eHealth2014 – Health Informatics Meets eHealth
A. Hörbst et al. (Eds.)

doi:10.3233/978-1-61499-397-1-40

A Reference Architecture for Semantic Interoperability and Its Practical Application

Christian ZUNNER[a, 1], Thomas GANSLANDT[b] Hans-Ulrich PROKOSCH[a, b] and Thomas BÜRKLE[a]
[a] *Chair of Medical Informatics, University of Erlangen-Nuremberg*
[b] *Medical Center for Information and Communication, Erlangen University Hospital, Erlangen, Germany*

Abstract. Objective: Reusing EPR data for secondary purposes often requires mapping to classifications and vocabularies such as ICD, LOINC or NCI thesaurus. We aimed for a common architecture which supports the use of different vocabularies and mapping tools. Methods: We integrated the components clinical data warehouse, vocabulary resources and mapping tools with the EPR and client applications. Results: In two projects we used this architecture to map laboratory parameters from the LIS to LOINC, and to map clinical data elements from the Soarian EPR to the cancer registry system using the NCI-Thesaurus®. Conclusion: The approach was successful in both projects. The reference architecture does not resolve the mapping task, but provides reusable integration links between the different components and thus facilitates further mapping activities.

Keywords. Controlled Vocabulary, Documentation, Semantics.

1. Introduction

Today, the focus of medical information processing shifts from system implementation and digital documentation towards reuse of information from the multitude of stored patient data inside the electronic patient record EPR [1, 2]. EPR data comprising free text, structured data fields and some percentage of coded data, is mostly stored in non-standardized format within different clinical information systems. A clinical data warehouse may combine the data from different sources but it does not solve the problem of semantic interoperability [3]. Therefore the use of classifications and controlled vocabularies is an important step in the composition of a single source approach [4]. Unfortunately, there are many classifications and vocabularies with different focus and philosophy such as ICD for diseases, LOINC for laboratory results or ATC for drug substances. There is yet no practical guideline how a single source platform can be implemented on this basis [5]. Cimino [6] determined that the challenges for the realization of a single source approach are the availability of suitable vocabulary, the coding of the data and the adoption of standards for representation. Controlled vocabularies are available, but the coding or mapping to such vocabularies

[1] christian.zunner@med.stud.uni-erlangen.de

remains difficult and laborious and terminological implementation practices are non-existent [3, 4, 5].

Erlangen University Hospital (EUH) is active in several single source projects [7, 8, 9]. Semantic mapping of different sources proved to be a tedious and repeating task, prompting the question if a generic framework-like concept for the mapping of clinical EPR data to different existing controlled vocabularies could be devised. Thus, our research objective was a generic component based reference architecture which enables semantic mapping of various EPR elements to different catalogues and vocabularies. Such a reference architecture should facilitate the implementation of ontological representations and thus support semantic interoperability.

2. Methods

EUH is a maximum care facility with 1316 beds situated in the south of Germany near Nuremberg. Patient treatment takes place in different highly specialized departments and functional units. EUH uses the Soarian® Clinicals electronic patient record system by Siemens Inc. which is interfaced with many other specialized information systems for laboratories, radiology, surgical theatre, and functional units. Soarian supports development of customized documentation forms with free text or structured items [10]. A Cognos®-based clinical data warehouse (DWH) system has been implemented to collect data from these various subsystems, a research data warehouse using i2b2 [11, 12] is currently being established. The i2b2 platform (Informatics for Integrating Biology and the Bedside) is an open-source platform for intuitively querying large biomedical datasets, based on a generic entity-attribute-value data model. Laboratory services are performed in different laboratories, with some overlap in the diagnostic spectrum. All laboratories use the Swisslab® commercial laboratory information system (LIS), with disparate but non-unified codes for all observations. Laboratory data are transferred to the clinical workstation via HL7 interface.

Incrementally, the infrastructure for data processing in EUH has been enhanced with a reference architecture for semantic annotation of EPR data using different controlled vocabularies. Figure 1 shows the top level architecture.

The following components and their functional interfaces constitute this architecture:

First a common database for analyzing and querying EPR data was required. This task is accomplished within the Cognos®-based clinical data warehouse which reads raw EPR data from most relevant source systems on a daily basis.

To assist semantic mapping of the EPR data, the respective vocabularies must be made available for digital processing within the component “vocabulary resources” on the right hand side. Vocabulary resources comprise several components which currently enable mapping to either NCI thesaurus [13] or LOINC [14]. More components will be added in our continuing semantic mapping efforts. Essentially, vocabulary resources store the controlled vocabularies content, the mapping tables and vocabulary services such as LexEVS.

To support the actual mapping between EPR data and the vocabulary sources a set of active components is required marked “Mapping tools” in Fig. 1. Similar to vocabulary resources, mapping tools, which currently comprise mechanisms for the mapping to LOINC (RELMA) and to NCI Thesaurus (Metamap) will be complemented with additional tools.

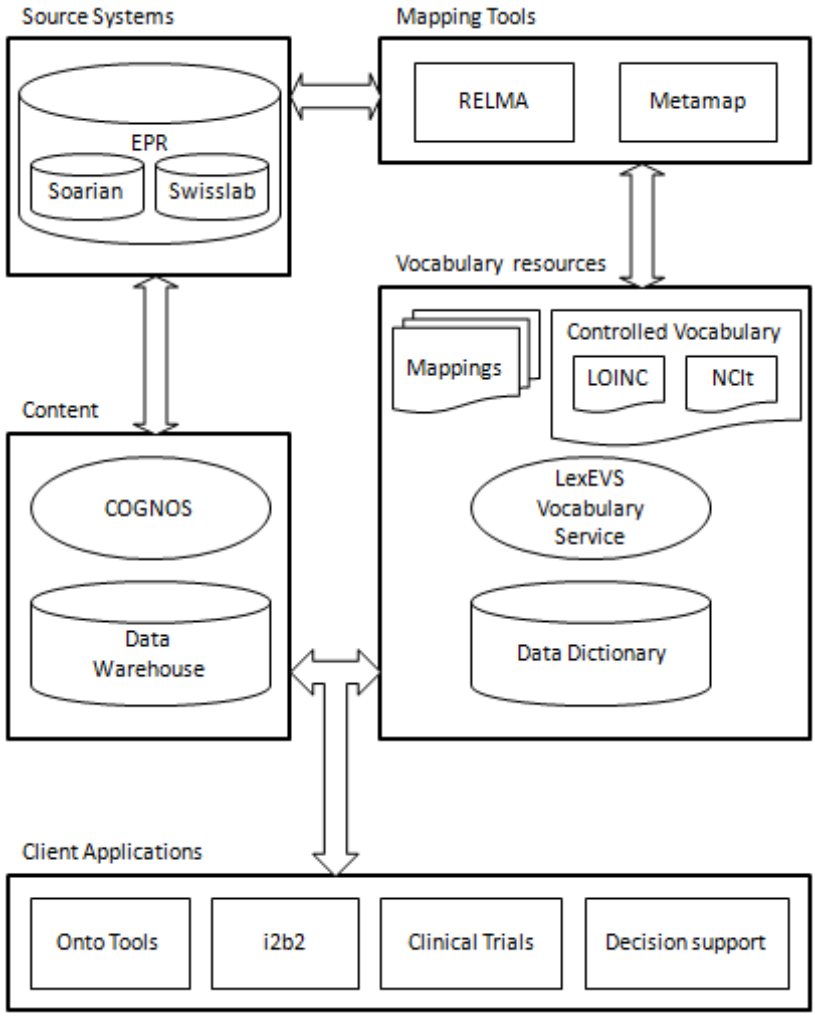

Figure 1. Reference architecture for semantic annotation in a single source environment. The boxes show the components of the framework, the arrows represent the implemented data loading and querying directions.

To harmonize unstandardized raw data with the available mappings we use Onto-Tools [9], a locally developed tool suite and support e.g. the transfer of LexEVS® mappings into secondary systems such as the research data warehouse i2b2®. The Onto-Tools [9] are a suite of tools for creating automated ETL-jobs (extraction - transformation - loading) based on mapping ontologies to load data from a source system into i2b2.

An essential feature within this component architecture is that the linkage between the components (the arrows) remains standardized and should need no or only minimal adaptions if further mappings are being implemented.

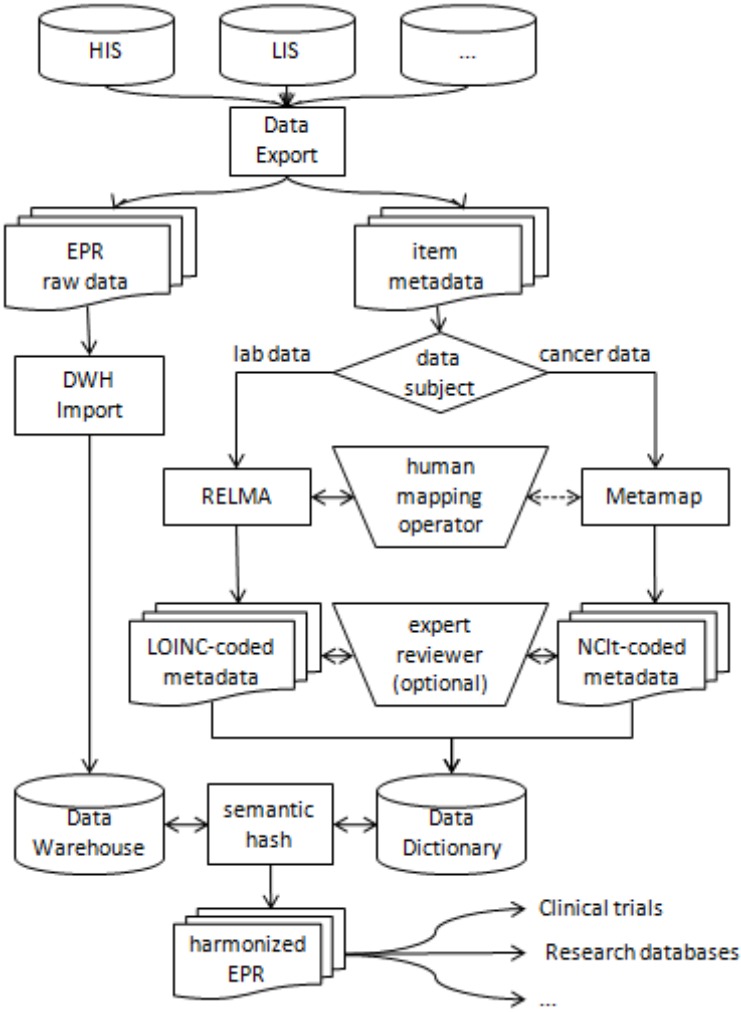

Figure 2. Workflows for LOINC and cancer data mapping within the reference architecture. Rectangles represent (existing) ETL and mapping tools, trapezoid shapes represent human mapping efforts. The reference architecture is characterized by the automated data flow which has been symbolized with arrows.

3. Results

The reference architecture has been prototypically implemented to support mapping of local laboratory interface terms to LOINC [15] and to map Soarian based cancer documentation data fields from the EPR to the GTDS tumor registry application used at Erlangen cancer center [16] via the NCI-Thesaurus. An Oracle database was used to implement the vocabulary resources. Interconnection of mapping tools and the linking components to the EPR have been programmed in Java Version 6 on a workstation running Windows 7.

Figure 2 demonstrates the workflows to use the reference architecture for the two pilot mapping projects. On the left the typical data warehouse ETL (extraction –

transformation – loading) process to load raw clinical EPR data from various sources into the DWH is depicted. On the right, the LOINC mapping pathway uses the RELMA tool from the mapping tools for semi-automated mapping (mapping proposals are displayed to a human mapper for approval), whereas in the cancer data mapping process an automated mapping with the Metamap tool takes place. Both pathways are enabled for optional expert review to ensure appropriate mapping quality.

In the LOINC mapping project [15] a total of 10.206 laboratory interface terms have been mapped to 2.564 LOINC codes. Quality controlled expert review has been performed for a sample of 100 terms. In the cancer data mapping project [16] we mapped a total of 1.142 clinical data items from the Soarian EPR system to GTDS cancer registry documentation system using the NCI-Thesaurus. These mappings are used for pooling and querying laboratory data respectively for interfacing the Soarian EPR cancer data items to the GTDS cancer registry system.

The correctness of the LOINC®-mapping using RELMA® was higher (98%) than the correctness of the NCI-Thesaurus®-mapping using Metamap® (79%). This is possibly due to the fact that RELMA® is a semiautomated mapping tool which requires a human user for mapping, whereas Metamap supports fully automated mapping. Finally our Onto-Tools [9] provide the possibility to semantically rehash raw data using mapping ontologies and load it into secondary systems.

4. Discussion

We present a proposal for a reference architecture to support semantic mapping of EPR data items to different vocabularies. The architecture comprises a clinical data warehouse, a vocabulary resources component and a mapping tools component plus client applications with their respective logic interconnections. In this scenario mapping tools include manual, semi-automated and fully automated mapping approaches. The prototypical implementation of this architecture has proven suitable for two mapping activities with several thousand data items of the clinical EPR.

Semantic mapping is not something which comes without effort. According to our results, fully automated mapping may still result in too many errors, thus requiring manual intervention and trained human mappers. Furthermore, expert review of the mapping results is a tedious but essential task in the mapping process to ensure reliability of the results. Our lessons learnt in the project are that semantic mapping can be facilitated when the source meta data are in most parts complete and understandable, which is not always the case. If meta-information is missing - we experienced e.g. initially missing information which methods are used for certain laboratory parameters – it may be impossible to assign the correct mapping to LOINC codes. In the given example this resulted in tedious communication with the consulting experts from the respective laboratories. Success of semantic mapping depends on the application area, which is reflected in the achieved mapping quality – see results section: Structured laboratory parameters are better suited than semi-structured data fields for cancer documentation. For the latter we noticed also that a true gold standard for cancer documentation does not exist. The more complex and unstructured the terminology of the application area, the more essential it is to rely on experienced human mapping operators with an in depth knowledge of the application area to achieve acceptable results. Consequently, good cooperation with the involved clinical departments is absolutely necessary.

The advantage of the reference architecture approach however is that the linkage components between mapping tools, vocabulary resources, the EPR sources and the clinical data warehouse can be reused, requiring only minor adaptations when a new mapping task needs to be performed. Most components except COGNOS® by IBM Inc. are available for free for academic use. The proposed reference architecture can be transferred to another environment subject to the following conditions:

1. a clinical data warehouse is required as the basis for development
2. a data export from the EPR for raw data and meta data is available
3. a component "vocabulary resources" must be established which enables storage of controlled vocabularies and of mapping results
4. a "mapping tools component" must be implemented which comprises the required mapping tools and permits dynamic integration of further mapping services
5. these components must be integrated with each other in a generic way so that future mapping tasks may be accomplished with minimal alterations in the integration components.

Former single source projects such as [17] just used locally developed standards to connect their system instead of standardized terminology, whereas newer projects such as [18, 19] use ontologies together with a controlled vocabulary. In comparison, our approach focuses on the facilitation of the mapping and on the implementation of an ontological representation of the semantic annotation for use by client applications. This has been described as an important prerequisite by Cimino [6] who determined that the main challenges for the realization of a single source approach are the availability of suitable vocabulary, the coding of the data and the adoption of standards for representation.

An essential prerequisite for semantic mapping efforts of the EPR data is the existence of structured items. Ontological approaches focusing on Vocabulary Servers or Medical Data Dictionaries cannot work when large parts of the EPR are unstructured free text. Our approach works well for items with predefined or discrete values. In principle, free text items can be mapped to concepts as well, but only if the EPR stored information has been analyzed for its content using e.g. text mining approaches.

References

[1] Herzberg S, Dugas M. Single source information systems can improve data completeness in clinical studies: an example from nuclear medicine. Stud Health Technol Inform. 2011;169:872e6.

[2] Berges I, Bermudez J, Illarramendi A. Towards Semantic Interoperability of Electronic Health Records. IEEE Trans Inf Technol Biomed 2012 May;16(3):424-31.

[3] McDonald CJ. The barriers to electronic medical record systems and how to overcome them. J Am Med Inform Assoc. 1997;4:213e21.

[4] Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med 1998;37:394e403.

[5] Tao C et al. Terminology representation guidelines for biomedical ontologies in the semantic web notations. J Biomed Inform (2012), http://dx.doi.org/10.1016/j.jbi.2012.09.003

[6] Cimino, James J. "Collect Once, Use Many: Enabling the Reuse of Clinical Data through Controlled Terminologies." Journal of AHIMA 78, no.2 (February 2007): 24-29.

[7] Prokosch HU, Ries M, Beyer A, Schwenk M, Seggewies C, Köpcke F, Mate S, Martin M, Bärthlein B, Beckmann MW, Stürzl M, Croner R, Wullich B, Ganslandt T, Bürkle T. IT infrastructure components to support clinical care and translational research projects in a comprehensive cancer center. Stud Health Technol Inform. 2011;169:892-6.

[8] Köpcke F, Kraus S, Scholler A, Nau C, Schüttler J, Prokosch HU, Ganslandt T.Secondary use of routinely collected patient data in a clinical trial: An evaluation of the effects on patient recruitment and data acquisition. Int J Med Inform. 2013 Mar;82(3):185-92.

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

[10] Haux R, Seggewies C, et al. Soarian - workflow management applied for health care. Methods Inf Med. 2003;42(1):25-36.

[11] Murphy SN, Weber G, Mendis M, Gainer V, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med InformAssoc. 2010; 17(2):124-130

[12] Ganslandt T, Mate S, Helbing K, Sax U, Prokosch HU. Unlocking Data for Clinical Research - The German i2b2 Experience. Appl Clin Inf. 2011; 2(1):116–127.

[13] http://ncit.nci.nih.gov/ (accesed 27th January 2014)

[14] http://loinc.org/ (accessed 27th January 2014)

[15] Zunner C, Bürkle T, Prokosch H-U, Ganslandt T. Mapping local laboratory interface terms to LOINC at a German university hospital using RELMA V.5: a semi-automated approach. J Am Med Inform Assoc. 2013 Mar-Apr;20(2):293-7. doi: 10.1136/amiajnl-2012-001063. Epub 2012 Jul 16.

[16] Zunner C, Bürkle T, Prokosch H-U, Ganslandt T. 56. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (gmds), 6. Jahrestagung der Deutschen Gesellschaft für Epidemiologie (DGEpi). Mainz, 26.-29.09.2011. Düsseldorf: German Medical Science GMS Publishing House. DOI: 10.3205/11gmds444 (published 20 September 2011)

[17] Rebecca Kush, Liora Alschuler, Roberto Ruggeri, et al. Implementing Single Source: The STARBRITE Proof-of-Concept Study. J Am Med Inform Assoc. 2007;14:662– 673. DOI 10.1197/jamia.M2157.

[18] El Fadly A, Rance B, Lucas N, Mead C, Chatellier G, Lastic PY, Jaulent MC, Daniel C. Integrating clinical research with the Healthcare Enterprise: from the RE-USE project to the EHR4CR platform. J Biomed Inform. 2011 Dec;44 Suppl 1:S94-102. doi: 10.1016/j.jbi.2011.07.007. Epub 2011 Aug 25.

[19] Doods J, Botteri F, Dugas M, Fritz F; EHR4CR WP7. A European inventory of common electronic health record data elements for clinical trial feasibility. Trials. 2014 Jan 10;15(1):18. doi: 10.1186/1745-6215-15-18.