Deliverable No. 4.2
Requirements for Semantic Access to Clinical Trial Data and HIS

Grant Agreement No.: 270089
Deliverable No.: D4.2
Deliverable Name: Requirements for Semantic Access to Clinical Trial Data and HIS
Contractual Submission Date: 31/01/2012
Actual Submission Date: 31/01/2012

<table>
<thead>
<tr>
<th>Dissemination Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PU</td>
<td>Public</td>
</tr>
<tr>
<td>PP</td>
<td>Restricted to other programme participants (including the Commission Services)</td>
</tr>
<tr>
<td>RE</td>
<td>Restricted to a group specified by the consortium (including the Commission Services)</td>
</tr>
<tr>
<td>CO</td>
<td>Confidential, only for members of the consortium (including the Commission Services)</td>
</tr>
</tbody>
</table>
**ABSTRACT:**
This deliverable reviews semantic aspects that are relevant for the p-medicine platform to interact with external health information systems, as hospital information systems and clinical trial repositories. First a state-of-the-art review on semantic resources that are commonly used in health care and clinical research is provided and their usage in health information systems is described. Then current approaches regarding the secondary usage of data from hospital information systems and clinical trial repositories are reviewed. Furthermore, based on the state-of-the-art analysis, an initial description of the p-medicine semantic layer is provided that lays the foundation for the p-medicine platform to interact with external health information systems and integrate their data semantically.

**KEYWORD LIST:** hospital information system, secondary usage of healthcare data, electronic health record, semantic interoperability, clinical trial management system

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no 270089.

The author is solely responsible for its content, it does not represent the opinion of the European Community and the Community is not responsible for any use that might be made of data appearing therein.

---

1 $R=\text{Report}, \ P=\text{Prototype}, \ D=\text{Demonstrator}, \ O=\text{Other}$

2 $PU=\text{Public}, \ PP=\text{Restricted to other programme participants (including the Commission Services)}, \ RE=\text{Restricted to a group specified by the consortium (including the Commission Services)}, \ CO=\text{Confidential, only for members of the consortium (including the Commission Services)}$
**MODIFICATION CONTROL**

<table>
<thead>
<tr>
<th>Version</th>
<th>Date</th>
<th>Status</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>05/07/11</td>
<td>First proposal for ToC</td>
<td>Gabriele Weiler</td>
</tr>
<tr>
<td>0.5</td>
<td>05/12/11</td>
<td>First Draft</td>
<td>Gabriele Weiler, Ulf Schwarz, Lefteris Koumakis, Alberto Anguita, Miguel García, Gaston Tagni, Corina Strössner</td>
</tr>
<tr>
<td>0.6</td>
<td>22/12/11</td>
<td>Draft, Chapter 3 and Conclusion updated</td>
<td>Gabriele Weiler, Ulf Schwarz, Stelios Sfakianakis, Corina Strössner</td>
</tr>
<tr>
<td>0.7</td>
<td>23/12/11</td>
<td>Draft, Chapter 4 and Conclusion updated</td>
<td>Anca Bucur, Aisan Maghsoodi, Gabriele Weiler</td>
</tr>
<tr>
<td>1.0</td>
<td>10/01/12</td>
<td>Draft, Formatting harmonized, Chapter 6 and Conclusion updated</td>
<td>Gabriele Weiler</td>
</tr>
<tr>
<td>1.3</td>
<td>17/01/12</td>
<td>Pre-Final for internal review</td>
<td>Gabriele Weiler, Fatima Schera</td>
</tr>
<tr>
<td>1.5</td>
<td>27/01/12</td>
<td>Final</td>
<td>Gabriele Weiler</td>
</tr>
</tbody>
</table>

List of contributors:
- Gabriele Weiler, IBMT
- Ulf Schwarz, IFOMIS
- Corina Strössner, IFOMIS
- Lefteris Koumakis, Forth
- Stelios Sfakianakis, Forth
- Alberto Anguita, UPM
- Miguel García, UPM
- Gaston Tagni, Philips
- Anca Bucur, Philips
- Aisan Maghsoodi, Philips
- Fatima Schera, IBMT
- Stephan Kiefer, IBMT
- Holger Stenzhorn, USAAR
- Benjamin Jefferys, UCL
Contents

CONTENTS.................................................................................................................................. 4
1 EXECUTIVE SUMMARY .................................................................................................................... 6
2 INTRODUCTION................................................................................................................................ 7
3 SEMANTIC RESOURCES AND HOSPITAL INFORMATION SYSTEMS .................................................. 9
  3.1 Semantic Resources...................................................................................................................... 9
    3.1.1 ICD - International Statistical Classification of Diseases and Related Health Problems .................................................................................................................. 9
    3.1.2 ICHI / ICPM and OPS.......................................................................................................... 12
    3.1.3 DICOM- Digital Imaging and Communications in Medicine .................................................. 12
    3.1.4 SNOMED CT...................................................................................................................... 13
    3.1.5 Health Level Seven (HL7) ................................................................................................. 14
  3.2 Exploitation of Semantic Resources in Hospital Information Systems ........................................... 15
    3.2.1 Current Situation in Hospital Information Systems .......................................................... 16
    3.2.2 openEHR .......................................................................................................................... 17
    3.2.3 CEN 13606, Electronic Health Record Communication ................................................... 18
    3.2.4 Integrating the Health Care Enterprise (IHE) ..................................................................... 20
    3.2.5 HL 7 CDA, CCD ............................................................................................................. 21
  3.3 Conclusion.................................................................................................................................. 22
4 SEMANTIC RESOURCES IN CLINICAL RESEARCH ................................................................. 25
  4.1 Semantic Resources...................................................................................................................... 25
    4.1.1 LOINC .............................................................................................................................. 25
    4.1.2 Gene Ontology .................................................................................................................. 26
    4.1.3 MedDRA .......................................................................................................................... 27
    4.1.4 MIAME/Mage ................................................................................................................... 29
  4.2 Exploitation of Semantic Resources in Clinical Research ............................................................ 29
    4.2.1 Current Clinical Trial Management Systems ..................................................................... 30
    4.2.2 Clinical Trial Repositories ............................................................................................... 32
    4.2.3 CDISC ............................................................................................................................. 33
    4.2.4 BRIDG ............................................................................................................................. 33
    4.2.5 ObTiMA ........................................................................................................................... 35
    4.2.6 Transcend ....................................................................................................................... 36
  4.3 Conclusion.................................................................................................................................. 37
5 REUSING EHR / CTMS DATA IN DATA WAREHOUSES AND FOR DATA MINING .................... 40
  5.1 Reusing EHR data ..................................................................................................................... 40
  5.2 Reusing CTMS Data ................................................................................................................... 41
  5.3 DARTNet .................................................................................................................................... 41
  5.4 cabIG Clinical Trial Suite ........................................................................................................... 43
    5.4.1 cabIG Clinical Connector .................................................................................................... 43
    5.4.2 cabIG® Integration Hub ..................................................................................................... 44
  5.5 EHR4CR ..................................................................................................................................... 45
  5.6 Information Warehouse ............................................................................................................. 46
  5.7 Stanford Translational Research Integrated Database Environment (STRIDE) .............................. 47
6 LINKING ELECTRONIC HEALTH RECORDS AND TRIAL MANAGEMENT SYSTEMS .................. 50
  6.1 Extracting Data from Electronic Health Records to Clinical Trial Management Systems .......... 50
    6.1.1 Direct Extraction .............................................................................................................. 50
    6.1.2 Extraction and Investigator Verification ............................................................................ 51
  6.2 Single Source Concept .............................................................................................................. 53
  6.3 Reusing EHR Data for the Recruitment of Patients for Clinical Trials ........................................ 55
  6.4 Initiatives to fully Integrate Healthcare and Clinical Trial Systems ............................................ 56
    6.4.1 EHRCR Project .............................................................................................................. 56
    6.4.2 Healthcare Link Initiative ............................................................................................... 58
    6.4.3 Clinical Research Value Case Workgroup ......................................................................... 59
  6.5 Conclusion.................................................................................................................................. 60
CONCLUSION: APPROACH TAKEN BY P-MEDICINE

7.1 The p-medicine Semantic Layer
7.1.1 The Health Data Ontology Trunk
7.1.2 The Ontology Annotator
7.1.3 The ObTiMA Ontology-based CRF Creator
7.1.4 The Data Translator

7.2 Approach for Push and Sync Services
7.2.1 Push Services
7.2.2 Sync Services

Appendix 1 - Abbreviations and acronyms
1 Executive Summary

The p-medicine platform will provide the possibility to interact with clinical trial repositories and hospital information systems via the sync and the push services. The push services will retrieve data from clinical information systems into the data warehouse for further research. The sync services will allow to reuse data from Hospital Information Systems in running clinical trials in the ontology based trial management system ObTiMA.

In order to reuse data from external repositories and information systems in the p-medicine platform, the semantics of the data needs to be clearly defined. Hence, in this deliverable we will lay the foundation to provide the semantic means to develop the sync and the push services. Firstly, we will review semantic resources that are commonly used in hospital information systems and clinical research and describe their exploitation. Then we will describe the state of the art regarding the push services, i.e. approaches to reuse EHR and CTMS data in data warehouses and for data mining. Furthermore, we will describe the state of the art regarding the sync services, i.e. approaches that realize the linking of CTMS and EHRs. We conclude this deliverable by providing an initial description of the p-medicine semantic layer that lays the foundation for the push- and the sync services. A detailed description of the semantic layer will be given in D4.3 taking into consideration the state of the art review presented in this deliverable.
2 Introduction

Nowadays, a tremendous amount of data is collected and stored in different heterogeneous hospital information systems, electronic health records and clinical trial information systems. This data provides a precious resource for various exploitations in clinical research, especially in the field of personalized medicine. Integrating this data offers an enormous chance for novel, individualized treatments for patients that vastly advance their prognosis and outcome. Although several researchers have endorsed such secondary usage of clinical data, this potential remains largely untapped and the data is rarely looked at after the process it was originally collected for is completed, as e.g. the direct course of patient care or the clinical trial. One of the reasons for this fact is that clinical data are stored in different formats by different hospitals and information systems and that the semantics of the data is mostly not clearly defined, making its reuse difficult or even impossible [1.1], [1.2].

The p-medicine project aims to overcome these challenges and to provide a platform that facilitates to exploit the data stored in the disparate data sources in hospitals. In particular, in this direction two scenarios will be realized that will foster and support the advancement of personalized medicine (s. Deliverable D2.2):

On the one hand, p-medicine will enable the syntactic and semantic integration of data from different biomedical data sources, such as in particular clinical trials and electronic patient records from hospital information systems, syntactically and semantically in a secure and scalable data warehouse. The data warehouse will store, manage and integrate large data sets and will provide the main resources to reuse the data for knowledge discovery, VPH modelling, data mining and simulations and decision support scenarios. All kind of analysis services developed in p-medicine can exploit the data stored in the data warehouse seamlessly. Tools will be provided, which allow owners of data (i.e. clinicians, trial chairmen) to upload their data into the data warehouse, in order to make so far unexploited data resources available to research.

On the other hand, p-medicine will support the secondary usage of data from hospital information systems in running clinical trials, which are conducted with p-medicine's trial management system ObTiMA. The p-medicine platform will enable that patient CRFs in ObTiMA can be filled automatically with data that was previously collected in hospital information systems to avoid double data entry and to enhance clinical trial processes.

In order to enable the two scenarios the data from the external heterogeneous medical information systems needs to be integrated semantically. This will be realized by the semantic framework that will be developed in WP 4. In this deliverable we lay the foundations for that work and review state of the art work to foster the semantic interoperability of health care systems. We especially focus on reviewing semantic resources that are used in hospital information systems and clinical trial management systems and their exploitation. Based on this review, the deliverable will describe the approach taken by p-medicine to enable the reuse of HIS and CTMS data. For that purpose a semantic layer will be developed by WP 4. This layer will provide the foundation for the p-medicine platform to interact with external health information systems, integrate their data semantically and provide the base to realize the two scenarios described above. The deliverable is structured as follows:

Chapter 3 describes semantic resources that are commonly used in patient care and their exploitation in hospital information systems and electronic health records.

Chapter 4 describes semantic resources that are commonly used in clinical trials, as e.g. in clinical trial management systems and clinical trial repositories.

Chapter 5 describes state of the art work regarding the reuse of CTMS and HIS data in data warehouses and for data mining. The focus is in particular on the semantic issues.
Chapter 6 provides a state of the art analysis of approaches and research projects that focus on reusing data from hospital information systems in clinical trial management systems to enhance clinical trial processes. We especially concentrate on approaches that avoid double data entry into these systems.

Chapter 7 concludes the deliverable by outlining the approach taken by p-medicine. The chapter gives an initial idea of the p-medicine semantic layer and describes the approach to reuse CTMS and HIS data.

References


3 Semantic Resources and Hospital Information Systems

In this section we describe semantic resources that are commonly used in hospital information systems, discussing also their exploitation in these systems. Generally speaking, with "semantic resources" we mean those standards for the bio-medical domain in which the content of the information is been explicitly stated. Indeed, in order to integrate and compare different data coming from different information systems, semantic agreements must be stated between different agents’ communities. Among such semantic resources ontologies play a specific role. Unlike terminologies, ontologies provide a high degree of semantic specification of terms both in human readable text and in computer language, facilitating the semantic interoperability between different information systems.

3.1 Semantic Resources

The semantic resources have been reviewed extensively in deliverable D4.1. In the following we describe them briefly concentrating on aspects relevant for their exploitation in Hospital Information Systems (HIS). In this context we pay close attention to the current initiatives to provide interoperability standards for the capture, representation and communication of clinical data. Electronic Health Records (EHRs) are of course in the focus of these initiatives.

Whenever clinicians or biomedical professionals produce information they generate semantic content, which is usually encoded in biomedical language, but up to date often in an unstructured free text form. Initiatives to standardize such semantic content of clinical data operate on three different levels on which this semantic content can be distinguished and represented:

- **Level 1**: Generic reference models for representing clinical data, for instance ISO/EN 13606 Part 1 [3.1], HL7 CDA Release 2 [3.7], the openEHR Reference Model [3.2]
- **Level 2**: Agreed clinical data structure definitions, for instance openEHR archetypes [3.3], ISO/EN 13606 Part 2 [3.15], HL7 templates [3.10], user driven generic templates, data schemas and data sets
- **Level 3**: Clinical terminology systems, for instance ICD, SNOMED CT, etc.

Even though it would be highly desirable for making considerable progress towards (semantic) interoperability in and amongst HISs and their components like Radiology Information Systems (RIS) and Picture Archiving and Communication Systems (PACS) semantic content solutions on level 1 are hardly used in running clinical information systems. The same is true for standardized semantic content of level 2 which is generally dominated by the use of locally developed data schema and data sets.

Up to now only semantic resources on level 3 have found a fairly widespread application in HISs and related components. In particular, ICD and, to some degree, SNOMED CT are integrated in clinical systems, but quite often only to enable standardized billing and accounting processes. Even though HL7 is widely used as a (syntactic) message exchange standard its application for the standardisation of the semantic content of such messages has been very limited.

3.1.1 ICD - International Statistical Classification of Diseases and Related Health Problem

The International Statistical Classification of Diseases and Related Health Problem is a systematic classification of diseases, namely a system of categories to which morbidity entities...
are assigned according to established criteria. Such a classification has been sought with the aim to provide a system of comparable nomenclatures for all countries. It has been revised periodically to incorporate changes in the medical field and to date there have been ten versions of it.

Through the use of ICD, diagnoses of diseases and other health problems are translated from words into alphanumeric code, in order to store, retrieve and analyze information in computer systems. The ICD includes also the analysis of the general health situation of different population groups and the monitoring of the incidence and prevalence of diseases and other health problems in relation to other variables, for example the characteristics and circumstances of the individual affected.

ICD contains information related to diagnoses, symptoms, abnormal laboratory findings, injuries and poisonings, external causes of morbidity and mortality, factors influencing health status from all the different branches of medicine: Oncology, Dentistry and Stomatology, Dermatology, Psychiatry, Neurology and so on.

The structure of such a classification was developed by William Farr, who grouped all data in:

1. epidemic diseases
2. constitutional or general diseases
3. local diseases arranged by site
4. developmental diseases

ICD-10

The 10th version of ICD (ICD-10) is a single coded list of three-character categories (from A00 to Z99), each of which can be further divided into up to ten four-character subcategories (for example, A00.0, A02.2, B51.9 and so on). Three main information units are related to each disease, namely Chapter, Block and Title. The classification is divided into 22 chapters and each of them is defined by a letter. Chapters I-XVIII deal with diseases and Chapters XIX-XXII with injuries. Using “blocks of diseases” each chapter defines in the best way its entities. The title of the chapter defines its content.

Examples of the first five chapters with related blocks and titles:

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Blocks</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>A00-B99</td>
<td>Certain infectious and parasitic diseases</td>
</tr>
<tr>
<td>II</td>
<td>C00-D48</td>
<td>Neoplasms</td>
</tr>
<tr>
<td>III</td>
<td>D50-D89</td>
<td>Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism</td>
</tr>
<tr>
<td>IV</td>
<td>E00-E90</td>
<td>Endocrine, nutritional and metabolic diseases</td>
</tr>
<tr>
<td>V</td>
<td>F00-F99</td>
<td>Mental and behavioural disorders</td>
</tr>
</tbody>
</table>

The letter U is currently not used in the classification, but it is used to describe the XXII chapter (U00-U99), which is defined as Codes for special purposes. Indeed, codes from U00 to U49 are used for the provisional assignment of new diseases of uncertain etiology, and codes from U50 to U99 may be used in research, for example when testing an alternative subclassification for a special project.

ICD-10 stresses a difference between inclusion and exclusion terms: the former are used, in addition to the title, as examples for the diagnostic statements to be classified within a specific rubric. They may refer both to different conditions or synonyms. The latter are used
to indicate a different location of a specific term in the classification itself. The code associated to the disease shows where the term has been classified.

For example, the ICD-10 code E10:

**Insulin-dependent diabetes mellitus**

**Includes:**
- diabetes (mellitus):
  - brittle
  - juvenile-onset
  - ketosis-prone
  - type I

**Excludes:**
- diabetes mellitus (in):
  - malnutrition-related (E12)
  - neonatal (P70.2)
  - pregnancy, childbirth and the puerperium (O24)

- glycosuria:
  - NOS (R81)
  - renal (E74.8)

- impaired glucose tolerance (R73.0)
- postsurgical hypoinsulinaemia (E89.1)

ICD-10 is widely used in hospital information systems, mainly for documentation purposes. In Germany for example, it is mandatory for hospitals to code diagnosis in ICD-10-GM-Version for the regulated hospital billing procedure. (See the appendix of „Vereinbarung über die Übermittlung von Daten nach § 21 KHEntgG” paragraph 4 and 5 [3.24]).

**ICD-9**

The 9th version of ICD was used between 1979 and 1998. It classified diseases in a single coded list of three-character categories, but it used only numbers (from 001 to 799), instead of different letters for different categories of diseases.

Diseases were classified in 16 different chapters and the last part was a supplementary classification of external causes of injury and poisoning, which was intended as a specification of the environmental events, circumstances and conditions related to diseases themselves.

**ICD-O**

The International Classification of Diseases for Oncology is an extension of ICD for the topography and the morphology of neoplasms. To date, the used version is the third, ICD-O-3.

The topography section has been adapted from the malignant neoplasm section of Chapter II of ICD-10 in order to gain a greater specificity for the site of nonmalignant neoplasms than in ICD-10. Topography terms have four-character codes (from C00.0 to C80.9).

The morphology code describes the histological cell type and its behaviour, and it is related to specific histological terms. Morphology terms have five-digit codes (from M-8000/0 to M-9989/3). The first four digits indicate the specific histological term. The fifth digit is a behaviour code, which indicates whether a tumour is malignant, benign, in situ, or uncertain whether malignant or benign. A separate one-digit code for histological grading or
differentiation is provided. For a lymphoma or leukemia, this element of the code is used to identify T-, B-, Null-, and NK-cell origin (cf. [3.4]).

3.1.2 ICHI / ICPM and OPS

The International Classification of Health Intervention is a WHO standard which has been used since 1971 to report and to analyze the distribution and evolution of health interventions for statistical purposes. ICHI is structured with various degrees of specificity for application at different levels of the health systems, and uses a common accepted terminology in order to permit comparison of data between countries services.

Its first version was published in 1978 as International Classification of Procedures in Medicine (ICPM) and was limited to surgical procedures. Nevertheless, to date, it covers a wide range of measures taken for curative and preventive purposes by medical, surgical and other health-related care services. Currently, the Family Development Committee of the Network of WHO Collaborating Centres for the Family of International Classification is working on the ICHI adaptation to medical recognized standards and on its update to the newest technological improvements in medical science.

Furthermore, ICPM has been modified by the German Institute of Medical Documentation and Information (DIMI) in the Operationen- und Prozedurenschlüssel (OPS). In the beginning of procedure coding in Germany, only surgical procedures were coded, but since January 2002 OPS versions have been used for general procedure coding. To date, OPS is the official classification of operational procedures in German hospitals.

3.1.3 DICOM- Digital Imaging and Communications in Medicine

The Digital Imaging and Communications in Medicine is developed by the American College of Radiology (ACR) and the National Electrical Manufactures Association (NEMA). It aims at handling, storing, printing and transmitting information in medical imaging processes.

DICOM is not just an image or file format, rather it can be defined as an all-encompassing data transfer, storage, and display protocol, built and designed to cover all functional aspects of digital medical imaging.

The most important features of DICOM are:

1. (It) is applicable to a networked environment using the industry standard networking protocol TCP-IP;
2. (It) is applicable to an off/line media environment using industry standard media such as CD-R and MOD and logical file systems such as ISO 9669 and PC File System (FAT16);
3. (It) specifies how devices claiming conformance to the Standard react to commands and data being exchanged;
4. (It) specifies levels of conformance, i.e. the general requirements which must be met by any implementation claiming conformance;
5. (It) is structured as a multi-part document, in such a way that its evolution consists just in the addition of new features;
6. (It) introduces explicit definition for all the terms in the DICOM Dictionary;
7. (It) specifies a technique for uniquely identifying any data.

Conclusion

DICOM is a universal standard of digital data, since all digital image-acquisition devices used in contemporary medicine produce DICOM images, and communicate through DICOM networks. Moreover, the DICOM data dictionary [3.5] facilitates the representation of clinical information, since it provides clinicians with device-independent terms.
3.1.4 SNOMED CT

Snomed Clinical Terms (SNOMED CT, Systematized Nomenclature of Medicine) is the result of the combination of SNOMED Reference Terminology (SNOMED RT), developed by the College of American Pathologists, with the Clinical Terms Version 3 (CTV3), developed by the National Health Service of the United Kingdom. It is not a medical record system, a database or encoding software, but a source terminology that is being optimized for computer storage in and retrieval from clinical information systems. Its purpose is to serve as a standardized terminology in healthcare software applications, since it enables clinicians, researchers and patients to share comparable data.

The vocabulary of SNOMED CT

Terms in SNOMED CT are classified using concepts, descriptions and relationships between concepts:

1. **Concepts** are clinical meanings identified by a unique numeric identifier (ConceptID) that never changes. Each ConceptID is represented by a unique human readable Fully Specified Name (FSN).

2. **Descriptions** are terms associated to concepts which are used to describe the concepts themselves. Each description is provided with three main characteristics:
   - The **Fully Specified Name** (FSN) provides an unambiguous way to name a concept clarifying its meaning. Each FSN has a “semantic tag” in parentheses, which indicates the semantic categories to which the concept belongs.
   - A **Preferred Term** is a word or a phrase used by clinicians to name a specific concept.
   - **Synonyms** add terms that represent the same concept as the FSN. Synonyms must not be different names, but could also be differences in the spelling of the word, or acronyms.

3. **Relationships** improve the definition of concepts stating relations between them. Besides the formal subsumption relation (“is_a”), there are relationships (called “attribute relationship”), which add more specific information about the concepts, for example “Associated morphology” or “Finding site”.

Top-level hierarchies

The amount of data stored in SNOMED CT is organized into top-level hierarchies and the concepts become increasingly specific. The Clinical finding hierarchy contains the sub-hierarchy of Disease, which subclasses are considered as the results of clinical observations. **SNOMED CT Concept** is the single topmost root concept.

The use of SNOMED-CT with HL7

There are two relevant kinds of conflicts between SNOMED-CT and HL7 (see below) which hamper a seamless integration when HL7 and its Reference Information Model (RIM) are used [3.21]. The first kind of problem occurs when representations of the same domain have different semantic structures. Unfortunately, this has been shown to occur between HL7 and SNOMED-CT: HL7 follows an act-centred view whereas SNOMED-CT aims at modelling clinical constructs. This is a consequence of the two information models different aims and purposes. A second problem occurs when either the same name refers to different concepts in ontologies, or the same concept is named differently in ontologies. Both of these problems occur when HL7 and SNOMED-CT information models are used together. Both these problems are of course related to issues concerning the unambiguous definition of terms in these semantic resources.
Conclusion

SNOMED CT aims to be used in EHR systems to incorporate medical knowledge and clinical data, hence detailed requirements for its integration into particular applications depend on the intended uses of the system.

3.1.5 Health Level Seven (HL7)

Health Level Seven (HL7) is a nonprofit organization which aims at developing protocols for the exchange of healthcare information in clinical settings. Its members are organized in a Working Group, which comprises a “Technical Committee” (TC) and a “Special Interest Group” (SIG). The former is responsible for the content of the Standards, while the latter tries to find out new areas that may need coverage in HL7’s published standards. In 1994 HL7 was accredited by the American National Standards Institute as American National Standards.

HL7 aims to provide standards for interoperability that improve care delivery, optimize workflow, reduce ambiguity and enhance knowledge transfer among all of their stakeholders, including healthcare providers, government agencies, the vendor community, fellow SDOs and patients. In all of their processes they exhibit timeliness, scientific rigor and technical expertise without compromising transparency, accountability, practicality, or their willingness to put the needs of their stakeholders first [3.11].

Three versions of HL7 have been released and the core of the HL7 version 3 is the Reference Information Model (RIM).

The Act-Centered View

The relevant healthcare information in the RIM is organized in six “blackbone” classes: Act, Entity, Role, Participation, Act-Relationship, and Role-Link. According to [3.23], these categories are defined as:

1. **Act Class**: Intentional actions documented by a healthcare professional in either a clinical or administrative context that has happened, can happen, is happening, is intended to happen, or is requested/demanded to happen;
2. **Entity Class**: Physical things or groups of physical things that can participate in an action as perpetrator, target or beneficiary (e.g. living subject, organization, material and places, and their specialization). It does not indicate the roles played, or the acts that these entities participate in;
3. **Role**: the competency of an entity, which can participate in an Act in a particular Role;
4. **Participation**: an association between an Entity in a Role and a specific Act;
5. **ActRelationship**: relates acts such as an order for an observation and the observation event as it occurs. Also relates an act to its component acts;
6. **RoleLink**: a connection between two roles such as patient and provider that expresses a dependency between those roles

SAIF

The HL7 leadership noted that the Version 3 was not achieving significant uptake and initiated a Services-Aware Interoperability Framework (SAIF), which has been seen as providing a framework to achieve working interoperability in the E-Health Domain (cf. [3.16]). The main goal of such an interoperability paradigm was to define standards for “messaging” (method for exchanging data in healthcare environment), “documents” (referring to HL7’s XML-based Clinical Document Architecture standard) and “services” (the meaning of data is preserved while details of the internal service implementation are hidden from the users).
Deployment of HL7

Currently, there are two message protocols supported by HL7, Version 2 and Version 3, the first parts of which were approved in 2004 by the American National Standards Institute [3.1]. HL7 Version 2 Messaging Standard is to date still the most widely implemented standard for healthcare information in the world. However, the compliance to HL7 Version 2 does not directly enable interoperability between healthcare systems because Version 2 messages lack a well-defined stable underlying information model. Thus most definitions for data fields are rather vague or ambiguous and there is an overboarding variety of optional data fields. These features of underspecification and optionality provide very good flexibility. At the same time they require very detailed and sometimes tedious bilateral agreements among the communications partners to achieve any meaningful level of (semantic) interoperability. A big effort has been made to amend this situation by the development of HL7 Version 3 basing it on the RIM that is an object-oriented data model.

There are only very few countries in which HL7 is already implemented in biomedical or healthcare information systems as a complete standard, i.e. messaging standards plus data model standards (RIM) allowing high level of semantic interoperability. Amongst the most important deployments are the following:

- Australia was one of the first countries to adopt the HL7 V2.x standards, and it is now widely used in Australian public and private healthcare organizations.
- In Finland HL7 CDA R2 and V3 messaging is used in the national ePrescribing and patient archival service deployed by the national social insurance institution whereas HL7 V2 is used in the majority of hospitals and healthcare institutions. HL7 CDA is much less common and only plays a role in the communication of patient records on a regional basis.
- In the Netherlands HL7 V3 is being implemented in the design of the national ICT-infrastructure for communication between healthcare providers. This system is called AORTA and will enable the exchange of patient data captured in Electronic Patient Records and its deployment benefits from an almost ten year long experience of Dutch hospitals using HL7 V2 in their internal workflow management.

However due to the lack of semantic content specifications of HL7 V2 and the on-going inconsistencies in HL7 V3 RIM (see [3.41]) these initiatives are still struggling to achieve a high level of semantic interoperability amongst their resources.

Conclusion

The HL7 RIM has gained importance as an information model of healthcare delivery in its own right. Through an analysis of the debate about the RIM effectiveness, it emerges that the main disagreement is not about how the RIM has been designed; rather it is about those principles which have guided the development of the RIM itself.

3.2 Exploitation of Semantic Resources in Hospital Information Systems

Hospital Information Systems are important repositories of medical data. Due to different reasons (like treatment, research, legal obligations and billing) hospitals collect and store many medical relevant data. Semantic resources are a promising tool to get the most benefit out of the data. This section describes the current implementation of semantic resources which is not as prevailed as one could hope for, especially with respect to semantic interoperability. First of all, we describe the current situation in hospital information systems. As an example we will describe the situation for the Saarland University Medical Center, which coordinates p-medicine and plays a key role in utilizing and exploiting the p-medicine
infrastructure. Furthermore, we exemplary analyze which terminologies are used in German HISs, since the E-Health profile of Germany is rather typical for Europe (cf. [3.7]). After that, we review the possible efforts of some specific standards and initiatives, like openEHR, or CEN 13606.

3.2.1 Current Situation in Hospital Information Systems

A Hospital Information System (HIS) is the totality of information systems in a hospital. Hospital Information Systems aim at a communication of different parts of a hospital. Patient management, administration and medical information systems are main tasks which need to be supported by any HIS. They should help to archive and access medical data, e.g. diagnoses and drug usage, enhance administration, for example by supporting the authoring of admission notes, order entries, electronic health records etc. Another crucial part of hospital administration is the billing. There is no software or special approach which is universally used. The provided solutions range from an entire HIS from one provider for all tasks to different software systems for sophisticated tasks which communicate through a communication server.

Interoperability is still a problem for European hospitals, not only for the external communication with other hospitals or medical practices but also for the integration of different sub-systems in one hospital. The study on the use of e-Health in European hospitals showed that “[m]edical Directors found that the main barrier encountered during implementation to be that EHR systems in their departments were not interoperable and/or could not be integrated with new solutions” and that “many had encountered different types of interoperability problems (at technical, semantic and organisational levels)” ([3.7], p. 101).

Though this is the case, little work is done on the question which particular role semantic resources can play in enhancing interoperability. For example [3.6], a study on e-Health, gives no attention to terminologies and ontologies.

The German Situation as an Example – Terminologies used in German HIS

There are little general norms for Hospital Information Systems in Germany. However, the legal background of funding makes the use of some classifications necessary. For German hospitals reimbursement is based on the German Diagnosis Related Groups (G-DRG) since 2004. They are based on the International Statistical Classification of Diseases and Related Health Problems (ICD-10) and the Operationen- und Prozedurenschlüssel (OPS). Since billing is a part of a hospital information system the use of these classifications can be found in any HIS.

Apart from the legally necessary use of DRG the role of the HL7 messaging standard is important. Almost every clinical information system in Germany uses and supports version 2 messaging standards of HL7. Furthermore, this is the most used messaging standard in hospitals worldwide and plays an important role for the integration of subsystems of a HIS (cf. [3.13]). However, HL7 V2.x is not used outside of hospitals. The communication between hospitals and medical practices is outside of this communication standard. The third version of the HL7 was meant to integrate all institutions of medical care but, to date, V3 is, compared to the previous version, rarely implemented.

Many hospitals integrate a Picture Archiving and Communication System (PACS) into their Hospital Information System. Such systems use the standard for Digital Imaging and Communications in Medicine (DICOM) (cf. [3.22], p. 59). DICOM is an open standard which is used by different systems which are related to medical imaging.

HIS at Saarland University, Medical Center Homburg

The Saarland University Medical Centre in Homburg is part of the Saarland University. The communication and information system, including the hospital information system, is run by the Zentrum für Informations- und Kommunikationstechnik (ZIK). Historically, it is founded on
the “Rahmenkonzept für das Klinikinformationssystem der Universitätskliniken des Saarlandes” from 1994. Homburg's HIS is organised as a combination of different parts, mainly SAP Industrial Solutions Healthcare (IS-H) with i.s.h.med modules by Siemens Medical Solutions and some additional software. All parts of the Homburg HIS interact by the central communication server Cloverleaf, a hospital's communication software.

Cloverleaf provides an interface to other hospital applications and is a nodal point of HIS. Other parts are connected to the communication server. Cloverleaf contains the following standards: HL7 (Version 2.1 to Version 3.0), the United Nations Electronic Data Interchange for Administration, Commerce and Transport (EDIFACT), the American National Standards Institute Accredited Standards Committee X12 and XML (cf. [3.17], p. 5). SAP’s Business Application Programming Interfaces (BAPIs) are also supported.

The core of the hospital information system in Homburg is the SAP IS-H (Industrial Solution Healthcare) which is used for the patients' management, admission, transfer and discharge from hospital as well as service recording and billing. It is supplemented by the associated IS-H*med (also: i.s.h.med) module which provides a specific user interface for clinicians. Together they build a complete HIS software which is used in almost 300 hospitals. The relevant data for the DRG-based founding are also captured with IS-H and i.s.h.med. The software is programmed in the language ABAP, a proprietary high-level programming language by SAP but provides interfaces for expansion and integration of further software. The Picture Archiving and Communication System (PACS) is integrated with the Radiology Information System (RIS) and runs with the software GE Centricity which also works with DICOM.

**Conclusion**

Hospital information systems still struggle with interoperability of systems in one hospital and even more of several hospitals and outside healthcare institutions. Nevertheless, little effort is made to use expressive standard terminologies. The messaging standards of HL7 are widely used in clinical information systems and enhance inner-hospital integration but not the exchange with outer systems. Demands for billing induce uniform coding of diagnosis and procedures. However, classifications like ICD-10 are not as powerful semantic resources as ontologies, taxonomies or thesauri. To date, none of such resources is commonly used in HIS software.

### 3.2.2 openEHR

openEHR is a specification for sharable health information systems which provides a foundation to build interoperable and modular software applications. Its intellectual property belongs to the openEHR Foundation, with the founding partners being University College London (CHIME department) and the company Ocean Informatics.

It aims to be implemented in a number of ways:

1. Scalable EHRs: from Personal Health Records to small/medium/large organizations to regional or state clinical record systems;
2. Message-based, web-service-based, middleware application;
3. Integrating existing clinical systems, including virtual federation of data for research or public health purposes.

The openEHR Specification Project is responsible to develop the specifications on which the openEHR Health Computing Platform is based. Such specifications are: the Reference Model (RM), the Service Model (SM), the Archetype Model (AM).

The following features characterize openEHR:

1. Open source initiative available under an open license;
2. A two-level information model, i.e. it separates the technical from the clinical domains. The technicians manage just the technical aspects, while the clinicians cope with the
development of the clinical archetypes and templates. The former are descriptions of
the contents of recorded health information, i.e. models of some clinical content
expressed in ADL (Archetype Definition Language) formalism. The latter are
localized or specialized models that aggregate and choose elements of archetypes to
create a data set;
3. Knowledge-enabled, i.e. it captures the health information, thus the archetypes can
be revisited and versioned to reflect the changes in health domain knowledge;
4. Terminology agnostic, i.e. connects flexibility to any or all terminologies through either
archetypes or templates;
5. Semantic querying: archetypes-based querying enables true longitudinal processing
of health data, regardless of the originating system;
6. Language independent, i.e. archetypes are designed to be used in different countries;
7. Sustainable reference model and life-long EHRs: the openEHR reference model
consists only of generic data types, structures and a small number of generic
patterns, resulting in a small, stable and sustainable information model;
8. Ease of implementation;
9. Ongoing development and enhancement, i.e. it is undergoing continuing
development;
10. Governance of shared content: archetypes are created with broad agreement upon
11. Collaborative model rather than a standards-based path: openEHR is the result of
interested and motivated volunteers from a broad international community of
clinicians and software engineers.

Brief Description
The openEHR information architecture has the following form:

- Four levels of information organisation:
  1. The cognitive user interface;
  2. Templates;
  3. Archetypes;
  4. The reference model.
- Standardised querying capability based on archetype paths and terminology;
- Standardised interface to terminology for inferencing

The archetype model or constraint language is tightly linked to the reference model since
archetypes define constraints on the reference model. Similarly, each data instance in the
EHR is connected to exactly one archetype that specifies the constraints the data has to
adhere to in addition to the rules of the reference model. Thus an EHR system conforming to
this standard needs to offer three building blocks that are required by the archetype
modelling approach: an editor for creating and maintaining archetypes, a validator that
enforces the constraints at runtime, and a browser component that allows for an optimized
display of specific archetypes [3.8].

Conclusion
OpenEHR is currently not implemented in any European hospital information system. Future
prospects for OpenEHR mainly consists in its co-operation with other standardisation efforts
like CEN 13606 and HL7 that we discuss below.

3.2.3 CEN 13606, Electronic Health Record Communication
The Electronic Health Record Communication (CEN 13606) is an European Standard which
aims at providing a rigorous and stable information architecture for communicating part, or all
of the Electronic Health Record (EHR). Such a process of moving pieces of the EHR from
one system to another facilitates the semantic interoperability among different systems.
Indeed, the CEN 13606 does not specify the EHR data in particular settings or domains, rather it is a specification for exchange of models called EHR Extract. Such EHR Extracts can be used to define a message, an XML document or schema, or an object interface.

The CEN 13606 revision project

An initial version of the 13606 four-part pre-standard was published in 1999, and it was only in 2002 that CEN revised the 13606 pre-standard and transformed it into a complete European standard. The cornerstone of this standard is the adoption of the archetype modelling methodology advocated by the openEHR initiative together with the inclusion of parts of the openEHR Reference Model.

The five parts are:

1. Part 1: Reference Model: comprehensive, generic model for communicating part or all of an EHR between heterogeneous systems
2. Part 2: Archetype Specification: constraint-based approach for defining clinical business objects that are built from the Reference Model - adopted from openEHR
3. Part 3: Reference Archetypes and Term Lists: an initial set of inter-reference model conversion archetypes, mapping to openEHR and to the HL7 version 3 RIM Act classes and vocabularies for the Part 1 model
4. Part 4: Security: measures and models to share the access control, consent and auditability of EHR communications
5. Part 5: Interface specification: message and service interfaces to enable EHR and archetype communications

The subject of CEN 13606

CEN 13606 is designed to specify the exchange of EHR Extracts. Its scope does not include standards for a full grown EHR system. Thus it lacks specifications concerning for example version management, workflow management, interfaces to other systems etc. Some of these requirements impact upon the Extract.

For example, because the CEN EN13606 Extract does not yet include version information, it cannot be used between openEHR systems, or any other systems wanting to preserve versioning information.

CEN 13606 is not expected to specify all requirements medical professionals might have for EHR systems. However, it does pertain to those relating to sharing parts and pieces of information contained in EHRs between different systems.

Initiatives like openEHR on the other hand do not only provide specifications for the communication of EHR Extracts of various levels of complexity, but also a full specification for the creation, storage, maintenance, and querying of EHRs. Neither CEN 13606 nor the HL7 RIM defines a reference model for an EHR. The openEHR development process has moved on with the publication of release 1.0, and now includes some features that are very important for implementation but not included in CEN 13606. This is discussed further below.

CEN 13606 and open EHR

13606 has not been widely implemented. The openEHR Foundation has pioneered most of the standard’s content, including developing the constraint formalism, the Archetype Definition Language (ADL). However, some of openEHR’s current formalisms differ significantly from it, particularly with respect to extensions in the reference model, changes to the data types, and extensions to the archetype design method.
Neither 13606 nor openEHR currently provide complete and detailed directions or guidance for designing archetypes. As the basic archetypes design method does not provide internal quality assurance mechanisms or semantic consistency checks, current practice relies primarily on the clinical and logical modelling skills of individual professionals.

The semantic content

Current archetype design practice heavily relies on semantic models known implicitly within individual designers or review groups. The generic structures within 13606 are capable of expressing varying levels of detail within an individual archetype, but the depth and breadth of this semantic detail is completely dictated by the archetype authors. The formalism does not in itself provide strong support for semantic linkages across archetypes.

It is currently not possible to reference clinical evidence or other information at the archetype ‘node’ (element) level.

Summary of strengths and weaknesses

Potential benefits for using CEN 13606 to record and communicate clinical information requirements include [3.18]:

- A reference model based on common concepts related to record structures (sections, entries, etc.) that clinicians generally find intuitively simple to understand and use.
- Open source tools available that have at least partially adopted CEN 13606.
- Tabular output that may be adapted to document detailed mapping findings (to terminologies or other data models) and queries/answers about archetypes design.

Current weaknesses include:

- The 13606 reference model cannot in itself provide a full basis for semantic interoperability. For stronger semantic integrity, it must be supplemented with a model that provides structures for more detailed clinical semantics and the formalism may need to be extended to more strongly support linkage across concepts.

Implementation of CEN 13606

CEN 13606 has been previously published as a prestandard named ENV 13606 in 1999 was not as successfully implemented as expected due to various weaknesses, such as the outdated architectural concepts of applications that support EHRs that were recommended by the standard [3.8]. In 2001, CEN/TC updated ENV13606 and adopted the openEHR “archetype methodology” defined by the openEHR Foundation to make it a full European Standard EHR [3.19]. The result of this work is in CEN prEN13606 which has been adopted by 48 countries up to 2007.

In the UK for instance it is recommended [3.18] that if CEN 13606 is to be adopted by NHS Connecting For Health programme, the possibility of constraining either or both 13606 and HL7 V3 within the NHS should be explored so that translations of semantic contents between them become secure, i.e. semantically standardised.

Although much progress has been made in defining methods and formalisms to describe clinical semantics, no strict formalism or wide-spread process for clinical record “content” modelling has been applied within the NHS National Programme for IT (NPfIT) to date. The same is true for all other European countries.

3.2.4 Integrating the Health Care Enterprise (IHE)

IHE is an international initiative by healthcare professionals and industry to improve the way computer systems share information in the health care domain. IHE is sponsored by the Healthcare Information and Management Systems Society (HIMSS), the Radiological Society of North America (RSNA), and the American College of Cardiology (ACC). The eye care domain is sponsored by the American Academy of Ophthalmology. The IHE approach is to
promote the coordinated use of established standards such as DICOM, HL7 or LOINC to address specific clinical needs in support of optimal patient care. In 1997, a consortium of radiologists and information technology experts formed IHE and started a process through which interoperability of health care IT systems can be improved. This initiative collects case requirements from relevant stakeholders, re-uses available standards, and issues technical guidelines that help manufacturers’ implementation. IHE also stages so-called connectathons and interoperability showcases in which vendors gather to demonstrate and advertise the interoperability of their products.

Designing IHE Profiles form the core activity of the IHE initiative. They provide a common language for purchasers and vendors to discuss their integration needs and offer clear implementation guidelines for communication standards which are accepted by industry partners. IHE pays special attention to careful documentation, reviews and testing.

More than 300 industry partners have pledged support for one or more IHE Profiles [3.25].

IHE Integration Profiles describe a clinical information need or workflow scenario and document how to use established standards (e.g. HL7, DICOM, LOINC...) to accomplish it. A group of systems that implement the same Integration Profile address the need/scenario in a mutually compatible way.

Available profiles cover the following areas [3.26]:

- Anatomic Pathology
- Cardiology
- Eye Care
- IT Infrastructure
- Laboratory
- Patient Care Coordination
- Patient Care Device
- Pharmacy
- Quality, Research and Public Health
- Radiation Oncology
- Radiology

Integration Profiles describe basic functional elements and therefore are reusable in many IT applications. Integration Profiles are developed from relevant use-cases described by real users and are independent of manufacturers or products. They are stable once published and may be extended only by amendments. Integration Profiles are designed as shorthand for healthcare providers to specify integration requirements when acquiring information systems, but up to now recognition seems much greater on the side of the industry.

3.2.5 HL7 CDA, CCD

The Clinical Document Architecture (CDA) was initially developed in 1996 by a group of physicians outside the scope of HL7. The first draft, called the Kona Architecture, was developed in 1997 after the group had joined HL7. CDA introduced the concept of incremental semantic interoperability, i.e. the idea that “there is a range of complexity allowed within the specification and users must set their own level of compliance” [3.12].

The HL7 v3 Clinical Document Architecture specifies the structure and the semantics of clinical documents allowing the exchange between healthcare providers and patients. According to CDA, a clinical document contains the following characteristics: 1) Persistence, 2) Stewardship, 3) Potential for authentication, 4) Context, 5) Wholeness and 6) Human
readability. Examples of CDA documents would be Discharge Summary, Imaging Report, Admission & Physical, Pathology Report and so on. Furthermore, CDA supports the re-use of clinical data for public health reporting, quality, monitoring, patient safety and clinical trials, and it can be reused in multiple applications.

The CDA HL7 is not strictly an EHR standard, but forms its sub-component, that has already been harmonized with the equivalent structure in CEN13606 and openEHR (HL7 EHR 2004). Thus HL7 and CEN cooperate and the current areas of harmonization include data types from (a) HL7 Templates (b) HL7 CDA, (c) CEN13606 Reference Models, and (d) CEN/openEHR archetypes [3.19].

The most recent version of CDA is Release 2, and CDA Release 3 is currently under development.

The Continuity of Care Document (CCD) specification was developed by HL7 together with several members of the American Society for Testing and Material (ASTM), particularly with the technical committee responsible for the development and maintenance of the Continuity of Care Record (CCR) standard. The CCD specification is intended to facilitate the exchange of data between the ASTM's CCR and HL7's CDA specifications, since it specifies the encoding, structure and semantics of a patient summary clinical document for exchange.

The CCD specification enables greater interoperability or healthcare integration of clinical data, and allows physicians to send electronic medical information to other providers without loss of meaning. The CCD specification is limited to the U.S., because it contains U.S. specific requirements.

3.3 Conclusion

In this chapter we analyzed the state of art of semantic resources for clinical information systems. Ideally, such standards assure high degrees of semantic interoperability between different information systems.

Nevertheless, two main problems became clear. First of all, although such standards have usually been developed and further updated and maintained by different communities of experts, some of them are rarely applied. Second, a lot of terminologies' semantics have not been defined for computer systems. Thus, they provide deep specifications of terms in human readable format, but computer systems are not able to reason on such meanings. Such a lack of formalism compromises the possibility to interoperate with different systems. So, unfortunately, a lot of terminologies which are widely used, especially ICD-10, are not suitable as a semantic resource for automated reasoning. What is needed are machine-readable semantics.

References


[3.17] Leistungsbeschreibung Cloverleaf: [http://www.health-comm.de/pdf/produkte_loesungen/cloverleaf/Leistungsbeschreibung_Cloverleaf_5_x.pdf](http://www.health-comm.de/pdf/produkte_loesungen/cloverleaf/Leistungsbeschreibung_Cloverleaf_5_x.pdf)

[3.18] Sato L., Investigating implementing CEN 13606 with HL7 V3 and SNOMED CT – Final Report 2006-12-20 / Final / V1.0


[3.26] [http://www.ihe.net/profiles/index.cfm](http://www.ihe.net/profiles/index.cfm)
4 Semantic Resources in Clinical Research

In this chapter we describe semantic resources that are commonly used in clinical research. Furthermore, we discuss their exploitation in clinical trial management systems and trial repositories. The use of these standardized resources has the role to improve the exchange and reuse of information in clinical research and it is a first step towards interoperability. The same standards can further support a closer link between clinical care and clinical research.

4.1 Semantic Resources

In this section we describe several standardized terminologies/ontologies and coding systems that are relevant in clinical research. While LOINC is widely used both in care and research, other examples such as MedDRA and GO are more specific to clinical research. The semantic resources have been reviewed extensively in deliverable D4.1. In the following we describe them briefly concentrating on aspects relevant for their exploitation in clinical research.

4.1.1 LOINC

The Logical Observation Identifier Names and Codes (LOINC) [4.36] is a universal coding system used in reporting laboratory test results and other clinical observations. This database contains over 30000 different observations. The observations have a formal name, a code and synonyms in the LOINC database.

Assuming the observation as a question and the observation values as answers, LOINC provides codes for the questions. Other coding systems, e.g., International Classification of Diseases (ICD)-9, International Classification of Diseases for Oncology (ICDO)-3, Systemized Nomenclature in Medicine (SNOMED), and the Medical Dictionary for Regulatory Activities (MedDRA) provide codes for the answers. [4.37]

LOINC is compatible with HL7 messages and can derive observations from electronic messages, facilitating the classification of messages coming from multiple sources for putting into electronic health records. Including the LOINC codes to the HL7 messages of the laboratories enables the easy integration of test results in other data repositories.

The mapping of the local test codes to LOINC codes are possible by using a mapping program called Regenstrief LOINC Mapping Assistant (RELMA) [4.35] which comes with the database (s. Figure 1). Many organizations have used LOINC to standardize their multi-source information inputs.

LOINC has been used in Digital Imaging and Communication in Medicine (DICOM) ultrasound messages and Clinical Data Interchange Standards Consortium (CDISC) pharmaceutical industry messages as well as being used as a coding system for the observation identifier field (OBX-3) of the HL7 observation-reporting message. It can further be applied in clinical and research data repositories to identify the clinical and laboratory observations. [4.37]

LOINC can be beneficial for organizations by providing the possibility to store, retrieve and process information coming from different HL7 sources more easily.
4.1.2 Gene Ontology

The GO project [4.1] is a collaborative initiative in the area of bioinformatics whose aim is the standardization of gene and gene product attributes across species and databases. The GO project provides a controlled vocabulary of terms for describing the characteristics or properties of gene products, the Gene Ontology (GO ontology). The project contributes to the specification and maintenance of the GO ontology, the annotation of data sources with terms defined in the GO ontology and, to the provision of tools for managing the ontology and enabling efficient search and reasoning over the data. Current statistics (as of October 2011) set the number of terms in the ontology to approximately 33,300 terms of which roughly 21,300 terms are part of the biological process ontology, 2,800 belong to the cellular component ontology and 9,000 are part of the molecular function ontology.

Although typically referred to as a single ontology, the Gene Ontology can be viewed as both a single ontology or an ontology comprised of three sub-ontologies, each of them covering a separate domain of discourse, namely the domain of biological processes, the domain of cellular components and the domain of molecular functions. In this sense, GO is comprised of three structured, controlled vocabularies for describing gene products characteristics in terms of their relationship with biological processes, cellular components and molecular functions. It is important to stress that GO does not provide a nomenclature for biological objects nor does it provide a catalog or database of genes and gene sequences. Instead, GO aims at describing the properties or characteristics of biological objects.

Figure 1: A snap shot of RELMA mapping (taken from [4.37]).
The ontology is structured as an acyclic, directed graph with terms arranged in a hierarchy-like structure according to generalization and specialization relations (is-a relation) as well as other types of relations (see below). Terms are represented by nodes and relationships among terms are captured by arcs between the nodes. GO contains three top-level terms which represent the roots of the three sub-ontologies. These upper-level terms are not related to each other by any means, i.e. they share no common parent and no relationship exists between them. GO terms are specified by an unique ID and referred to by their ID and a namespace that identifies the domain (sub-ontology) the term belongs to. In addition, terms must contain a textual representation of its definition which can include a set of references to the source(s) of information the definition is taken from. Terms can be related through several types of relations, the most common of which is the is-a relation, however, other types are also possible and these include the part-of, regulates, negatively regulates and positively regulates relations. Additionally, GO terms can be related to terms or concepts defined in external ontologies. Finally, terms can have alternate IDs and synonyms, which are captured through exact, narrow, broad and related relationship types.

The Gene Ontology is available for download in a number of formats including an OBO-based representation of the ontologies (the current version as of October 2011 is v1.2), database files in SQL and MySQL format, three XML-based serializations that include RDF/XML, OBO/XML and OWL and, GO slims, a subset of terms in the ontologies that give a broad overview of its content without given details of its fine-grained terms. The GO database, which integrates the terms of the GO ontologies and the annotation of gene products provided as annotation files by several partners in the GO Consortium, is also accessible through ontology browsers. One such browser is called AmiGO and allows users to post queries against the GO database. See the GO's website for a list of other ontology tools including browsers and editors. Annotation files are also available for download as separate files. The GO database can also be accessed remotely using the GOOSE SQL web-based environment. In addition to these means of access the GO database can be accessed programmatically through Java and Perl APIs.

4.1.3 MedDRA

MedDRA (the Medical Dictionary for Regulatory Activities) is a medical vocabulary/terminology that focuses on the regulatory process of drug development and is used by regulatory bodies and the regulated biopharmaceutical industry for data entry, retrieval, evaluation and display. This terminology classifies adverse event information associated with the use of biopharmaceuticals and of other medical products such as medical devices and vaccines. MedDRA is used in clinical trials for reporting adverse events.

MedDRA was developed by the International Conference on Harmonisation (ICH) and is owned by the International Federation of Pharmaceutical Manufacturers and Associations (IFPMA) acting as trustee for the ICH steering committee. It is an international multi-lingual terminology that enables standardized communication between industry and regulators, provides support of electronic submissions, and addresses all phases of the drug development cycle. It provides a classification for a wide range of clinical information (Diseases, Diagnoses, Signs, Symptoms, Therapeutic indications, Investigation names and qualitative results, Medical and surgical procedures, Medical, social, family history) and support for multiple medical product areas.

The MedDRA hierarchy consists of five levels as depicted in Figure 2. System Organ Class (SOC) is the highest level of the hierarchy. SOCs are identified by anatomical or physiological system, etymology and purpose. The High Level Group Term is subordinate to SOC and is a super-ordinate descriptor for one or more High Level Terms (HLTs).HLTs are

---

3 Technically speaking, the terms in the ontology do not form a hierarchy as they are allowed to have multiple parents.
super-ordinate descriptors for the Preferred Terms (PTs) linked to them. PTs are super-ordinate to Lowest Level Terms (LLTs). There is no limit to the amount of LLTs that can be associated with a single PT. For each PT an identical LLT is created. Each path to a SOC from a PT should have exactly one HLT and HLGT (there is a single path from a PT to a SOC), but a PT can appear in multiple SOCs (called multiaxality).

LLTs are meant to support consistent coding. An LLT is a synonym or a lexical variant of a PT. MedDRA was built from several other terminologies and any terms from included terminologies are represented at the LLT level. The included terminologies are: MCA Medical Terminology, COSTART (5th ed.), WHO-ART (96:4), J-ART (1996), ICD-9, ICD-9-CM (4th revision), HARTS (release 2.2). Each MedDRA term is assigned an 8-digit numeric code.

MedDRA is fully implemented in the WHO global safety database allowing entry and retrieval of information in either MedDRA or WHO-ART. A mapping bridge is kept updated by WHO and ICH, to allow conversion of WHO-ART coded data into MedDRA, allowing users to readily convert their data and use MedDRA.

MedDRA is widely used and brings significant benefits in terms of cost reduction and efficiency. However, inconsistencies are also reported in the literature. In [4.7] several paradoxical features are described and evaluated. An example of a reported issue is that though all of the LLTs linked to a PT are equivalent to one another – and this includes the PT itself – and even though they refer to the same medical concept, LLTs are subordinate to their PT and occupy a distinct level in the MedDRA structural hierarchy. Even the LLT that is identical to the PT is also subordinate to it and the PT is the “parent” of the LLT, and that although the LLT and PT have the same MedDRA code as they are identical. It seems that a PT is subordinate to itself, which is not possible. The author argues that the underlying problem of this paradox is deep, and has consequences for how MedDRA should be represented and used in various contexts as it encourages the view that LLTs (being “subordinate” to PTs) are “less general” or “more specific” than PTs (rather than being truly equivalent to them), and that this distinction must appear explicitly in any representation or use. This view then generates a number of confusions in applying MedDRA e.g. in data mining and knowledge representation, especially when used in conjunction with other similar terminologies. The author also gives several examples when such confusions propagate in publications.
4.1.4 MIAME/Mage

The Functional Genomics Data - FGED Society [4.8] defines several standards related for biological research data quality, annotation and exchange. The “Minimum Information About a Microarray Experiment” (MIAME) [4.9] standard defines the information that is needed to enable the interpretation of the results of the experiment unambiguously and potentially to reproduce the experiment. The six most critical elements contributing towards MIAME are [4.9]:

1. The raw data for each hybridisation (e.g., CEL or GPR files)
2. The final processed (normalised) data for the set of hybridisations in the experiment (study) (e.g., the gene expression data matrix used to draw the conclusions from the study)
3. The essential sample annotation including experimental factors and their values (e.g., compound and dose in a dose response experiment)
4. The experimental design including sample data relationships (e.g., which raw data file relates to which sample, which hybridisations are technical, which are biological replicates)
5. Sufficient annotation of the array (e.g., gene identifiers, genomic coordinates, probe oligonucleotide sequences or reference commercial array catalog number)
6. The essential laboratory and data processing protocols (e.g., what normalisation method has been used to obtain the final processed data)

MAGE-TAB [4.10] (the current recommended best practice) defines a (tab-delimited) format conforming to the MIAME standard. A recommended practice is that the MGED Ontology [4.11] is used for description of the key experimental concepts, and where possible ontologies developed by the respective community for describing terms such as anatomy, disease, chemical compounds etc.

The primary purpose of the MGED Ontology is to provide standard terms for the annotation of microarray experiments (see [4.11]). These terms enable structure queries of elements of the experiments. Furthermore, the terms enable unambiguous descriptions of how the experiment was performed. The terms are provided in the form of an ontology which means that the terms are organized into classes with properties and are defined. A standard ontology format is used. For descriptions of biological material (biomaterial) and certain treatments used in the experiment, terms may come from external resources that are specified in the ontology.

4.2 Exploitation of Semantic Resources in Clinical Research

It is obvious that the semantic resources described in Section 4.1 evolved from a clear need in the scientific community to share information, to reduce ambiguity and to increase reuse. While some of those resources have reached wide use in industry and have become part of commercial products, others are still limited in use. While they are not sufficient to ensure interoperability, they are a first step towards being able to share data with meaning across systems.

While there is a clear trend towards improved interoperability, existing solutions offer only limited support. Main hurdles are represented by heterogeneity of solutions, lack of coherence in adoption of common standards, the large number of initiatives that try to solve the same issues in isolation, and the many legacy systems.

In this section we describe several research and commercial CTMS and research infrastructures and refer to the use of standards in each of them. We also present several prominent standardization initiatives.
4.2.1 Current Clinical Trial Management Systems

4.2.1.1 Oracle Clinical

Oracle Clinical [4.12] is a clinical trial management system with remote data capture functionality. In Oracle Clinical, protocols can be designed (including study objectives, investigator and site information, enrolment plans, drug treatment regimens, randomization schedules, and visit definitions), and it provides a complete infrastructure for electronic data capture and clinical data management.

Internally, data is managed in an Oracle database with a design following the design of a CRF (Case Report Form) book for a clinical trial.

The following tables are defined:

- DVG – Discrete Value Group. The DVG table is used to specify the set of possible answers when a question has a limited set of answers (e.g. coded answers).
- Question. The Question table is used to specify the CRF questions. Each row represents a specific question on a CRF form.
- QG – Question group. A question group is used to group (medically) related questions. This makes it possible to reuse complete question groups between CRFs.
- DCM – Data Collection Module. A DCM represents all QG’s that should be answered during a single visit. Typically, these are represented as (sections of) CRF screens.
- DCI – Data Collection Instrument.

In addition to collecting data which represents the Case Report Forms, Oracle Clinical also provides the definition of so-called derivation procedures. These procedures create additional (derived) variables in the database to support analysis.

Oracle also supports the “Oracle Thesaurus Management System” [4.13] product, a product to manage and classify free text captured during clinical trials, out-of-the-box supporting a wide variety of commonly used dictionaries like MedDRA, MedDRA-K, MedDRA SMQs, SNOMED, ICD9, WHO-ART and WHO-drug. The Oracle Thesaurus Management system can be used to classify “raw” (textual) patient data in Oracle Clinical [4.14]. The product tries to auto-classify the data (that is marked for processing by the TMS), if it fails to classify, it will flag the entry for manual handling. The results of the classification will be stored in derived variables. This is often used in for instance the coding (into MedDRA terms) of adverse events in the Adverse Events Case Report Form of a clinical trial.

4.2.1.2 OpenClinica

OpenClinica [4.34] is a web-based software platform developed by Akaza Research for managing diverse and multi-site clinical research studies through a unified interface. It facilitates management and collection of clinical trial data, and serves as a centralized clinical data repository that manages data securely while providing distributed web-based access for collaborative research.

OpenClinica generates web-based case report forms (CRFs) dynamically for clinical assessment instruments from user-defined clinical parameters, protocols and validation logic. It provides built-in standard-based CRF templates based on which new CRFs can be developed.
A study can be designed by submitting different workflows according to different research policies, adding clinical assessment instruments and CRFs, query and export datasets and assigning privileges to people according to their roles in the study and administrating their accounts. The intuitive workflows and user interfaces are all integrated within a web-based environment that results in high accessibility as well as high usability due to the fact that no technical background is required to program a study. The system also supports sharing resources across projects in a secure and transparent manner.

OpenClinica has a robust security model. Moreover, it enables comprehensive auditing and provides an advanced data integrity. It uses standardized formats for exchanging data, therefore, interoperability and cooperation is made possible.

OpenClinica supports HIPAA, 21 CFR Part 11, and other regulatory guidelines, and is designed as a standards-based extensible, modular, and open source platform. Compliance with these guidelines brings along capabilities such as definition of hierarchical user roles and privileges within projects, SSL web access, de-identification and encryption of Protected Health Information (PHI), and auditing of data updates/access.

---

**Figure 3: Data flow in OpenClinica (figure taken from [4.15])**

**Figure 4: System end of OpenClinica (picture taken from [4.15])**
4.2.2 Clinical Trial Repositories

4.2.2.1 I2b2

I2b2 [4.16] (Informatics for Integrating Biology and the Bedside) is an NIH-funded National Center for Biomedical Computing based at Partners HealthCare System. I2b2 provides a platform allowing clinical researchers to import clinical (research) data for discovery research.

I2b2 uses a service oriented approach, its core services (named “cells”) are project management, file repository, identity management, workflow framework, ontology management and data repository. Of these services, the ontology management [4.17] and data repository [4.18] are relevant with respect to interoperability.

The data repository uses a data warehouse approach based on the star scheme structure. The star has a central data table (called a “fact” table). In the data repository this is the table representing observations on a patient (containing the basic attributes about the observation, like the patient number, provider number, a code for the concept observed, a start date and an end date). The fact table is related to a number of other tables (called “dimension tables”) which can further describe attributes of the fact table. There are five dimension tables defined: the patient, concept, visit, and provider dimension.

Each record in the patient_dimension table represents a patient in the database. The table includes demographics fields such as gender, age, race, etc. Most attributes of the patient dimension table are discrete (i.e. Male/Female, Zip code, etc.).

The visit_dimension table represents sessions where observations were made. Each row represents one session (also called a visit, event or encounter.) This session can involve a patient directly, such as a visit to a doctor’s office, or it can involve the patient indirectly, as when several tests are run on a tube of the patient’s blood.

The concept_dimension table contains one row for each concept. Possible concept types are diagnoses, procedures, medications and lab tests. It can store virtually any concept type, such as demographics and genetics.

Each record in the provider_dimension table represents a physician or provider at an institution. The provider_path is the path that describes how the provider fits into the institutional hierarchy. Institution, department, provider name and a code may be included in the path.

The ontology management cell manages all ontology/terminology related aspects. Conceptually, the i2b2 ontology cell organizes the i2b2 ontology conceptually as a tree (with concepts and (subsumption) relations, see Figure 5).

Existing ontologies can be imported into the i2b2 structure (e.g. as a sub tree) as long as they are mapped onto a tree structure.
4.2.3 CDISC

The Clinical Data Interchange Standards Consortium (CDISC) [4.19] focuses on clinical research and is an organisation that aims to develop and support global, platform-independent data standards that enable information system interoperability to improve medical research and related areas of healthcare. The FDA encourages the usage of CDISC standards for (clinical trial) result reporting. In order to facilitate an efficient submission of trial results to regulatory bodies, CDISC has defined the Study Data Tabulation Model (SDTM). The SDTM is a general framework describing the organization of the information that is collected during clinical trials. The SDTM consists of a set of clinical data file specifications and underlying guidelines. These different file structures are referred to as domains. Each domain describes a type of data associated with clinical trials, such as demographics, vital signs or adverse events. CDISC also provides a standard for running a clinical trial, namely the Operational Data Modelling (ODM) standard. ODM supports interchange between applications used in collecting, managing, analysing and archiving data. ODM provides a format for representing study metadata, study data and administrative data associated with a clinical trial. Various clinical trial management systems like SAS [4.20], Oracle clinical [4.12] and ObTiMA [4.23] support the ODM standard for data exchange.

4.2.4 BRIDG

The BRIDG project [4.22] is a collaborative effort of several institutes and organizations that includes the US National Cancer Institute (NCI), the Cancer Biomedical Informatics Grid of the NCI (caBIG), the US Food and Drug Administration (FDA), the Clinical Data Interchange Standards Consortium (CDISC) and the HL7 Regulated Clinical Research Information Management Technical Committee (RCRIM TC). The project aims at developing a shared view of the data, relationship, and processes found within the domain of protocol-driven research and its associated regulatory artifacts. The main focus is on capturing the semantics of both static and dynamic data structures and relationships and, business processes related to the project's domain of interest in order to support semantic interoperability both among humans and computer systems within the domain of interest and
between this domain and others, e.g. the Public Healthcare domain or the Life Sciences domain. The domain of interest of the BRIDG project is defined as [4.22]:

“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 biologic subject or substance plus all associated regulatory artifacts required for or derived from this effort.”

The main contribution of the project is the definition of an implementation-independent reference model for capturing the semantics of the project's domain of interest. Such a model, referred to as the BRIDG model, is a type of Domain Analysis Model that captures a shared view of the static and dynamic semantics of the underlying domain and is realized through a series of graphical “views” expressed in UML, the Unified Modelling Language. The artifacts modelled by the BRIDG model, i.e. the domain's business processes, the data structures and relationships between domain entities stem from projects within the BRIDG project's stakeholder organizations (for a list of the projects that contribute content to the BRIDG project and model the interested reader is referred to the BRIDG project's documentation [release 3.0.3 documentation]) [4.22]. The BRIDG model provides the means for capturing both a domain's static and dynamic semantics and employs a harmonization process to “bridge” the semantic gap between a (sub) domain's model, i.e. a project, and the BRIDG reference model. This process enables the mapping between a project's static content and the BRIDG model's static component. In addition, the process enables the representation of a project's dynamic content and the set of entities that cannot be mapped to the reference model. For details of the harmonization process the interested reader is referred to the model's documentation (version 3.0.3) found at its website [4.22].

To achieve its goal the current version of the BRIDG model (version 3.0.3) uses a multi-perspective approach to modelling the static semantics of protocol-driven research. Each perspective captures the semantics of the underlying domain from a different point of view, e.g. from a domain expert's point of view, and in that way becomes accessible to a diverse audience. The current version of the BRIDG model is structured in the following four perspectives: the Canonical Perspective, the Ontological Perspective, the HL7 Perspective and the Subject Matter Expert (SME) Perspective.

- **Canonical Perspective**: This perspective defines how the semantics in the SME perspective are integrated and gives a comprehensive view of the semantics of BRIDG.
- **Ontological Perspective**: This perspective provides an OWL DL-based representation of the semantics of BRIDG and is intended as an approach to validate the semantics of BRIDG against that of other ontologies such as for example HL7's RIM.
- **HL7 Perspective**: This perspective represents the semantics of BRIDG in HL7's RIM, the Reference Information Model proposed by HL7.
- **Subject Matter Expert (SME) Perspective**: This perspective is tailored to domain experts. From this point of view domain models appear as the domain experts expect them to be described, i.e. using the domain expert's language. The SME perspective can contain multiple sub domain models, each of which captures a specific area within a domain. Currently it contains five sub domains:
  - **Protocol Representation**: this refers to the domain of research protocols, including their design and planning stages. A model of this (sub) domain will capture the relevant concepts, attributes and relationships among the concepts related to research protocols such as arms and epochs.
**Study Conduct**: this sub domain involves the execution of a research study. A model of this (sub) domain will cover the activities carried out while conducting a study and the results of the study.

**Adverse Event**: this is the domain of safety-related activities within research protocols and/or after the execution of a protocol.

**Regulatory**: this (sub) domain covers the activities pertaining to the creation and submission of documents to regulatory organisms.

**Common**: this domain captures the semantics that is common to all the previous domains.

BRIDG uses both automatic and manual transformation methods for transforming the information between perspectives. In particular, the transformations between the canonical and ontological perspectives and between the canonical and HL7 perspective are done manually while the transformation between the canonical and the SME perspectives is automated.

The NCI uses the BRIDG model as an enabling technology to facilitate the interoperability among the NCI's caBIG program's Cancer Clinical Trial Suite of tools and applications. In this sense it is being used as the model for the "representation of the shared static semantics of the Clinical Trial Management System (CTMS) Workspace". BRIDG will also be used for specifying the underlying domain model of the CDISC's SHARE (Shared Health and Research Electronic) project, a project that aims at building a global, accessible electronic library for use within the biomedical research and healthcare domains.

### 4.2.5 ObTiMA

ObTiMA [4.23] is an ontology-based trial management system intended to help design and conduct clinical trials in an end-user friendly way. It comprises all features of a full-fledged trial management system and achieves data integration by utilizing an ontology.

ObTiMA has been developed within the integrated European project ACGT (Advancing Clinico Genomic Trials on Cancer) [4.24], which provided a European biomedical grid infrastructure for cancer research. One of the main objectives of the project was the ontology-based integration of heterogeneous biomedical databases using a semantic mediator. Therefore, a mediator was developed, that is able to query different heterogeneous data sources in terms of a shared ontology for cancer trials, the ACGT Master Ontology (ACGT-MO) [4.25]. A new data source is integrated into the semantic mediator by creating a set of rules for mapping this data source onto the ACGT-MO. For databases, this task is simplified by a graphical tool that assists in mapping from database tables and columns onto appropriate ontological classes and relations. Still this process remains a complex task that needs to be performed by users who are experts both in the domain, the database, and the ontology. They must be able to realize the subtle differences between similar ontological classes and how this is mirrored in the data sources (and vice-versa).

The process is necessary for legacy data sources but it is desirable that databases of newly developed data management systems are set up during creation in an ontology compliant way to allow a seamless integration of the data collected in these systems into the ACGT mediator architecture. To explore this approach ObTiMA has been developed, a trial management system that integrates the ACGT-MO already at the beginning, in the design process of a clinical trial, in order to guarantee that the data collected during the trial has comprehensive meta data in terms of the ontology without the need to perform a separate mapping process. In Figure 6 the main components of ObTiMA, which are the Trial Builder and the Patient Data Management System, and their interaction with the semantic mediator are shown.
The Trial Builder allows the trial chairman to define the master protocol, the Case Report Forms (CRFs) as well as the treatment plan for the trial, in a way that is both semantically compliant with the ACGT-MO and user-friendly. From these definitions, the Patient Data Management System can be set up automatically in such a way that a clinician can collect the patient data during the trial according to the defined treatment plan. The data collected in the trial is stored in trial databases whose comprehensive meta data has been rendered in terms of the ACGT-MO. Furthermore, from these definitions the mapping file for the mediator can be created automatically and the data collected during the trial can thus be seamlessly integrated into the mediator architecture.

**Figure 6: ObTiMA System Components and Integration into the ACGT Infrastructure (taken from [4.23]).**

Beside the enhanced data integration, an advantage of ObTiMA is that by using a shared ontology to create a data model, the collected data becomes consistent to the knowledge of the underlying domain, coded in the ontology, and thus data quality increases. Nevertheless, as shown above, ObTiMA has been designed to make the details of the Master Ontology and Semantic Mediator transparent for the user to concentrate on the clinical trial workflow and to make the system as user-friendly as possible. Also, the assembled ontology descriptions can be used to determine attributes necessary to setup the database, as e.g. the data types for items and thus enables the user to set up the trial database in an user friendly and semantically compliant way.

### 4.2.6 Transcend

**TRANSCEND (TRANslational Informatics System to Coordinate Emerging Biomarkers, Novel Agents, and Clinical Data)** is a comprehensive, scalable bioinformatics infrastructure developed to support the rapid data collection and analysis required by the I-SPY 2 TRIAL. The I-SPY 2 TRIAL [4.26] *(Investigation of Serial Studies to Predict Your Therapeutic Response with Imaging And MoLecular Analysis 2)* is a multi-center study (currently, 20 centers are involved) that combines biomarkers to predict response to therapy in patients with locally advanced disease undergoing neo-adjuvant treatment. I-SPY 2 uses an adaptive design where treatment response is used to randomize future subjects, thus dramatically accelerating drug evaluation in breast cancer.

TRANSCEND leverages caBIG® tools and 3rd party open source products, such as an electronic health record system created by Tolven [4.27], which are connected via the caBIG® Integration Hub [4.28] to achieve uniform data collection across the 20 sites participating in I-SPY 2 trial and to ensure interoperability and compliance with standard biomedical terminologies such as SNOMED CT [4.29] and caDSR. Only large sites were selected to participate to be able to ensure the availability of the required technical infrastructure, the expertise, and the enrolment capability.
The Tolven system, initially developed as a PHR, is customized to create CRFs for the study. The system is used as a CDMS and was chosen because it is easy to move data around, which is not the case with standard CDMS. The sites attach the source documentation to the eCRF so that at the centre the filled in data can be verified.

TRANSCEND uses the following open source caBIG tools:

- **caTissue**: caTissue Suite is caBIG’s biorepository tool for biospecimen inventory management, tracking, and annotation [4.30].
  
  In TRANSCEND, it is used to:
  
  - Track all biospecimens collected from multiple trial sites;
  
  - Create and track derivative samples; Tolven eCHR creates specimen IDs that are electronically transmitted to caTissue via a service unit in Integration Hub (caXchange). caTissue has an automated interface via Integration Hub.

- **caBIG Integration Hub (formerly caXchange)**: enables tools in the caBIG® Clinical Trials Suite to interface seamlessly with one another [4.28].

- **caIntegrator**: caIntegrator is a web-based software package that allows researchers to set up custom, caBIG®-compatible web portals to conduct integrative research, without requiring programming experience. These portals bring together heterogeneous clinical, microarray and medical imaging data to enrich multidisciplinary research. caIntegrator leverages the Cancer Data Standards Registry and Repository (caDSR) to map experimental data to well-defined datatypes and utilizes caGrid and Java client APIs to access data from caBIG® applications such as caArray, the National Biomedical Imaging Archive (NBIA). caIntegrator is also integrated with caBIO to perform queries on genes and pathways [4.31]. In I-SPY2, caINTEGRATOR is used for storage and retrieval of trial data for each participant or in aggregate.

- **caARRAY**: is an open-source, web and programmatically accessible array data management system [4.32].

  In I-SPY 2, caARRAY is used to store gene expression profiling (GEP) raw data in MAGE-ML format. One of the enhancements the group wants to achieve is the connection of caINTEGRATOR with caARRAY in order to allow the access to both the raw data and the results obtained from their analysis/other patients’ data. caARRAY has a manual interface.

- **caAers**: The Cancer Adverse Event Reporting System is an open source software tool that is used to collect, process, and report adverse events that occur during clinical trials. [https://cabig.nci.nih.gov/tools/caAERS](https://cabig.nci.nih.gov/tools/caAERS). Currently, Adverse Events are collected and communicated to Data Monitors (CRO) in Tolven eCHR. Data Monitors manually upload Serious Adverse Events into caAERS for reporting to Regulatory Agencies.

### 4.3 Conclusion

In this chapter we have described several semantic resources relevant in clinical research, prominent standardization initiatives, and clinical research-specific systems and
infrastructures. While our survey was not meant to be exhaustive, it is sufficient to enable us to identify several issues that hamper the main goal of using semantic resources and standards, which is to enable efficient machine-processable and machine-understandable sharing of information across systems and organizations, and by a large community of users. Many of the systems developed are limited in scope and solve a specific problem for a specific community of users whose requirements were at the basis of the development. There is therefore a strong need towards unifying the existing initiatives in order to increase the achievable impact.

There are promising efforts from large standardization bodies such as HL7 and CDISC towards a unifying approach, but achieving semantic interoperability in healthcare research across all required boundaries and being able to deal with all sources of heterogeneity are still challenging issues.

References

[4.16] https://www.i2b2.org/
Holger Stenzhorn, Gabriele Weiler, Mathias Brochhausen, Fatima Schera, Vangelis Kritsotakis, Manolis Tsiknakis, Stephan Kiefer and Norbert Graf: The ObTiMA System - Ontology-based Managing of Clinical Trials. in Proceedings of the 13rd World Congress on Health (Medical) Informatics (Medinfo 2010), Cape Town, South Africa, September 2010


Different countries, regions, even hospitals use different systems for storing patients’ data. Also in many countries, clinical and administrative data are captured in different ways. An EU task force for Information Society and Media Directorate in health and social care carried out a survey of over 1000 hospitals in 27 EU member countries and Croatia, Iceland and Norway [5.1] and provided an overview of how Europe’s acute hospitals use eHealth. According to the study, sharing electronic clinical data between hospital organisations, consulting physicians in the community and other community health care providers is essential in order to track the level of take-up by hospitals and the implementation of eHealth aims at European level. Improved levels of medical exchanges across countries also remain relevant in the view of the single market and the mobility of patients and workers across countries for personal or professional purposes.

However, the external exchange of patient-level information takes place in only three out of ten hospitals. This is especially the case in public hospitals: over one-third of them exchange information in contrast to fewer than one-quarter of private hospitals. University hospitals are also more likely to exchange information than non-university hospitals. Furthermore, the larger the hospital is, the more likely it is to exchange information. In around half of all hospitals there is, however, simply no active data exchange with other providers.

This external data exchange remains very low within and between hospitals, and especially between countries. It is possible that this low level of usage can be explained by such continuing barriers as insufficient interoperability, and inhibiting structural, organisational, and other behavioural or motivational mechanisms.

In the following sections the reuse and exchange of Electronic Health Record (EHR) data and Clinical Trial Management System (CTMS) data are discussed. The chapter concludes with efforts related to re-use of heterogeneous EHR and CTMS data like DARTNet, caBIG Clinical Trial Suite tools (clinical connector and Integration Hub), EHR4CR, Information Warehouse (IW), and STRIDE.

### 5.1 Reusing EHR data

EHR systems aim at achieving the best possible support towards the health care of patients. The clinical decision making, and drug and treatment prescribing are the first hand uses of EHR data. But there are some alternate uses of EHR data also. One of the potential benefits of standardised computerisation of health data is the opportunity of combining, aggregating and analysing electronic records for clinical research.

While the benefits of EHRs in direct patient care are widely recognised, the benefits from secondary use of data that has been de-identified and aggregated for medical research purposes is grossly underestimated, or even overlooked. In many hospitals the integration of data within data warehouses is still limited to administrative data sources and applied within hospital management [5.10].

Currently health care requires researchers to carry out research studies in isolated fashion at different locations. The use of the EHRs varies widely across different nations worldwide. According to the Electronic Health Records/Clinical Research (EHR/CR) Working Group [5.2] moving from the current state of electronic health records and data capture for clinical research, to the ultimate desired state of clinical research as a natural progression of healthcare, will be a long process with many stages.

Data standards are essential for data collection, interpretation and exchange within the medical and research communities. Collaboration on common data standards and data transfer standards will be critical to support both the implementation of electronic national health records for national health information networks and clinical research. To support interoperability and provide better access, several metadata standardisation projects have
been initiated. The Clinical Data Interchange Standards Consortium (CDISC) standard aims to develop a XML-based metadata model to support standard data interchange between medical and biopharmaceutical. Also, the Health-Level 7 (HL7) standard represents an effort to define an EHR standard for the healthcare industry. Finally, the Integrating the Healthcare Enterprise (IHE) initiative that creates the framework for passing vital health information seamlessly – from application to application, system to system, and setting to setting – across multiple healthcare enterprises by driving the adoption of existing standards, such as HL7, to address specific clinical needs.

5.2 Reusing CTMS Data

Every year, vast amounts of resources are invested into clinical trials to discover new mechanisms of disease, evaluate the effectiveness of new therapies, and assess the safety of therapeutic interventions [5.3]. As scientific discovery becomes more complex, it becomes increasingly important to systematically analyse—and integrate across different trials—data collected as part of clinical trials. Unfortunately, the valuable data gained from these studies are often collected in disparate databases that limit the ability to exchange, share, and systematically analyse and integrate clinical trials data.

The ECRIN (European Clinical Research Infrastructure Network) data management working group conducted a two-part standardized survey for the heterogeneity of clinical trial data management in Europe [5.11]. The questionnaires were answered by nearly 80 centres/units (with an overall response rate of 47% and 43%) from 12 European countries and EORTC (European organisation for research and treatment of cancer). According to the survey the solutions used for clinical data management are very heterogeneous: 20 different commercial CDMS products (7 Open Source solutions) in addition to 17 proprietary systems are in use.

In a heterogeneous environment the use of data exchange standards (like CDISC, HL7, IHE) becomes a necessity for cooperation in international clinical trials. Using standards we can simplify data exchange, increase the quality of data and prepare centres for new developments such as the use of EHR for clinical research.

One of the most promising models for data exchange nowadays is the Biomedical Research Integrated Domain Group (BRIDG) project (s. Section 4.2.4) to develop a model of the shared understanding of the semantics of clinical research.

One of the early projects supported by the caBIG™ project was the development of a structured protocol representation that could be used to exchange clinical trial protocol information among caBIG™ participants. Rather than create “yet another standard”, the caBIG™ structured protocol representation project joined forces with CDISC and HL7 to understand the underlying semantics of the data collected as part of clinical research and to develop a shared model that could be used by all the stakeholders. Using the initial model developed by CDISC as part of the harmonization efforts between CDISC and the HL7 RCRIM TC as the basis for the initial modelling effort, the BRIDG project was initiated.

5.3 DARTNet

The Distributed Ambulatory Research in Therapeutics Network [5.12] is a federated network of electronic health record data, designed as a platform for next-generation comparative effectiveness research in real-world settings. DARTNet links information from non-integrated primary care clinics that use EHRs to deliver ambulatory care to overcome limitations with traditional observational research.

The aims of the initial DARTNet project were to:

- develop a federated network of 200+ primary care clinicians, all using electronic health records (EHR);
analytically demonstrate how existing large-scale data sets can be enhanced by patient-level EHR data to inform and expand knowledge of effective and safe medical therapeutics;

demonstrate the ability to collect specific data from clinicians or their staff on a clinically defined set of patients to enrich the EHR data set and answer effectiveness and safety questions concerning medical therapeutics.

DARTNet’s architecture is based on Grid computing. Distributed networks allow a central site to simultaneously query data stored locally for each member of the Grid. The primary interface for accessing the Grid system is called the electronic Primary Care Research Network (ePCRN) Portal [5.15]. The ePCRN portal provides the architecture that accesses the Gateway databases for all users, as well as the query capabilities and the security systems. DARTNet uses the connectivity and distributed query capabilities of the ePCRN Portal, which are handled through a specifically designed application over the grid middleware.

Federated networks link separate databases in such a way that a single query can run on the separate databases and return results while conforming to each organization’s privacy and confidentiality standards. This structure facilitates sharing and interchanges of data among autonomous databases, such as EHRs located within different organizations. The single interface provided by a federated database system allows a user to retrieve data from multiple geographically decentralized and heterogeneous databases with a single query. The federated architecture provides mechanisms for sharing data and transactions, for combining information from several components of the system, and for coordinating activities among autonomous components.

DARTNet has elected to work only with EHRs that include coded problem lists, electronic prescribing, and laboratory interfaces. The EHR system must also allow read-only access to a data extraction/standardization system. Virtually all EHRs that meet these minimum requirements can be supported. EHRs that are known to be compatible with the current data extraction system include Allscripts Professional®, Allscripts Enterprise®, eMDs Chart®, GE Centricity®, Meditech®, Misys EMR®, NextGen®, Practice Partner®, Medent®, NextGen®, SmartClinic®, and SOAPware®.

The core of DARTNet is a network of medical practices that use electronic health records (EHR). Data from each practice’s EHR are standardized and stored in a local database. This local database, called the Clinical Data Repository (CDR), connects to other databases, such as practice management databases, hospital databases, and pharmacy fill/refill databases, thus centralizing and standardizing data across disparate systems. The CDR prepares data for distributed queries. The CDR is primarily populated with data elements used for clinical decision support. All data elements in the CDR are standardized (cross-walked) to one of several coding systems; ICD-9 CM for diagnoses, RxNORM and GCN codes for drugs and SNOMED CT codes for all other data elements.

Data elements identified as applicable to the clinical decision support process or for current or future studies are mapped within each EHR and standardized upon being imported into the onsite CDR. The manual labor-intensive process of mapping includes identifying the variations on data content and storage locations occurring within each EHR and deciding how to apply standardized nomenclature to each data variant. The EHR mapping process which uses the CINA Mapper® (http://www.cina-us.com/techmapper.html) utilizes pattern matching to locate, verify and translate both codes and text into the standardized nomenclature maintained within the CDR.

The movement of data from the CDR to the Gateway database is based on the Continuity of Care Record, which is a core dataset of the most relevant administrative, demographic, and clinical information about a person’s health care. The CINA tools access data in the EHR, standardize the data, and then create Continuity of Care Records locally for each eligible
patient at each of the clinical organizations. Continuity of Care Record files consist of an XML string, which is passed to the ePCRN Gateway database (created in MySQL), and the file is parsed into fields that are selectively available to outside Grid enabled queries, effectively de-identifying the dataset. Figure 7 below summarizes the relationships between data sources and data access points.

![Figure 7: Relationship of data sources and the DARTNet (taken from [5.15]).](image)

5.4 caBIG Clinical Trial Suite

The cancer Biomedical Informatics Grid [5.6] initiative was launched by the National Cancer Institute, aiming to create a virtual network of interconnected data, individuals and organizations that collaborate in order to redefine the way that cancer research is conducted. Several tools have been developed under this initiative that assist in collecting, analysing, integrating and disseminating data information that is related with cancer care and research. Objective of these tools is to promote data sharing in a syntactically interoperable manner.

The caBIG® Clinical Trials Suite [5.8] is an enterprise clinical trials system that has been developed (and continues to be enhanced) primarily for use in trial sites. The Suite is comprised of a collection of interoperable modules covering a broad range of key areas in cancer clinical trials management. These include patient registration via caBIG® Central Clinical Participant Registry (C3PR), patient scheduling via caBIG® Patient Study Calendar (PSC), adverse events reporting via caBIG® Adverse Event Reporting System (caAERS), lab analysis via caBIG® Lab Viewer, and clinical data management via the caBIG® Clinical Connector. Integration of these applications is centered around five key scenarios: Study Creation, Registration of Subject, Loading Labs in CDMS, Lab-driven Adverse Event Creation, and Adverse Event-Triggered Schedule Change.

![Figure 8: caBIC clinical Trials Suite (taken from [5.8]).](image)
account the diversity of clinical research activities and local practices that exist among trial sites.

The implementation is based upon the caGrid infrastructure with caBIG® Integration Hub as the Enterprise Service Bus for reliable message routing and GAARDS providing robust security.

5.4.1 caBIG Clinical Connector

The caBIG® Clinical Connector is a component of the caBIG® Clinical Trials Suite that provides a conduit (a semantically integrated service layer) from the caBIG® Clinical Trials Suite to multiple Clinical Data Management Systems (CDMS). The connector uses a data model based on the Biomedical Research Integrated Domain Group (BRIDG) model, and defines the service operations that can be implemented by CDMS.

The Clinical Connector is made up of multiple connector services. The first connector service provides a mechanism for external applications to enrol patients. The second connector service provides a conduit for external applications to load laboratory test result data. The third connector service facilitates data extraction.

Features (Services)

- Allows a patient registered in the caBIG® Central Clinical Participant Registry (C3PR) to be enrolled on the corresponding study in the CDMS
- Allows the caBIG® Lab Viewer tool to transfer laboratory test results into the CDMS and populate electronic Case Report Forms (eCRFs)
- Allows study design metadata to be extracted from the CDMS along with context-related information

![Figure 9: Architecture of caBIG Clinical connector (taken from [5.9]).](image)

The Clinical Connector for C3D has an Administrator’s User Interface. Administrators can use this application to manage the Clinical Connector settings and Login credentials to be used by the patient registration process.

5.4.2 caBIG® Integration Hub

The caBIG® Integration Hub [5.7] enables tools in the caBIG® Clinical Trials Suite to interface seamlessly with one another. This software supports integrated clinical trials workflows and offers users the flexible interoperability associated with a service oriented architecture. The caBIG® Integration Hub is designed to create integrated business workflows using existing applications and databases and facilitates the ability of organizations to integrate caBIG® Clinical Trial applications with their existing applications.
caBIG Integration Hub, which was originally called Lab Information Hub/caXchange, is a more generic platform for exchanging all types of clinical trial data and messages using service invocations and data exchange and is leveraged in the caBIG Clinical Trials Suite project. Furthermore, this application is the enterprise service bus for exchanging clinical trial information between applications and systems. The Java Business Integration (JBI) compliant Apache ServiceMix Enterprise Service Bus (ESB) is the platform on which various routing and transformation services and components will be deployed.

The following are the key features of the caBIG Integration Hub:

- Leverages open-source standards based on Apache Service mix to facilitate adherence to standards, vendor independence, and collaboration
- Integrates Clinical Connector for generic integration to CDMS vendors such as Oracle Clinical and OpenClinica
- Provides multiple connectivity options by supporting integration standards such as Web services, JEE, JMS, FTP, CSV files, and e-mail
- Provides an extensible and flexible ESB-based architecture using the JBI framework; new JBI components can be plugged in to extend and complement existing features
- Simplifies integration of multiple disparate applications
- Supports multiple eXtensible Markup Language (XML) file formats including Biomedical Research Integrated Domain Group (BRIDG) and Health Level Seven (HL7)
- Supports integration with caGrid and non-caGrid environments
- Supports synchronous and asynchronous processing
- Provides reliable messaging, reliable transactions, and high availability, as well as message transformation and notification capabilities
- Meets enterprise-class performance and reliability requirements
- Provides a configurable mechanism for quickly and easily adding and modifying integration scenarios using configuration files

5.5 EHR4CR

The EHR4CR (Electronic Health Records for Clinical Research) project [5.14] aims to design and demonstrate a scalable and cost-effective approach to interoperability, or the ability of health information systems to work together within and across organizational boundaries, between Electronic Health Record systems (EHRs) and Clinical Research. The project is divided into four work streams which addresses an entirely new approach to reusing EHR data (e.g. semantic interoperability, privacy enhancing techniques, and standards) for supporting medical research, underpinned by a comprehensive business model for governance, acceptance, adoption and sustainability.

The project will build a platform to enable the use of EHR systems for more efficient medical research and run pilots (on interoperability, security, data quality, data storage solutions, organisational issues, accreditation and certification) to demonstrate the viability and scalability of an EHR4CR business model.

The EHR4CR platform will:

- Enable trial eligibility and recruitment criteria to be expressed in ways that permit searching for relevant patients across distributed EHR systems and initiate participation requests confidentially via the patients’ authorised clinicians
- Support the feasibility, exploration, design and execution of clinical studies and long-term surveillance of populations
- Provide harmonised access to multiple heterogeneous and distributed clinical (EHR) systems and integration with existing clinical trials infrastructure products (e.g. EDC systems)
- Facilitate improvements of data quality to enable routine clinical data to contribute to clinical trials and vice versa thereby reducing redundant data capture

Figure 10: EHR4CR concept overview (taken from [5.14]).

- Enable clinical trials to be established and delivered more cost effectively at greater scale.

The EHR4CR project supports the IMI strategic agenda with an information gateway solution to enhance clinical research efficiency and innovation. A key IMI aspect is the development of a knowledge management capability that can provide information management support for other research on personalized medicines. EHR4CR also supports other IMI R&D projects by enabling the use (and reuse) of large amounts of health data – in an ethical and cost-effective way.

The EHR4CR project consortium draws its expert partners from academia, with 20 organisations and 4 SMEs working with 10 EFPIA companies and is an example of the scale of collaboration made possible through IMI.

5.6 Information Warehouse

The Information Warehouse (IW) [5.5] of the Ohio State University Medical Center (OSUMC) is the combination of four integrated components: a clinical data repository containing over a million patients; a research data repository housing various research specific data; an application development platform for building business and research enabling applications; a
business intelligence environment assisting in reporting in all function areas. The IW is structured and encoded using standard terminologies such as SNOMED-CT, ICD, and CPT. The IW was initially built in 1997 in order to replace an existing Decision Support System hosting data feed from a home-grown accounting system and to provide a seamless integrated environment for data for financial analysis with added provision for data cleansing and validation. The long term goal of the IW was to provide the physicians, administrators and analysts alike with a tool that allows them to monitor the hospital operations in an effort to gain efficiencies and render better cost effective clinical care for the patients, as well as plan and monitor the financial health of the organization. The scope of the IW was later expanded to support the research mission of the medical centre and to provide an informatics environment that helps facilitate both translational research and advances in personalized medicine.

The basic architecture of IW consists of the following parts:

- The Data Acquisition processes that extract or receive data from different operational and/or transactional systems and internal or external databases. Currently the IW collects data in real time from systems such as Admission, Discharge, and Transfer (ADT), Laboratory Information System (LIS), and Radiology, Cardiology, and Dictated reports. Less frequently collected data include Computerized physician order entry (CPOE), Operating Room System, dictated and Pathology Reports, and patient management data, and billing information. In addition, as the IW expands its service into areas such as basic and translational researches, data is also acquired in various intervals on Cancer Genetics, Tissue, genomics and proteomics. The IW maintains its Master Patient Index and Master Physician Index that are populated from incoming source systems. They are updated in real time as part of the data extraction, transform, and load (ETL) process.

- Data Transfer and Transformation: The data is verified against standard codes and values to check for its validity. The processes are automated to correct the values where possible and the errors are communicated back to the source system via an electronic report. Data is transformed for uniformity in data definitions, naming conventions and consistency in business rules.

- Data Storage and Management: Logical data model, created through the data-modelling tool to meet the business requirements, is converted into physical structures in a relational database. All details of the metadata relating to the refresh process or each of the data sources, including the mapping of the source system data elements to target columns are captured in a metadata storage within the IW.

- Data Access: A key factor in determining IW success is its ability to provide customers with data access tools that are intuitive, easy to use and cater to the very unique requirements of administrators, analysts, and clinicians. The IW data is available to its users (administrators, analysts, and clinicians) via the web, online analytical processing (OLAP) tool for multidimensional analysis and an ad hoc query and reporting tool, and a Business Intelligence (BI) tool for parameterized reporting.

### 5.7 Stanford Translational Research Integrated Database Environment (STRIDE)

The Stanford Translational Research Integrated Database Environment (STRIDE) [5.4] is a research and development project at Stanford University to create a standards-based informatics platform supporting clinical and translational research. STRIDE consists of three integrated components: a clinical data warehouse, based on the HL7 Reference Information Model (RIM), containing clinical information patients cared for at Stanford University Medical Center (SUMC); an application development framework for building research data management applications on the STRIDE platform and a biospecimen data management
system. STRIDE’s semantic model uses standardized terminologies, such as SNOMED, RxNorm, ICD and CPT, to represent important biomedical concepts and their relationships. The system is in daily use at Stanford and is an important component of Stanford University’s CTSA (Clinical and Translational Science Award) Informatics Program.

STRIDE is built on the Oracle 11g relational database platform and uses an n-tiered architecture. Data is stored using an Entity-Attribute-Value (EAV) model and is represented by object-oriented data structures (entities, roles and acts) derived from the HL7 Reference Information Model (RIM). The system includes a Master Person Index (MPI) that is dynamically populated from clinical, research and biospecimen data.

The STRIDE physical database layer is organized into three logical database partitions: (1) a Clinical Data Warehouse (CDW); (2) Research Database Management supporting multiple logically separate research databases and (3) the Biospecimen Data Repository, which supports multiple separate biospecimen databases. All three components rely on the same underlying architecture and services with data be linked across all three partitions.

STRIDE’s semantic layer consists of a framework supporting multiple terminologies, including ICD9-CM, ICDO, CPT, RxNorm and SNOMED. This mixed terminology model supports standards-based data entry, data integration, hierarchical concept-based retrieval and data interoperability. As an example, using RxNorm to represent pharmacy data in STRIDE allows the system to merge drug information from the two different vendor drug models used at SUMC, with dynamic integration of pediatric and adult medication orders within the CDW. Additionally, the semantic model allows linkage from RxNorm to SNOMED drug classes.

References


[5.7] https://cabinet.nci.nih.gov/tools/caBIGIntegrationHub

[5.8] https://cabig.nci.nih.gov/adopt/CTCF

[5.9] https://cabinet.nci.nih.gov/tools/C3DClinicalConnector


6 Linking Electronic Health Records and Trial Management Systems

In this chapter we focus on reusing electronic health record data in clinical trial management systems to enhance clinical trial processes and especially to avoid double data entry into these systems. In the last years, the time and costs for clinical research by investigators, academia, government agencies, and industry have escalated. This has resulted in “fewer new innovative medicines, more expensive and late delivery of therapies, failure to explore niche markets of high medical need, and the tendency to focus research resources on environments offering higher return on investment” [6.9]. This results partly from the fact that currently there is a major gap between clinical care and clinical trial processes at the health care provider site [6.1]. Different, incompatible systems are used for clinical care and clinical research and often the same data need to be entered several times. The current systems impede research activities by introducing inefficient processes making clinical research prohibitively expensive and slow. Such inefficiencies in clinical trial data collection cause delays, increase costs, and may reduce clinician participation in medical research.

To overcome these obstacles, many authors have endorsed secondary uses of healthcare data in CTMS [6.7], [6.3]. There are expressed goals for the direct use of EHR data for clinical trial management systems to a) facilitate clinical research for sites/investigators by enabling the entry of data once for research and healthcare; b) reduce the number of duplicate samples taken from subjects of trials who also are patients receiving healthcare; c) maximize the use of information from healthcare for the research benefit of the population as a whole and others. Creating an environment for more efficient clinical research integrating secondary uses of healthcare data in CTMS can contribute to increased quality of health care and clinical research, reduce transcription errors, increase sponsor and site personnel efficiency, facilitate information flow and improve timelines of data. Such approaches have the potential to reduce cost and time and increase productivity in clinical research.

In the following we will describe several research initiatives to avoid redundant data entry into clinical trial management systems.

6.1 Extracting Data from Electronic Health Records to Clinical Trial Management Systems

The eSDI project is a joint initiative of FDA and CDISC to encourage the use of eSource data (e.g. EHRs) in regulated clinical research, leveraging CDISC standards. In [6.2] they examine the current regulatory framework for clinical trials, user requirements and current practice. Based on this analysis they have developed several scenarios that they believe will permit to deploy new technologies for the capture of eSource clinical trial data within the context of existing regulations while ensuring quality and integrity of the collected data. In three of these scenarios they especially focus on the reuse of electronic health record data for clinical research, to facilitate interoperability between clinical research and healthcare systems and information sharing between these two arenas. We will describe these scenarios in the following sections.

6.1.1 Direct Extraction

The Scenario Direct Extraction of Electronic Health Records [6.2] addresses the implementation of direct extraction of data from electronic health records for clinical research as shown in Figure 11.
However, in [6.2] it is clarified that in the context of the existing regulations, the direct extraction of data from electronic health records for the reuse in clinical trials requires that the EHR system meets the requirements of FDA 21 CFR Part 11. This is mostly not realistic for current EHR systems since it puts too much requirements on the system. E.g. if the EHR system interfaces with other systems that may include such data as that used for billing, admissions, and insurance; the hospital will need to comply with the required validation process according to FDA 21 CFR Part 11 for the entire system. However, if an EHR can meet the existing regulations and the requirements (e.g. a stand-alone EHR application designed for clinical research), then it is acceptable to reuse the data in CTMSs and to extract clinical research data from the EHR [6.2].

6.1.2 Extraction and Investigator Verification

To allow reuse of data stored in EHR systems that do not meet the requirements in FDA 21 CFR Part 11 in a CTMS, the eSDI initiative has described the scenario Extraction and Investigator Verification (Electronic Health Records) that is compliant to today’s regulations for clinical trials. It is similar to the previously described scenario, but in this scenario the data extracted from the EHR needs to be validated manually before it is stored in the clinical trial management system. Therefore, an additional process step is added in this scenario as an interim step to allow a more direct use of EHR in clinical trials without the necessity to validate the entire EHR system. This step is for the investigator to verify that the extracted data, for clinical research use, accurately reflects the source data for that subject before it is included as part of the clinical trial data record.

A pilot implementation of the Extraction and Investigator Verification scenario, called Siemens Integrated Clinical Trial Solution, is described in [6.1] for a hospital in Munich. The solution bridges the clinical care and clinical trial world by re-using electronically available data in the HIS that are relevant to a clinical trial in the according CTMS.
Technically, the Integrated Clinical Trial Solution is realized with a 21 CFR part 11 compliant integration engine that bridges the HIS and the CTMS, as shown in Figure 13. For scheduled visits in a trial, data are documented in the HIS (IS-H med, Lab, PACS) and from there transferred to the integration engine by a HL7 message. The integration engine translates the trial relevant data automatically into the CDISC ODM format and sends the data to a validation buffer. The translation is possible, based on a manual mapping of the hospital specific HL7 messages to the trial specific CDISC ODM files that needs to be done once, when setting up the trial. The validation buffer requires a human interaction of qualified personnel to review the data, check the match with the patient as well as the source data at the clinical documentation site and then manually confirm the data transfer into the CTMS [6.1].

Zahlmann et al. [6.1] evaluated the Integrated Clinical Trial Solution on the clinical trial process design and on data quality and costs, by using a real life clinical trial conducted following a specific clinical trial protocol. They report that within their evaluated trial a considerable high amount of the eCRF data could be filled in automatically from the HIS into the CTMS using their solution. They state that for screening visits around 48% and for chemotherapy visits around 69% of the required data could be filled in automatically.

Furthermore, they report that the time spent per week by personnel involved in the trial was reduced in comparison to a traditional paper-based trial process. They observed that on average, the personnel spent 375 minutes per week on the trial in the traditional process. The integrated solution required 98 minutes.

However, they do not report on the effort that is needed for the manual mapping of the CDISC ODM to the HL7 messages that is necessary for the approach and, furthermore, not on the terminologies needed in the CDISC ODM or the HL7 messages.
6.2 Single Source Concept

The third scenario that was outworked by the CDISC eSDI initiative is the Single Source Concept shown in Figure 14. It is a solution to simultaneously populate an electronic healthcare record and a clinical trial management system while adhering to existing regulations for clinical trials and healthcare and leveraging healthcare and clinical research standards. In this scenario, data are entered into an electronic source document (typically as an interface to the electronic health record (EHR) system but conceivably as an interface to an EDC system). All of the eSource data can flow into the EHR database, while the clinical trial data (as identified by the protocol) can be simultaneously passed into eSource repository and passed onwards to the clinical trial database.

In [6.2], the eSDI initiative concluded that this scenario is not the ideal future methodology to facilitate clinical research by investigators; however, it does offer a viable means for data to be entered just once for multiple purposes (research, patient care, safety surveillance, etc.) within the context of existing regulations. This would presumably facilitate the processes at an investigative site and also eliminates data transcription, which is a point of potential error introduction.

A proof-of-concept has been implemented in the single source project by CDISC [6.7]. Kush et al. examine in this project, the feasibility of using point-of-care data capture to populate simultaneously both a hospital information system (clinical document repository) and a clinical trial management system in the setting of a working clinical trial. They aim to increase reuse of patient data, eliminate redundant data entry, and minimize disruption to clinic workflow.

They have developed an application to record data during patient visits on CRFs based on Microsoft InfoPath. The application stores data in an HL7 CDA document that forms the single source. The source CDA (CDA CRF) was transformed to the CDISC ODM for integration into the CTMS and was furthermore transformed into a clinical note (CDA CN). The pre-populated clinical note needed to be completed by a physician and was then stored...
automatically in the clinical document repository. They evaluated their system within a running clinical trial, the STARBRITE trial.

![Diagram of Organisation Boundary, Site, EHR, Clinical Research, Sponsor, and Investigator sphere of control over eSource data]

**Figure 14. Scenario Single Source Concept described by eSDI initiative, taken from [6.2].**

To define semantics in the CRFs and in the clinical notes they used the semantics defined in the CTMS, i.e. they reused the unique keys in the database of the CTMS as identifiers into the CRFs and the clinical note. That means that they did not use a commonly understandable semantics, as a shared vocabulary or ontology. Therefore, their approach required a manual non semantic mapping between the CRF, clinical note, ODM, and CDA for the data collected by the study. They concluded that this is neither desirable nor scalable for use in prospective multicenter clinical research; however, until standards for content and computable semantics exist and are commonly used for health care data, it will remain a necessity. They state that in future projects they would use the CDISC submission data model. However, they have not used it in the single source project since it had not been defined at project inception. They describe the advantages of a possible usage of the CDISC submission data model as follows:

“If the ODM for STARBRITE had used the data definitions in the CDISC submission data model, a semantic mapping to CDA and the RIM would have been possible, although it would still have been specific to STARBRITE. Use of a submission dataset could have provided a common semantic; this would constitute a good test for future projects. The lack of a higher-level definition of semantic structures to create a reusable map that would retain its usefulness beyond a single trial protocol led directly to efforts to create a Clinical Research domain model under the RIM. Partly as a result of this project, the CDISC ODM has now been mapped to the HL7 RIM. The semantics, however, warrant consideration at a higher level of abstraction. A key finding is that in the structured narrative note, the metadata lacked the semantic structures and controlled terminology required for direct machine processing. Thus, a manual mapping of data elements was necessary. Unfortunately, this situation exists for healthcare data in narrative text and structured form, as standards for the clinical content and the expression of that content in an electronic format reusable across computer applications do not yet exist for most therapeutic areas in healthcare. As long as this situation persists, there will be no large-scale interoperability, and reuse of healthcare data in general will exist as isolated implementations.” [6.7]
They evaluated their system within the STARBRITE trial, with five sample CRFs and clinical notes. The system was integrated in the workflow of the ongoing trial and worked in parallel with data collection procedures already in place for the trial. It was tested in two live patient encounters.

They observed that for the sample CRFs and clinical notes they evaluated (five sample cases) there was a significant overlap between clinical note and CRF, approximately 75% of data fields, but that neither constituted a superset of the other. However, their analysis also showed that even in cases where the same data were present in both clinical note and CRF, presentation and sometimes even values differed. For example, medications might be recorded using either generic or brand names. Exposition of data also differed widely: for instance, vital signs were recorded in a table on the CRF but were captured as part of a narrative in the clinical note.

They concluded that although subject to the limitations of a small feasibility study, their study demonstrated that electronic patient data can be reused for prospective multicenter clinical research and patient care, and demonstrate a need for further development of therapeutic area standards that can facilitate researcher use of healthcare data.

### 6.3 Reusing EHR Data for the Recruitment of Patients for Clinical Trials

Another promising scenario to reuse EHR data in clinical trial management systems is the recruitment of patients for clinical trials. Today, efficient patient recruitment in clinical trials is a common problem and often the recruitment process is delayed. Campbell et al. have observed in a review of more than one hundred trials that less than one third of the studies managed to recruit their original target within the time originally specified. A solution for this problem is to integrate EHRs into the recruitment process, since such systems contain data relevant for inclusion or exclusion criteria of these studies. Therefore, several researchers have explored new ways of patient recruitment based on extracting data from existing EHRs [6.4], [6.4].

In [6.4] an overview about the various approaches is given. It is described that examples are the application of computer-based decision support algorithms to detect patients with HIV/AIDS for clinical-trial eligibility, to check for eligibility in breast cancer studies, the use of an expert system with a Bayesian network concept to identify patients for clinical trial protocols, a sepsis-alerting system, usage of a decision support system, designed to provide best therapeutic recommendations for breast cancer patients, for eligibility screening and the successful integration of clinical trial decision rules in an electronic medical record. Recently, such approaches have also been applied in a German hospital information system [6.5] and the authors concluded that automated notification workflows as components of routine hospital information systems can be successfully used to support patient recruitment for clinical trials.

Nevertheless, so far only few studies have systematically investigated the effect of clinical trial recruitment that is supported by EHR data extraction and these reports illustrate, that results may vary and depend on the type, setting and complexity of a trial. While some researchers have reported significant benefits to clinical trial recruitment rates and streamlined patient recruitment by improvements in the speed of patient identification and higher accuracy of assessing eligibility as well as enhanced study investigator notification, others have mentioned a disappointing impact on patient accrual despite better identification of potential candidates. However, in [6.4] it is concluded that for the application of HIS systems in clinical trials, prerequisites, such as the provision of adequate clinical
terminologies and data interoperability between the worlds of medical care and clinical research are necessary [6.4].

6.4 Initiatives to fully Integrate Healthcare and Clinical Trial Systems

While the solutions we have described in the last sections aim to link current HIS and CTMS within the current regulations, most researchers do not see these approaches as the ideal future methodology to integrate clinical care and research. They describe the ultimate goal as fully integrated EHR and clinical trial management systems, ideally integrating clinical care and clinical research activities in one unified system environment harmonizing the processes in the two areas. [6.2] Already in the eighties, several researchers have implemented clinical research functionalities (e.g. analysis and retrieval possibilities) within hospital information systems, mainly in institution specific-approaches. However, combining clinical research and clinical care activities into one unified electronic information system requires integrating a substantial body of regulatory requirements and institutional policies. Because of those complexities, the early approaches to integrate clinical research functionalities in a hospital information system, normally have not tried to apply those methods for FDA regulated clinical trials. [6.4] Furthermore, it is necessary to harmonize the different semantics used in the two areas.

Several current research initiatives try to solve these issues. One of them is the EHR4CR project that is funded under the Innovative Medicine Initiative. As we have described in Section 5.5 one of the projects main goals is the harmonization of multiple heterogeneous and distributed clinical EHR systems with existing clinical trial management systems. Other prominent initiatives focusing on these issues are the EHRCR Project, the Healthcare Link Initiative and the Clinical Research Value Case Workgroup that we will describe in the following sections.

6.4.1 EHRCR Project

The global EHRCR Functional Profile Project is a collaborative effort (initially of eClinical Forum and PhRMA along with HL7 and EuroRec) to expand and adapt the functionality of EHR and associated systems, networks, and processes to support clinical research. The project aims to develop a Functional Profile that identifies critical capabilities for the conduct of clinical research utilizing EHR systems and establishes conformance to the HL7 EHR Functional Model and the Q-Rec EHR Certification Criteria. This project is the first step to build upon current work being done by HL7 and QRec, and to develop criteria at a global level that can facilitate the conduct of global clinical research [6.8], [6.9].

The main project objectives as described on their website [6.9] are:

- To further expand the use of EHR systems for clinical research processes such that clinical research is optimized for clinics and hospitals, allowing new therapies to be available to patients in the shortest time at the lowest cost.
- To ensure that EHR systems, when used to collect source data in support of claims made regarding the safety and efficacy of new medical products, can be trusted by regulatory authorities to be a ‘reliable’ data source.

The EHRCR project provides an integration roadmap that visualizes different tiers towards the aim of fully interoperable EHR and research systems (shown in Figure 15). Tier 0 describes current possibilities to link EHR and Research systems considering current requirements. The scenarios that we have described in Sec. 6.1 and 6.2 fall into Tier 0. The ideal future is described in Tier 3 as EHR and research systems that work seamlessly
together. In this scenario the EHR holds the complete patient medical record including all clinical study and research data.

![Emerging and Future EHR-Research Connectivity and Complexity](image)

Figure 15. EHRCR’s vision of emerging EHR – CTMS interoperability (taken from [6.8]).

The EHRCR project team has produced a number of deliverables including:

1. **A user Requirements document** outlining the project vision of fully integrated healthcare and research systems as well as clearly showing the minimum Regulatory-mandated clinical research requirements that must be met in order to use electronic data from electronic health records as source for clinical research.

2. **The EHRCR Functional Profile (approved by HL7, ANSI, and EuroRec, and under joint consideration by CEN and ISO)[6.8]** delineates the high-level requirements necessary, based on the User Requirements (above) and the succinct criteria that can be used to evaluate EHR systems for use with clinical research. The EHRCR Functional Profile aims to provide practitioners, research community and regulators with a level of confidence that the integrity of clinical research data is protected, source data are stored in a manner compliant with clinical research regulations and process redundancy is minimized. The EHRCR Functional Profile is intended to provide high-level requirements necessary for using electronic health record data for regulated clinical research, and to further provide a roadmap toward an evolutionary process of integrating the environment that provides both patient care and data for clinical research. This functional profile is aimed at encouraging EHR vendors and developers to incorporate functions into their products that are necessary to utilize the Electronic Health Records as a direct data source for clinical
It is intended to provide one overall view of the needs of regulated clinical research with respect to electronic patient records.

6.4.2 Healthcare Link Initiative

The CDISC Healthcare Link project [6.10] focuses on solutions to enable full interoperability between healthcare systems (e.g. the EHR) and clinical research systems. They especially aim to avoid double data entry of research data. The project was started in 2005 and based on the work of the eSDI project that we have described in Sec. 6.1 and 6.2.

Within the project CDISC and IHE have described a first solution to link EHRs and clinical research systems. This approach uses an IHE integration profile that is called “Retrieve Form for Data-capture” (RFD), together with CDISC standards to collect relevant data from the electronic health record for secondary uses such as Clinical Research, Safety Reporting and Disease Registries.

RFD provides a method for gathering data within a user’s local application to feed the data into an external system. RFD supports the retrieval of forms from a form source into the user’s local application, display and completion of a form in the familiar way, and return of data from the local application to the source application. RFD can e.g. support scenarios as the following. A hospital uses an Electronic Health Record (EHR) to document patient care. For the hospital’s personnel the EHR is the local home application. A CTMS requires data that is partly stored in the EHR database, partly requires additional data entry by the users of the EHR. RFD enables the EHR user to retrieve a data capture form from the CTMS, to fill out the form, and to return the data back to the CTMS without leaving the familiar EHR.

Reaching through to the EHR in this way to pull key data of interest to clinical research that is already existing in the EHR has the advantage that it creates system interoperability and improves data quality and most importantly timeliness of data sharing while alleviating the Investigator site from supporting and entering data in to multiple redundant data collection tools for the purpose of the secondary uses.

RFD permits automatic form population and provides a generic mechanism by which this can be accomplished. However, the profile does not speak to the issue of content, remaining silent on normative vocabularies and other enablers of semantic interoperability. Specific domain groups – clinical trials, drug safety, bio-surveillance – will build on RFD by contributing content specifications or by evaluating and recommending existing content.

![Figure 16. Health Care – Clinical Trial Scenario with and without RFD integration (taken from [6.12]).](image-url)
standards that will operate within RFD. When RFD, as an infrastructure profile, integrates with domain-specific content standards, a much greater level of interoperability will result. Furthermore, RFD offers the capability to leverage industry standards that address both the structure and content of forms used for data capture. HL7’s Individual Case Safety Record (ICSR) and CDISC’s Operational Data Model (ODM) provide examples. The CDISC Healthcare Link project aims to augment RFD soon by additional IHE profiles to form an ever growing toolkit. They anticipate that the efforts made thus far by the Healthcare link Initiative will provide a foundation for the harmonization of standards between EHRs and clinical research.

Several life interoperability demos of RFD have been shown as e.g. described in [6.12]. These demos show several use case scenarios which demonstrate how clinical data can be transferred seamlessly between care EHR systems and systems used for clinical research, disease registries, safety surveillance, and disease surveillance.

6.4.3 Clinical Research Value Case Workgroup

The Clinical Research Value Case Workgroup [6.13] has been founded by the American National Standards Institute (ANSI) to promote convergence within the global clinical research and healthcare arenas. The aim of the workgroup is to identify priorities for the harmonization of the technical standards that are necessary to ensure the interoperability of EHRs and clinical research applications based on their potential for providing value to stakeholders. Priorities identified by the workgroup will then be transmitted to the Healthcare Information Technology Standards Panel (HITSP) for harmonization and the development of HITSP Interoperability Specifications, which represent standards and provide implementation guidance.

The workgroup has developed a detailed use case [6.14] regarding the use of Electronic Health Records in Clinical Research based on the needs of the American Health Information Community (AHIC), called "Core Research Data Element Exchange Detailed Use Case". The scope of the use case is focused according to [6.14] on:

- The ability to communicate study parameters, eligibility information, results, and case report forms within the research community and
- The ability to exchange a core dataset of de-identified or anonymized information from the EHR for use in clinical research.

In [6.14], the workgroup discusses several issues and obstacles that are applicable for the realization of the use case, as e.g. that there is currently a lack of financial, network, technical, and policy infrastructures to enable information exchange that is secure and consistent. As one of the main obstacles the gap between clinical care and research standards is identified. They point out that clinical research data standards are developing independently from certain standards for clinical care data and a lack of harmonized standards including consistent terminology, nomenclature and semantics used to exchange clinical research hampers interoperable exchanges of that information. They identify the need to standardize terminology for all clinical research related information and to harmonize these standards with those developing for clinical care and the need for a core dataset of patient-level clinical information between EHRs and clinical research systems. Therefore, the focus of the document is on the harmonization of standards leveraging a core dataset of widely useful clinical care-data from EHR systems to increase the effectiveness and the efficiency of clinical research activities.

Furthermore, the use case stakeholders and the information flow for the workflow and information exchange process of different types of clinical studies are described in detail. They especially focus on the extraction of data from the EHR into the CRFs of the CDMS during the phase of study conduction. In particular, during recording all study related information into the CRFs, a core set of information can be extracted from the EHR, while
other information may be study specific and needs to be recorded on an ad hoc basis by the clinical care team into the CRFs. They estimate that the percent of information coming from the EHR varies from 5% to 40%, depending on the specifics of the particular study design. A prerequisite for the extraction of data is, that in the EHR for each trial patient is indicated that the patient is enrolled in the trial and his trial id must be stored. Furthermore, the EHR must be capable to exchange the anonymized core data set for the patients.

The workgroup provides considerations to create a harmonized core dataset and describe prerequisites for such a data set, e.g. they enumerate categories of data elements which are generally found in case report forms and EHRs. However, they do not suggest such a data set and identify that as important future work.

6.5 Conclusion

In this chapter we have described several projects that aim to link EHR and CTMS systems, focusing especially on approaches to avoid redundant data entry into these systems. One of the main obstacles is the lack of semantic interoperability between current EHR systems and CTMSs partly resulting from a lack of harmonized semantic standards in the areas of health care and clinical research.

On the one hand we have described several approaches to link current HISs and CTMSs in current healthcare settings. These approaches solve the semantic interoperability issue by mapping the data model of the EHR onto the CRFs used in the CTMS. In general a tedious, manual mapping is required for each conducted trial.

On the other hand we have reviewed research initiatives that describe requirements to achieve integrated EHR and clinical trial management systems, with the aim to achieve full semantic interoperability of the systems in the future. In these scenarios CTMS and EHR systems in general need to fulfill special requirements, as e.g. be compliant to specific data sets, which are mostly not fulfilled of current systems.

The approaches described in this chapter are mainly research initiatives and are not widely applied in practice. Providing full semantic interoperability between HIS systems and CTMS systems is still a challenging issue.

References


[6.2] Clinical Data Interchange Standards Consortium, Electronic Source Interchange (eSDI) Group, Leveraging the CDISC Standards to Facilitate the use of Electronic Source Data within Clinical Trials, v. 1.0, 2006

[6.3] John Powell, Iain Buchan, Electronic Health Records should support clinical research, In Journal of Medical Internet Research, 7(1), 2005


[6.11] IHE IT Infrastructure, Technical Framework Supplement, Retrieve Form for Data Capture (RFD), Trial Implementation, August 2010


7 Conclusion: Approach taken by p-medicine

In p-medicine, there is a need to gain access to clinical data stored in different data sources, such as hospital information systems and clinical trial management systems, in a transparent way, following appropriate pseudonymization and security procedures. It is of outmost importance to harmonize the data semantically in this process, in order that the semantics of the data is clearly defined, a precondition to reuse the data for reliable analysis and simulations.

The state of the art review described in the last chapters has shown that currently, a variety of semantic resources that describe different aspects of medicine, health care and clinical trials exist for this purpose and are utilized in different medical information systems. However, in our investigation several problems became clear. Firstly, a lot of standards which provide semantics and could improve semantic interoperability are only rarely applied. The few standards which are commonly in use, on the other hand, mostly provide no deep specifications of terms in a machine-readable format. For example, the widely used system ICD-10 is not suitable as a semantic resource for automated reasoning which would be a desideratum for machine-readable semantics. Furthermore, semantic standards for clinical care and clinical research are often developing independently, resulting in a lack of harmonized standards.

The approach in p-medicine to solve these problems is to develop a new semantic reference for p-medicine’s purposes, the Health Data Ontology Trunk (HDOT). HDOT will not be developed from scratch, but harmonize and link existing semantic resources, which are described in the last chapters, under one umbrella. It especially takes advantage of the fact that it integrates resources, which are widely applied and well-known by medical staff, as e.g. ICD-10. The provision for allowing the use of such terms and codes makes HDOT user-friendly. In particular, HDOT will provide a machine-readable semantic background for these resources. That means that as opposed to many existing resources, HDOT has underlying machine-readable axioms and supports automated reasoning.

HDOT will provide the semantic basis for the data integration and sharing in p-medicine. In order to enable the integration of data from different data sources based on the HDOT ontology, the p-medicine semantic layer will be developed, that is responsible to resolve semantic interoperability issues. This approach will foster the secondary usage of HIS and CTMS data for various use cases important to realize personalized medicine as e.g. simulations, data mining and the reuse of HIS data in the ontology-based clinical trial management system ObTiMA. To enable such scenarios push and sync services will be developed based on the p-medicine semantic layer to support the reuse of the data.

In the following we will give an initial idea of the p-medicine semantic layer and describe the push and the sync services. Details of the p-medicine semantic layer will be described in deliverable D4.3 taking into consideration the state of the art review presented in this deliverable.

7.1 The p-medicine Semantic Layer

The p-medicine technological platform is a framework comprised by tools and services aimed at biomedical researchers and biostatisticians. The platform includes a federated Data Warehouse (DW) for storing heterogeneous data stemming from external repositories. These repositories range from private databases from hospitals and research institutions to public biomedical databases accessible through the Internet. With the aim of enabling semantic integration of data in the DW and allowing users to perform integrated queries over the data, the p-medicine Semantic Layer (SL) of tools and services will be developed. This software
layer will follow an ontology-based approach for achieving semantic integration of data, with the HDOT ontology acting as global schema of the integrated data sources.

Given that a data warehouse-based approach has been selected for the project’s data managing process, the SL will be in charge of implementing the operations of an Extract-Transform-Load architecture. In that sense, the SL will comprise two main kinds of tools:

1. On the one hand, tools for harmonizing data sources with HDOT will be provided. The SL supports the annotation of existing heterogeneous data sources with HDOT as well as the HDOT-compliant set up of new data sources for clinical trials. To support the first scenario the Ontology Annotator Tool will be provided, which is aimed at external users (mainly database administrators) who wish to include their databases in the project framework. The second scenario is supported by the ontology-based CRF Creator, a component of p-medicine’s trial management system ObTiMA.

2. On the other hand, the SL comprises a tool that provides translation services to the DW for homogenizing data, the so-called Data Translator.

The SL will additionally include a tool for aggregating semantic resources in an automated fashion under HDOT (the Ontology Aggregator Tool). The goal is to explore ways of creating HDOT-modules automatically from a set of pre-selected ontologies. Additional details, characteristics, and features of this tool will be provided later in the project. In the following we will describe the components of the semantic layer in more detail.

7.1.1 The Health Data Ontology Trunk

In this section we will summarize the main design principles and advantages of HDOT. We refer to deliverable D4.1 for a detailed discussion.

HDOT is developed as a middle-layer ontology that provides the base to harmonize and link existing semantic resources under one umbrella. Its design is governed by three main related structural considerations in order to achieve the highest level of semantic interoperability between heterogeneous data sources, maintain a high level of ontological soundness, and ensure a high degree of expandability:

1. The HDOT level of generality is designed in such a way that HDOT classes and relations cover all areas of the health-care domain, i.e. there is a meaningful ontologically well-defined HDOT super-class under which all necessary parts and pieces of semantic data descriptions (annotations, metadata) can be directly subsumed or otherwise represented.

2. The core ontological structure integrates different modular ontologies at different levels of granularity. Each class is provided with a deep axiomatization, which guarantees p-medicine’s work-flow high degrees of both, representation and reasoning, together with the ability to construct defined classes and composite terms. The semantics of the HDOT central body will change in the further development of the project only in case problems related to HDOT’s application to the project itself or clinicians’ needs emerge.

3. HDOT’s modules for specific applications can be obtained by stating further specifications of HDOT classes, i.e. by inserting subclasses into provided HDOT slots (super-classes). Generally speaking, an ontological module can be defined as the specification of one ontology’s class in more classes related to a specific portion of a particular domain for some specific purposes.

It is likely that users are able to employ existing semantic resources which already cover these more specific domains. Such new classes constitute a “module” because they can be considered both as part of the ontology and as an ontology by itself. Indeed, each module can be navigated in an ontology browser together with the import of HDOT, or by itself.
In other words, some classes in an ontology can be retrieved from an external semantic resource and subsumed under appropriate superclasses provided in HDOT as integration nodes of numerous other potential ontological modules. However, one should always bear in mind that this integration is only a partial one because it will not be possible to include all axioms in the integrated module these classes might have as this could lead to ontological inconsistencies depending on the approach to axiomatization in the semantic resources from which the integrated classes are taken.

The main advantages of such an approach are the following:

- **Greater ontological flexibility:** We do not aim to incorporate all necessary annotation values or metadata in one single monolithic resource. Instead, we enable a user driven expandability with the help of the Ontology Aggregator Tool to be developed in WP4.

- **Re-use of resources and deeper level of axiomatization:** We provide an axiomatized broad conceptual frame in which users can import relevant parts of pre-existing semantic resources, thus on the one hand allowing their re-use and on the other simultaneously providing their classes (or terms) with deeper axiomatic (at least hierarchical) specifications. In most cases, HDOT provides for the expression of more relations between classes than just a subsumption relation.

- **Provision of a machine readable semantics:** We provide a machine readable semantic axiomatizations for class definitions in order to facilitate the realisation of a very high level of semantic interoperability amongst heterogeneous data sources.

- **Utilizing defined relations:** We provide an intuitive method to use defined relations between classes to generate new defined classes, thereby enabling the definition of the reference of complex terms within one conceptual frame and without the need for a dedicated mark up language.

### 7.1.2 The Ontology Annotator

The Ontology Annotator Tool (OA) is aimed at providing users with a GUI to annotate an existing database schema in terms of the HDOT ontology. The results of this annotation process will allow to seamlessly access heterogeneous data in terms of the HDOT model, thus achieving semantic integration of such data. The tool will be responsible of assisting users in establishing semantic correspondences between an external database schema and the HDOT model.

Heavy effort will be put on usability. The goal is to allow users lacking deep knowledge on semantic technologies to easily use the tool. The graphical interface will hide in a high degree the complexities of RDF and OWL models, making it easy to use for relational, Access, or Excel database administrators. For that purpose, a thorough study of existing applications on the area and a deep analysis on graphical representation of RDF models will be carried out prior the actual implementation of the tool. The application will be accessible through the Internet with any traditional web browser, facilitating users the access to the tool. It will allow creating annotation projects with several types of databases (Excel, Access, SQL or RDF formats will be accepted) or editing existing projects.

The process of annotating a database in terms of an ontology consists on establishing semantic equivalences between views of the two schemas. In this regard, the schema provided by the user will be automatically translated into an RDF-based form. Both schemas will then be graphically presented to the user so he can establish the required correspondences. This process will be supported by two tools: i) the MappingAPI 2.0 and ii) the OwlBasicModel 2.0. These are both evolutions from previous developments carried out in
the former European project ACGT. The MappingAPI 2.0 is a Java-based API for creating semantic equivalences of two RDF models and serializing them into an XML format. The OwlBasicModel 2.0 is also a Java-based API for browsing RDF and OWL models in an efficient manner. The projected architecture of the OA is depicted in Figure 17.

The OA will be supported by the security framework to ensure access to the tool and its resources is only achieved by validated users.

7.1.3 The ObTiMA Ontology-based CRF Creator

ObTiMA has been chosen as p-medicine’s trial management system, because of its ability to integrate an ontology into the trial building process. While the current version of ObTiMA integrates solely the ACGT master ontology (s. Sec. 4.2.5), in p-medicine ObTiMA will be enhanced to integrate the HDOT ontology.

Therefore, in p-medicine, the ObTiMA ontology-based CRF Creator will allow a clinical trial chairman to design CRFs compliant to the HDOT model. While creating the CRFs, all information can be defined which are necessary for the data integration, i.e., each CRF item is described based on HDOT concepts together with metadata, like data type and measurement unit. Based on these definitions the trial database for conducting the trial can be configured that is compliant to the HDOT ontology. Therefore, data that is collected during a trial can be accessed in terms of the HDOT model, thus achieving seamless integration into the semantic layer of p-medicine. The ontology annotation for a trial can be automatically created from the trial definitions using the Mapping API 2.0.

7.1.4 The Data Translator

The Data Translator service (DT) will be in charge of performing the actual translation of data from annotated databases to an HDOT compliant format. The DT takes as input the data of one database and its ontological annotation (previously generated with the OA or ObTiMA), and returns the data translated to an HDOT compliant form. It will be accessible through a REST interface.

The data translation process will imply analyzing the given data according to the provided ontology annotations and generating a data model with the results of that analysis. The DT service will be supported by ad-hoc software modules for accomplishing these tasks. The DataAnalyzer module will be responsible for driving the analysis of the input data. The ModelConstructor module will generate an RDF model from the results of those analyses. In addition, the DT service will make use of the MappingAPI and OwlBasicModel modules, described above. Figure 18 presents the initial architecture of the DT.
The process of translating data must also take into account data previously translated from other sources. The DT service will be able to identify duplicate information. To accomplish this, the DT will query the DW in search of common identifiers from already translated data. The ModelConstructor module will take these results into account for building the RDF model prior submitting it to the DW.

More detailed descriptions of the tools and services that will comprise the p-medicine Semantic Layer will be provided in deliverables D4.3 and D4.5.

7.2 Approach for Push and Sync Services

The secondary use of data from HIS and CTMS in p-medicine will be enabled by the so-called “push” and “sync” services, which are based on the p-medicine semantic layer. The relevant data flow is sketched in Figure 19. Push services will support data owners to retrieve data from heterogeneous clinical information systems and other biomedical databases and import them into the p-medicine’s data warehouse. They integrate the data semantically in terms of the HDOT ontology by using the tools of the semantic layer so that the data in the DW can be stored in a form compliant to this ontology. The data warehouse supports the secondary use of the data by providing an SPARQL interface that allows querying the data in terms of the HDOT ontology. This interface supports the seamless reuse of the data by the various analysis tools and services of the p-medicine platform, as e.g. data mining, simulation and decision support tools.

Furthermore, the sync services will allow the reuse of the data that are kept in hospital information systems in running clinical trials in ObTiMA. In the following we describe briefly the initial idea of the push and sync services. A detailed description of these services will be given in deliverable D8.5.

7.2.1 Push Services

The p-medicine platform aims to provide the tools and interfaces for bridging with existing systems and gaining access to the available data. The semantic substrates of the platform, its integration and security requirements, its governance, etc. impose a tighter data integration approach than a simple “linking” solution. Therefore an “Extract, Transform, Load” (ETL) methodology has been adopted in order for the data stored in external databases to become available to the p-medicine users in the DW. In general this approach requires the export of the data from their original sources, their transformation to the necessary format, and their upload to the new repository.

The whole transfer of the data from medical information systems into the federated data warehouse will be implemented following a ‘push’ concept. That means data transfer is initiated and controlled by the owner of the data, who can be e.g. a trial chairman or a clinician. The reason for this is evident since the p-medicine tools do not necessarily have access to the original data sources and there can be a large variety of these systems. What is required from the original data sources is a way to export their data in a machine-readable format that is fully documented.
The next two phases of the ETL process, i.e. the transformation and the upload activities, are taken care by the specialized tools of p-medicine under the control of the data owners and curators. The transformation tools should first of all remove any personal identification and the data are pseudonymized according to the p-medicine security framework. This framework guarantees that data from same patients can be linked in the data warehouse. Further details of the security framework and the pseudonymization procedure are described in D5.1. Subsequently, the main transformation task is the semantic translation of the data in terms of HDOT supported by the Data Warehouse. For this task it is necessary that the extracted data are clearly annotated with semantic and syntactic information in order for a translation tool to perform the mappings of the fields and the data items to the p-medicine compliant attributes and concepts.

After the transformation process, the p-medicine platform will provide push-services, which support a data owner in the process of loading his data into the Data Warehouse. In terms of the origin of the data two kinds of scenarios will be supported:

1. **Pushing data from external data sources into the DW:**
   The p-medicine platform supports owners of external data sources to upload data into the data warehouse. A user-friendly “portlet” for the p-medicine portal will be implemented to support the upload-process as follows. When a user wishes to upload data from an external data source, it is firstly checked if there is already a source-specific ontology annotation available. If that is not the case the user is forwarded to the Ontology Annotator Tool that supports him in creating one as described above and storing it also for future use. As soon as this process is completed, he can extract selected data from his source and upload it. During this process the Data Translator is called and translates the data according to the appropriate ontology annotation into ontology triples compliant to the HDOT ontology, in order to integrate the data semantically.

2. **Pushing data from ObTiMA into the data warehouse:**
   It will furthermore be supported to upload data from ObTiMA into the Data Warehouse, which is a simpler case because ObTiMA is part of the p-medicine platform. Therefore, a user
interface will be provided in ObTiMA that allows data of a completed trial to be uploaded in the Data Warehouse. In this case the user only needs to select the appropriate data. Since the trial has been already annotated with HDOT during the design phase, the appropriate ontology annotation for the trial can be generated automatically. As in the last scenario, during the upload-process the data translator is called and translates the data according to the ontology annotation.

As a final note, the whole “push” process does not bear any “synchronization” semantics. The data owners and stewards are in full control to decide when to extract and upload their data and if this is done repeatedly or once, e.g. after the completion of a clinical trial. The issues of versioning and provenance of the data are of course important and it will be decided how to best support them by the semantic annotation tools and services.

7.2.2 Sync Services

The p-medicine platform will provide a sync service that allows to reuse data from hospital information systems in running clinical trials in the ontology-based trial management system ObTiMA. The sync service will retrieve the data not directly from the hospital information systems but from the p-medicine DW, since this data is already integrated compliant to HDOT. The HDOT compliance allows that the data can be mapped semi-automatically into the ontology-based CRFs of the clinical trial. We have chosen this approach since it does not require for each new trial a full manual mapping of the data models in the EHR onto the CRFs of the trial as in the approaches described in Sections 6.1 and 6.2. Furthermore, it does not put restrictions on the used hospital information systems, as e.g. compliance to standard data sets (cf. Chapter 6), that are in general not fulfilled by current systems.

In particular, the p-medicine sync service will support the following scenario with graphical user interfaces in ObTiMA. A clinician can select a patient in ObTiMA and can request to fill his CRFs with HIS data. The sync service will then retrieve all relevant HIS data for the patient from the data warehouse utilizing the DW SPARQL interface. The HIS data is then mapped automatically into appropriate items on the CRFs of the patient. This mapping is represented to the user. In this process it is assured that duplicate or outdated information in the DW will not be considered. Furthermore, the user can adjust the mapping to his needs and verify it. Finally, the data is filled automatically into the patient CRFs, after the mapping was confirmed by the clinician.

During this process only non-personal data can be filled into the CRFs since the data in the data warehouse is pseudonymized. Nevertheless, it is ensured by the p-medicine security infrastructure that appropriate non-personal data for patients in ObTiMA can be found in the DW, by linking the pseudonyms in ObTiMA and the DW (s. Deliverable D5.1).
### Appendix 1 - Abbreviations and acronyms

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABAP</td>
<td>Advanced Business Application Programming</td>
</tr>
<tr>
<td>ACC</td>
<td>American College of Cardiology</td>
</tr>
<tr>
<td>ACGT</td>
<td>Advancing Clinico Genomic Trials on Cancer</td>
</tr>
<tr>
<td>ACGT-MO</td>
<td>ACGT Master Ontology</td>
</tr>
<tr>
<td>ACR</td>
<td>American College of Radiology</td>
</tr>
<tr>
<td>ADL</td>
<td>Archetype Definition Language</td>
</tr>
<tr>
<td>ADT</td>
<td>Admission, Discharge, and Transfer</td>
</tr>
<tr>
<td>AHIC</td>
<td>American Health Information Community</td>
</tr>
<tr>
<td>AIDS</td>
<td>Acquired Immun Deficiency Syndrome</td>
</tr>
<tr>
<td>AM</td>
<td>Archetype Model</td>
</tr>
<tr>
<td>ANSI</td>
<td>American National Standards Institute</td>
</tr>
<tr>
<td>API</td>
<td>Application Programming Interface</td>
</tr>
<tr>
<td>ASTM</td>
<td>American Society for Testing and Material</td>
</tr>
<tr>
<td>BAPI</td>
<td>Business Application Programming Interfaces</td>
</tr>
<tr>
<td>BI</td>
<td>Business Intelligence</td>
</tr>
<tr>
<td>BRIDG</td>
<td>Biomedical Research Integrated Domain Group</td>
</tr>
<tr>
<td>C3PR</td>
<td>Central Clinical Participant Registry</td>
</tr>
<tr>
<td>caAers</td>
<td>Cancer Adverse Event Reporting System</td>
</tr>
<tr>
<td>caBIG</td>
<td>cancer Biomedical Informatics Grid</td>
</tr>
<tr>
<td>caDSR</td>
<td>Cancer Data Standards Registry and Repository</td>
</tr>
<tr>
<td>CCD</td>
<td>Continuity of Care Document</td>
</tr>
<tr>
<td>CCR</td>
<td>Continuity of Care Record</td>
</tr>
<tr>
<td>CDA</td>
<td>Clinical Document Architecture</td>
</tr>
<tr>
<td>CDA CN</td>
<td>CDA Clinical Note</td>
</tr>
<tr>
<td>CDISC</td>
<td>Clinical Data Interchange Standards Consortium</td>
</tr>
<tr>
<td>CDMS</td>
<td>Clinical Data Management Systems</td>
</tr>
<tr>
<td>CDR</td>
<td>Clinical Data Repository</td>
</tr>
<tr>
<td>CDW</td>
<td>Clinical Data Warehouse</td>
</tr>
<tr>
<td>CEN</td>
<td>Comité Européen de Normalisation</td>
</tr>
<tr>
<td>CEN/TC</td>
<td>CEN Technical Committee</td>
</tr>
<tr>
<td>CFR</td>
<td>Code of Federal Regulations</td>
</tr>
<tr>
<td>CHIME</td>
<td>Centre for Health Informatics &amp; Multiprofessional Education</td>
</tr>
<tr>
<td>COSTART</td>
<td>Coding Symbols for a Thesaurus of Adverse Reaction Terms</td>
</tr>
<tr>
<td>CPOE</td>
<td>Computerized physician order entry</td>
</tr>
<tr>
<td>CPT</td>
<td>Current Procedural Terminology</td>
</tr>
<tr>
<td>CRF</td>
<td>Case Report Form</td>
</tr>
<tr>
<td>CRO</td>
<td>Clinical Research Office</td>
</tr>
<tr>
<td>CSV</td>
<td>Comma-Separated Values</td>
</tr>
<tr>
<td>CTMS</td>
<td>Clinical Trial Management Systems</td>
</tr>
<tr>
<td>CTSA</td>
<td>Clinical and Translational Science Award) Informatics Program</td>
</tr>
<tr>
<td>DARTNet</td>
<td>Distributed Ambulatory Research in Therapeutics Network</td>
</tr>
<tr>
<td>DCI</td>
<td>Data Collection Instrument</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>DCM</td>
<td>Data Collection Module</td>
</tr>
<tr>
<td>DICOM</td>
<td>Digital Imaging and Communication in Medicine</td>
</tr>
<tr>
<td>DIMI</td>
<td>German Institute of Medical Documentation and Information</td>
</tr>
<tr>
<td>DRG</td>
<td>Diagnosis Related Groups</td>
</tr>
<tr>
<td>DT</td>
<td>Data Translator</td>
</tr>
<tr>
<td>DVG</td>
<td>Discrete Value Group</td>
</tr>
<tr>
<td>DW</td>
<td>Data Warehouse</td>
</tr>
<tr>
<td>EAV</td>
<td>Entity-Attribute-Value</td>
</tr>
<tr>
<td>ECRIN</td>
<td>European Clinical Research Infrastructure Network</td>
</tr>
<tr>
<td>EDC</td>
<td>Electronic Data Capture</td>
</tr>
<tr>
<td>EDIFACT</td>
<td>Electronic Data Interchange for Administration, Commerce and Transport</td>
</tr>
<tr>
<td>EFPIA</td>
<td>European Federation of Pharmaceutical Industries and Associations</td>
</tr>
<tr>
<td>EHR</td>
<td>Electronic Health Record</td>
</tr>
<tr>
<td>EHR/CR</td>
<td>Electronic Health Records/Clinical Research Working Group</td>
</tr>
<tr>
<td>EHR4CR</td>
<td>Electronic Health Records for Clinical Research</td>
</tr>
<tr>
<td>ECRIN</td>
<td>European Clinical Research Infrastructure Network</td>
</tr>
<tr>
<td>EMR</td>
<td>Electronic Medical Record</td>
</tr>
<tr>
<td>EORTC</td>
<td>European organisation for research and treatment of cancer</td>
</tr>
<tr>
<td>ePCRN</td>
<td>electronic Primary Care Research Network</td>
</tr>
<tr>
<td>ESB</td>
<td>Enterprise Service Bus</td>
</tr>
<tr>
<td>eSDI</td>
<td>Electronic Source Data Interchange</td>
</tr>
<tr>
<td>ETL</td>
<td>Extraction, Transform, and Load</td>
</tr>
<tr>
<td>EU</td>
<td>European Union</td>
</tr>
<tr>
<td>FAT</td>
<td>File Allocation Table</td>
</tr>
<tr>
<td>FDA</td>
<td>Food and Drug Administration</td>
</tr>
<tr>
<td>FGED</td>
<td>Functional Genomics Data</td>
</tr>
<tr>
<td>FSN</td>
<td>Fully Specified Name</td>
</tr>
<tr>
<td>FTP</td>
<td>File Transfer Protocol</td>
</tr>
<tr>
<td>GAARDS</td>
<td>Grid Authentication and Authorization with Reliably Distributed Services</td>
</tr>
<tr>
<td>G-DRG</td>
<td>German Diagnosis Related Groups</td>
</tr>
<tr>
<td>GE</td>
<td>General Electric</td>
</tr>
<tr>
<td>GO</td>
<td>Gene Ontology</td>
</tr>
<tr>
<td>GOOSE</td>
<td>GO Online SQL Environment</td>
</tr>
<tr>
<td>HARTS</td>
<td>Hoechst Adverse Reaction Terminology System</td>
</tr>
<tr>
<td>HDOT</td>
<td>Health Data Ontology Trunk</td>
</tr>
<tr>
<td>HIMSS</td>
<td>Healthcare Information and Management Systems Society</td>
</tr>
<tr>
<td>HIPAA</td>
<td>Health Insurance Portability and Accountability Act</td>
</tr>
<tr>
<td>HIS</td>
<td>Health Information System</td>
</tr>
<tr>
<td>HITSP</td>
<td>Healthcare Information Technology Standards Panel</td>
</tr>
<tr>
<td>HIV</td>
<td>Human Immunodeficiency Virus</td>
</tr>
<tr>
<td>HL7</td>
<td>Health-Level 7</td>
</tr>
<tr>
<td>HLT</td>
<td>High Level Term</td>
</tr>
<tr>
<td>I2b2</td>
<td>Informatics for Integrating Biology and the Bedside</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
</tr>
<tr>
<td>-----------</td>
<td>------------------------------------------------------------------</td>
</tr>
<tr>
<td>ICD</td>
<td>International Classification of Diseases</td>
</tr>
<tr>
<td>ICD-10-GM</td>
<td>International Classification of Diseases German Modification</td>
</tr>
<tr>
<td>ICD-O</td>
<td>International Classification of Diseases for Oncology</td>
</tr>
<tr>
<td>ICH</td>
<td>International Conference on Harmonisation</td>
</tr>
<tr>
<td>ICHI</td>
<td>International Classification of Health Intervention</td>
</tr>
<tr>
<td>ICPM</td>
<td>International Classification of Procedures in Medicine</td>
</tr>
<tr>
<td>ICSR</td>
<td>Individual Case Safety Record</td>
</tr>
<tr>
<td>ICT</td>
<td>Information and communications technology</td>
</tr>
<tr>
<td>IFPMA</td>
<td>International Federation of Pharmaceutical Manufacturers an Associations</td>
</tr>
<tr>
<td>IHE</td>
<td>Integrating the Healthcare Enterprise</td>
</tr>
<tr>
<td>IMI</td>
<td>Innovative Medicines Initiative</td>
</tr>
<tr>
<td>IS-H</td>
<td>Industrial Solutions Healthcare</td>
</tr>
<tr>
<td>ISO/EN</td>
<td>International Organization for Standardization/Europaeischen Normen</td>
</tr>
<tr>
<td>I-SPY 2 TRIAL</td>
<td>Investigation of Serial Studies to Predict Your Therapeutic Response with Imaging And MoLecular Analysis 2</td>
</tr>
<tr>
<td>IT</td>
<td>Information technology</td>
</tr>
<tr>
<td>IW</td>
<td>Information Warehouse</td>
</tr>
<tr>
<td>J-Art</td>
<td>Japanese Adverse Reaction Terminology</td>
</tr>
<tr>
<td>JBI</td>
<td>Java Business Integration</td>
</tr>
<tr>
<td>JEE</td>
<td>Java Platform, Enterprise Edition</td>
</tr>
<tr>
<td>JMS</td>
<td>Java Message Service</td>
</tr>
<tr>
<td>KHEntgG</td>
<td>Krankenhausentgeltgesetz</td>
</tr>
<tr>
<td>LIS</td>
<td>Laboratory Information System</td>
</tr>
<tr>
<td>LLT</td>
<td>Lowest Level Term</td>
</tr>
<tr>
<td>LOINC</td>
<td>Logical Observation Identifier Names and Codes</td>
</tr>
<tr>
<td>MAGE-TAB</td>
<td>Microarray Gene Expression Tabular</td>
</tr>
<tr>
<td>MCA</td>
<td>MCA - Medicines Control. Agency</td>
</tr>
<tr>
<td>MedDRA</td>
<td>Medical Dictionary for Regulatory Activities</td>
</tr>
<tr>
<td>MGED</td>
<td>Microarray Gene Expression Data</td>
</tr>
<tr>
<td>MIAME</td>
<td>Minimum Information About a Microarray Experiment</td>
</tr>
<tr>
<td>MPI</td>
<td>Master Person Index</td>
</tr>
<tr>
<td>NBIA</td>
<td>National Biomedical Imaging Archive</td>
</tr>
<tr>
<td>NCI</td>
<td>National Cancer Institute</td>
</tr>
<tr>
<td>NEMA</td>
<td>National Electrical Manufactures Association</td>
</tr>
<tr>
<td>NHS</td>
<td>National Health Service</td>
</tr>
<tr>
<td>NIH</td>
<td>National Institutes of Health</td>
</tr>
<tr>
<td>NOS</td>
<td>Not Otherwise Specified</td>
</tr>
<tr>
<td>NPfIT</td>
<td>National Programme for IT</td>
</tr>
<tr>
<td>OA</td>
<td>Ontology Annotator</td>
</tr>
<tr>
<td>OBO</td>
<td>Open Biological and Biomedical Ontologies</td>
</tr>
<tr>
<td>ObTiMA</td>
<td>Ontology-based Trilal Management Application</td>
</tr>
<tr>
<td>ODM</td>
<td>Operational Data Modelling</td>
</tr>
<tr>
<td>OLAP</td>
<td>Online Analytical Processing</td>
</tr>
<tr>
<td>OPS</td>
<td>Operationen- und Prozedurenschlüssel</td>
</tr>
<tr>
<td>OSUMC</td>
<td>Ohio State University Medical Center</td>
</tr>
<tr>
<td>------------</td>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>OWL</td>
<td>Web Ontology Language</td>
</tr>
<tr>
<td>PACS</td>
<td>Picture Archiving and Communication Systems</td>
</tr>
<tr>
<td>PHI</td>
<td>Protected Health Information</td>
</tr>
<tr>
<td>PHR</td>
<td>Personal Health Record</td>
</tr>
<tr>
<td>PhRMA</td>
<td>Pharmaceutical Research and Manufacturers of America</td>
</tr>
<tr>
<td>PT</td>
<td>Preferred Term</td>
</tr>
<tr>
<td>QG</td>
<td>Question group</td>
</tr>
<tr>
<td>R&amp;D</td>
<td>Research and Development</td>
</tr>
<tr>
<td>RCRIM TC</td>
<td>Regulated Clinical Research Information Management Technical Committee</td>
</tr>
<tr>
<td>RDF</td>
<td>Resource Description Framework</td>
</tr>
<tr>
<td>RELMA</td>
<td>Regenstrief LOINC Mapping Assistant</td>
</tr>
<tr>
<td>RFD</td>
<td>Retrieve Form for Data-capture</td>
</tr>
<tr>
<td>RIM</td>
<td>Reference Information Model</td>
</tr>
<tr>
<td>RIS</td>
<td>Radiology Information System</td>
</tr>
<tr>
<td>RM</td>
<td>Reference Model</td>
</tr>
<tr>
<td>RSNA</td>
<td>Radiological Society of North America</td>
</tr>
<tr>
<td>SAE</td>
<td>Serious Adverse Events</td>
</tr>
<tr>
<td>SAIF</td>
<td>Services-Aware Interoperability Framework</td>
</tr>
<tr>
<td>SHARE</td>
<td>Shared Health and Research Electronic</td>
</tr>
<tr>
<td>SIG</td>
<td>Special Interest Group</td>
</tr>
<tr>
<td>SL</td>
<td>Semantic Layer</td>
</tr>
<tr>
<td>SM</td>
<td>Service Model</td>
</tr>
<tr>
<td>SME</td>
<td>Subject Matter Expert</td>
</tr>
<tr>
<td>SMQs</td>
<td>Standardised MedDRA Queries</td>
</tr>
<tr>
<td>SNOMED-CT</td>
<td>Systematized Nomenclature of Medicine -- Clinical Terms</td>
</tr>
<tr>
<td>SOC</td>
<td>System Organ Class</td>
</tr>
<tr>
<td>SPARQL</td>
<td>SPARQL Protocol And RDF Query Language</td>
</tr>
<tr>
<td>SQL</td>
<td>Structured Query Language</td>
</tr>
<tr>
<td>SSL</td>
<td>Secure Socket Layer</td>
</tr>
<tr>
<td>STRIDE</td>
<td>Stanford Translational Research Integrated Database Environment</td>
</tr>
<tr>
<td>SUMC</td>
<td>Stanford University Medical Center</td>
</tr>
<tr>
<td>TC</td>
<td>Technical Committee</td>
</tr>
<tr>
<td>TCP-IP</td>
<td>Transmission Control Protocol / Internet Protocol</td>
</tr>
<tr>
<td>TMS</td>
<td>Thesaurus Management System</td>
</tr>
<tr>
<td>TRANSCEND</td>
<td>TRANslation Informatics System to Coordinate Emerging Biomarkers, Novel Agents, and Clinical Data</td>
</tr>
<tr>
<td>UML</td>
<td>Unified Modelling Language</td>
</tr>
<tr>
<td>WHO</td>
<td>World Health Organisation</td>
</tr>
<tr>
<td>WHO-ART</td>
<td>WHO Adverse Reactions Terminology</td>
</tr>
<tr>
<td>XML</td>
<td>eXtensible Markup Language</td>
</tr>
<tr>
<td>ZIK</td>
<td>Zentrum für Informations- und Kommunikationstechnik</td>
</tr>
</tbody>
</table>