publicly and privately supported clinical studies on a wide range of diseases and conditions. These data are provided to the National Library of Medicine by organisations and institutions that sponsor and implement the studies. The web site itself and the web services are maintained by the National Library of Medicine at the National Institutes of Health. Because ClinicalTrials.gov is a government web site, it does not host, or receive funding from, advertising or the display of commercial content.

4.2.1 Data in the registry

At the time of writing, the registry contains 141,132 clinical studies. The studies are located in 182 different countries around the world, with the oldest trial being from 1999 and the latest trials from 2013. The registry is continually updated by sponsors and principal investigators of the studies. While ClinicalTrials.gov does not contain all clinical studies because not all studies are required by law to be registered, it is by far the most comprehensive registry of clinical trials available, and therefore a valuable resource to EURECA.

ClinicalTrials.gov organizes information for each registered study as an integrated unit, displaying the study protocol information and, if available, the corresponding results information.

Each ClinicalTrials.gov record presents summary information about a study protocol and includes the following:

- Disease or condition
- Intervention (for example, the medical product, behaviour, or procedure being studied)
- Title, description, and design of the study
- Requirements for participation (eligibility criteria)
- Locations where the study is being conducted
- Contact information for the study locations
- Links to relevant information on other health web sites, such as NLM's MedlinePlus for patient health information and PubMed for citations and abstracts for scholarly articles in the field of medicine.

Some records also include information on the results of the study, such as:

- Description of study participants (that is the number starting and completing the study and their demographic data)
- Outcomes of the study
- Summary of adverse events experienced by study participants

Within EURECA, ClinicalTrials.gov will be used as an external relevant information source for the matchmaking algorithms as part of the personal medical recommender systems.

4.2.2 LinkedCT

The data on ClinicalTrials.gov described above has been translated into a structured format for Web data, RDF, and published as Linked Data on LinkedCT.org⁶⁵. The data can also be accessed through a SPARQL endpoint or as an RDF data dump. A script is continuously running to translate data from ClinicalTrials.gov, so technical difficulties aside we can expect LinkedCT to be as up to date as ClinicalTrials.gov. Considering the support for SPARQL queries in the EURECA architecture, it is likely that EURECA will access trial information through LinkedCT rather than through ClinicalTrials.gov directly.

4.2.3 clinicaltrialsregister.eu

The website clinicaltrialsregister.eu⁶⁶ hosts the EU Clinical Trials Registry, and can be seen as the European equivalent of ClinicalTrials.gov. It maintains a registry of all trials that are conducted at sites within the European Union (and trials outside the EU that are part of a Paediatric Investigation Plan). The registry currently contains 20.016 trials conducted between 2004 and the present. The data in the registry is based on information provided by trial organizations as part of their application to a national medicine regulatory authority for authorization to conduct a trial.

While the EU Clinical Trials Registry is smaller than ClinicalTrials.gov, it does contain trials that are not registered at ClinicalTrials.gov. Therefore, it is a valuable source of trial information for EURECA.

4.3 BioPortal SPARQL endpoint

Bioportal is an open repository of biomedical ontologies created and maintained by the National Center for Biomedical Ontology (NCBO)⁶⁷. For a description of its content and relevance to EURECA, we refer to deliverable D4.1 *“Requirements analysis and selection of the initial clinical scenarios for core datasets”*. In the present document, we focus on the *services* it provides. In particular, we are interested in the functionality it offers to access the ontologies and the links between them. As said in D4.1, Bioportal can be accessed in three ways. First, there is a website⁶⁸ on which users can search and browse the ontologies, mappings and annotated resources. Second, there are several RESTful services to request or search ontology content, concepts and terms. Third, there is a SPARQL endpoint⁶⁹ through which the data can be queried with the semantic web query language SPARQL. Note that EURECA plans to use only SOAP services so the REST interface will not be used. The SPARQL endpoint option provides the best functionality because it enables access to an RDF representation through complex and highly structured queries.

Querying ontologies through the Bioportal SPARQL endpoint

In EURECA we foresee three types of access to the Bioportal terminologies.

⁶⁵ <http://linkedct.org/>

⁶⁶ <http://clinicaltrialsregister.eu/>

⁶⁷ <http://www.bioontology.org/>

⁶⁸ <http://bioportal.bioontology.org/>

⁶⁹ <http://sparql.bioontology.org>

1. Retrieving all concepts of a given terminology.
2. Retrieving the sub- or super-concepts of a given concept.
3. Retrieving a concept URI based on its label

Figure 4-1, Figure 4-2 and Figure 4-3 show example SPARQL queries for each of the three above types of access. These examples are taken and/or modified from <http://sparql.bioontology.org/examples>.

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>

SELECT DISTINCT ?s ?label WHERE {
  GRAPH <http://bioportal.bioontology.org/ontologies/SNOMEDCT> {
    ?s a owl:Class .
    ?s skos:prefLabel ?label
  }
}
LIMIT 10
```

Figure 4-1: SPARQL query to retrieve classes and their preferred labels from SNOMED

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX snomed-term: <http://purl.bioontology.org/ontology/SNOMEDCT/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
SELECT DISTINCT ?x ?label
FROM <http://bioportal.bioontology.org/ontologies/SNOMEDCT>
WHERE
{
  ?x rdfs:subClassOf snomed-term:363664003 .
  ?x skos:prefLabel ?label.
}
LIMIT 10
```

Figure 4-2: SPARQL query to retrieve the subclasses of a SNOMED term, along with their labels.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT DISTINCT *
FROM <http://bioportal.bioontology.org/ontologies/SNOMEDCT>
FROM <http://bioportal.bioontology.org/ontologies/globals>
WHERE
{
  ?x rdfs:label ?label .
  FILTER (CONTAINS ( UCASE (str(?label)), "MELANOMA" ) )
}
LIMIT 10
```

Figure 4-3: SPARQL query to retrieve the subclasses of a SNOMED term, along with their labels

Querying mappings through the Bioportal SPARQL endpoint

In addition to querying the ontologies themselves, it is possible to query the mappings between ontological terms using the SPARQL language. We provide an example query that retrieves all mappings and their targets for an obsolete SNOMED concept:

```
PREFIX maps:
<http://protege.stanford.edu/ontologies/mappings/mappings.rdfs#>
SELECT DISTINCT * WHERE {
  ?s maps:source
  <http://purl.bioontology.org/ontology/SNOMEDCT/164075007>;
  maps:target ?target . }
LIMIT 10
```

Figure 4-4: SPARQL query to retrieve concepts mapped to SNOMED concepts

Usage data of the Bioportal SPARQL endpoint

The open availability of the Bioportal vocabularies, and the fact that they are available in a structured format on the Web, means that they *can be used* for a wide variety of purposes in EURECA and other projects. The USEWOD workshop⁷⁰ series studies how this type of data is *actually used*. This year, USEWOD will make usage data of the Bioportal SPARQL endpoint available for people to study. We aim to use this dataset to learn how often each vocabulary is requested, and how the mappings between them are used.

4.4 Trial feasibility data sources

The goal of the trial feasibility scenario is to assess whether a sufficient number of patients can be recruited in a specified timeframe given a trial proposal and a selected set of recruitment sites. The main sources for this determination are the (clinical) data sources of the respective recruitment sites. However, sometimes criteria relate to conditions which are not readily available in these data sources.

In order to get a prediction of the enrolment rates when no relevant patient data is available, the optional use cases UC.TS.PF.11 (*Compute eligibility criterion probability*) and UC.TS.PF.13 (*Compute trial path probability*) allow the use of external data sources to obtain probabilities.

Relevant external data sources for these optional use cases will be extended when it is further clarified which cancer will be used as carrier for the development

4.4.1 Literature

- Pubmed⁷¹: PubMed comprises more than 22 million citations for biomedical literature from MEDLINE, life science journals, and online books
- OMIM - Online Mendelian Inheritance in Man⁷²: OMIM is a comprehensive, authoritative, and timely compendium of human genes and genetic phenotypes.

⁷⁰ <http://data.semanticweb.org/usewod/2013/>

⁷¹ <http://www.ncbi.nlm.nih.gov/pubmed>

⁷² <http://omim.org/>

4.4.2 Cancer registries

A cancer registry is⁷³ an information system designed for the systematic collection, management, and analysis of data on persons with the diagnosis of a malignant or neoplastic disease (cancer). It should be investigated during the course of EURECA whether (1) the cancer registries allow access to their databases, and (2) the cancer registry databases provide sufficient data for stratification. Two relevant organizations/programs would be:

- The Surveillance, Epidemiology and End Results (SEER) Program⁷⁴, for cancer statistics in the United States. The SEER Program registries routinely collect data on patient demographics, primary tumor site, tumor morphology and stage at diagnosis, first course of treatment, and follow-up for vital status.
- The International Association of Cancer Registries⁷⁵, a professional society dedicated to fostering the aims and activities of cancer registries worldwide. It is primarily for population-based registries, which collect information on the occurrence and outcome of cancer in defined population groups (usually the inhabitants of a city, region, or country). For each new cancer case, registries record details of the affected individual, the nature of the cancer, information on treatment, and on follow-up especially with respect to survival from the disease.

4.4.3 Other relevant resources

Another resource that could be relevant is the FDA Adverse Event Reporting System (FAERS)⁷⁶, a database that contains information on adverse event and medication error reports submitted to FDA. The database is designed to support the FDA's post-marketing safety surveillance program for drug and therapeutic biologic products. The FAERS provides access to individual case safety reports which can be mined. A similar database in Europe⁷⁷ collects suspected side effects is derived from EudraVigilance⁷⁸, a European Medicines Agency database designed for collecting reports on suspected side effects.

⁷³ <http://www.ncra-usa.org/i4a/pages/index.cfm?pageid=3301#sub1>

⁷⁴ <http://seer.cancer.gov/>

⁷⁵ <http://www.iacr.com.fr/>

⁷⁶

<http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm>

⁷⁷ <http://www.adrreports.eu/EN/index.html>

⁷⁸

http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/document_listing/document_listing_000239.jsp

5 Conclusion

A wide variety of tools and data sources are available that are potentially useful for building functionality for the EURECA use cases. This deliverable is a snapshot of the components and resources of which we are aware, either already in use or under consideration by EURECA partners. Other types of resources, such as those from similar projects, are also listed. Although the list is not complete, it should help partners to understand each other, establish commonality, and provide a glimpse of the state of the art.