ICT for Health Science Research
A. Shabo (Shvo) et al. (Eds.)

doi:10.3233/978-1-61499-959-1-85

Approaching Clinical Data Transformation from Disparate Healthcare IT Systems Through a Modular Framework

Lo An PHAN-VOGTMANN [a,1,2], Alexander HELHORN [b,2], Henner M. KRUSE [b,2],
Eric THOMAS [b], Andrew J. HEIDEL [b], Kutaiba SALEH [b], Florian RISSNER [c],
Martin SPECHT [d], Andreas HENKEL [d], André SCHERAG [a] and Danny AMMON [b]

[a] *Institute of Medical Statistics, Computer and Data Sciences (IMSID), Jena University Hospital, Jena, Germany*
[b] *Data Integration Center, IT Department, Jena University Hospital, Jena, Germany*
[c] *Center for Clinical Studies, Jena University Hospital, Jena, Germany*
[d] *IT Department, Jena University Hospital, Jena, Germany*

Abstract. Many healthcare IT systems in Germany are unable to interoperate with other systems through standardised data formats. Therefore it is difficult to store and retrieve data and to establish a systematic collection of data with provenance across systems and even healthcare institutions. We outline the concept for a Transformation Pipeline that can act as a processor for proprietary medical data formats from multiple sources. Through a modular construction, the pipeline relies on different data extraction and data enrichment modules as well as on interfaces to external definitions for interoperability standards. The developed solution is extendable and reusable, enabling data transformation independent from current format definitions and entailing the opportunity of collaboration with other research groups.

Keywords. Electronic Health Records, Healthcare IT, Interoperability Standards, Data Transformation, Metadata

1. Introduction

The lack of interoperable data models in the predominantly proprietary data repositories of healthcare IT systems and electronic health records requires flexible data acquisition and conversion processes for data transfers and secondary uses [1]. Although most hospitals run a communication server to allow messaging and data conversion between IT systems, some of the deployed IT products do not support the desired data formats or

[1] Corresponding Author: Lo An Phan-Vogtmann, Institute of Medical Statistics, Computer and Data Sciences (IMSID), Jena University Hospital, Bachstraße 18, 07743 Jena, Germany; E-mail: loan.phan-vogtmann@med.uni-jena.de; Phone: +49 3641 9 398352.

[2] The authors contributed equally to this work.

there are obstacles in using those systems within a different context, e.g. for clinical research. For cross-institutional communication and secondary use of routinely collected clinical data, it is desirable to have standardised data integration processes throughout all the affiliated healthcare institutions. An example of the demand for such data integration processes appeared in the form of the German Medical Informatics Initiative, where health care data shall be made available for clinical research or health services research purposes [2]. Starting from university clinics within funded consortia, other partners in the health care system are to be gradually included as well. The "Smart Medical Information Technology for Healthcare" (SMITH) consortium is especially targeting the use of healthcare interoperability standards for data sharing between its members [3].

In this paper, we outline a common model for the development of a data Transformation Pipeline component, which implements syntactic, semantic interoperability and integrated data curation processes. This component is a proposal – developed at the Jena University Hospital – for a concrete implementation of the ideas and concepts developed jointly within SMITH. The paper should intensify a discussion on how to implement data retrieval and transformation processes with the Data Integration Centers (DICs) of the German Medical Informatics Initiative.

2. Methods

From the perspective of clinical data integration and consolidation, several operations are required for a pipeline transformation. First, the Transformation Pipeline extracts the data from the source system (e.g. Laboratory Information System). To achieve this, a connector must be used or implemented for each IT source system. In the next step, data is converted into one consolidated format which is then ready for data transformation.

The data transformation process needs to integrate technical standards and terminologies from external definitions and integrates process and provenance definitions as well. Transformation results must then be stored in an interoperable format such as a HL7 CDA document [4] in an IHE XDS Registry and Repository or as HL7 FHIR [5] resources for discrete data.

Thus, an important function in the transformation process is the conversion of stored data (e.g. a single bit representation of the patient's gender) into interoperable formats. In these formats defined syntax (e.g. the FHIR attribute definition for Patient.gender), semantics (e.g. the HL7 Value Set for AdministrativeGender) and provenance (e.g. a predefined OID [6] for the IT system as original data source) must be included, but each can be subject to revisions by Standards Developing Organizations (SDOs). Therefore, there is a need to include technical representations of these definitions over an interface which itself is not subject to these revisions. Currently, in most data extraction and transformation processes, certain versions of standardized or non-standardized formats are created directly (e.g. FHIR Release 3 (STU)), and with updates, the complete transformation must be implemented anew.

Since various healthcare IT systems can act as a data source for the pipeline, there is a strong need to support many different data formats and communication protocols. In order to fulfil this need and to integrate external definitions determining the outputs, we developed a transformation approach with fine-grained building blocks. Each in- / output protocol and format, as well as syntax, semantics and process / provenance definitions are represented by a single configurable module. The order of these modules can vary (see Figure 1).

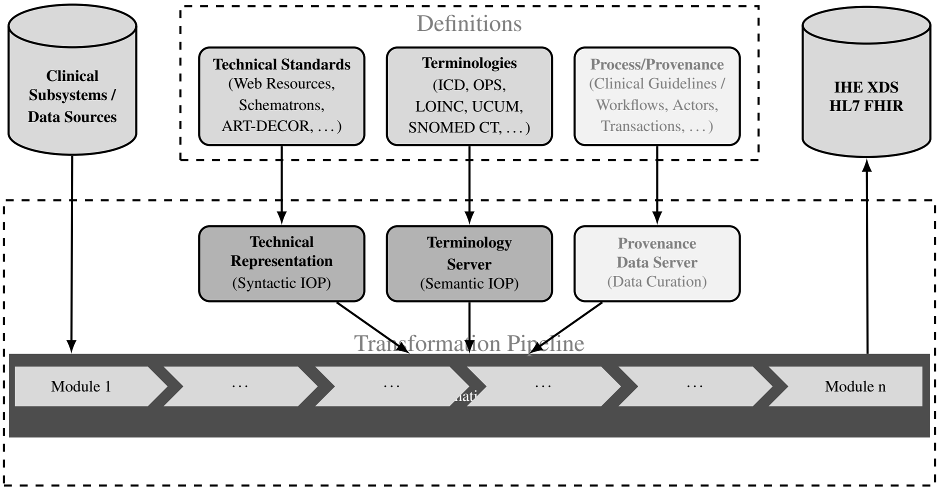
Definitions
Clinical Subsystems / Data Sources
Technical Standards (Web Resources, Schematrons, ART-DECOR, . . .)
Terminologies (ICD, OPS, LOINC, UCUM, SNOMED CT, . . .)
Process/Provenance (Clinical Guidelines / Workflows, Actors, Transactions, . . .)
IHE XDS HL7 FHIR
Technical Representation (Syntactic IOP)
Terminology Server (Semantic IOP)
Provenance Data Server (Data Curation)
Transformation Pipeline
Module 1
· · ·
· · ·
· · ·
· · ·
Module n

Figure 1. Context of the Transformation Pipeline

An object-oriented approach with Java as programming language was chosen for the prototypical implementation. To structure the source code, packages were created according to the pattern `LAYER.FUNCTIONAL_OBJECT`. `LAYER` is the first abstraction and represents the domain the source interacts with (i.e. configuration or extraction). A `FUNCTIONAL_OBJECT` restricts the domain (i.e database configuration or log configuration). The corresponding Java classes (modules) are created in the third level. The meaningful name of a class specifies the concrete function to this class, so that, for example the path for the data extractor of a patient data management system (PDMS) would be: `Extractor.Database.PDMSExtractor`. To avoid naming conflicts the structure described should also be in a Java package. The following prefix was selected for the prototype implementation within the SMITH project: `care.smith.dic.die`. The resulting fully qualified class name (FQCN) for the class mentioned above is accordingly `care.smith.dic.die.Extractor.Database.PDMSExtractor`.

The Transformation Pipeline executes conversions of medical data from sources into formats based on standardised information models. In addition to typical extracting, transforming and loading (ETL), our focus lies on the development of modules through which we are able to access most recent syntactical, semantical, and processual definitions to support a data preparation process and to enrich extracted data with necessary metadata without directly including these definitions in the transformation code.

3. Results

Using the described concept, a prototypical implementation of the Transformation Pipeline has been developed at Jena University Hospital. In this setup, a PDMS is used as a first clinical data source. The Transformation Pipeline gathers a patient's vital data and transforms them into HL7 FHIR resources loaded to a local FHIR Server. The initial state of the PDMS consists of a SQL database, which contains required clinical data and the PDMS currently features no interface to extract the stored information.

The final FHIR resource is defined by a profile. We used the official profiles in FHIR Release 3 (STU). The profiling is part of the definition of technical standards and can be changed and adjusted for any kind of data projection. The correct resource format will be validated by the FHIR server. The transformation algorithm uses methods defined by the HAPI FHIR API [7] in a special transformation module connecting that API. Combining the technical standards and the API methods from HAPI FHIR allows the system to provide FHIR resources that are always compliant to the desired formatting.

To provide further information about the transformed data, semantic coding is added by using an external terminology server. The correct semantic annotation can be added by either providing the specific terminology code or by searching the server entries using keywords and certain circumstances. A CTS2 conform terminology server was used in this prototype implementation, which feeds LOINC codes into the Transformation Engine. The terminology codes can be added to every transformed set of data.

Using this method, the system was able to create FHIR observation resources for some of the most important vital parameters of a patient. In the current implementation these include: body temperature, blood pressure, heart frequency, and oxygen saturation.

4. Discussion

With the first examples of vital parameter transformations, we were able to show that our structural concept can be implemented. Additionally, through standardised data formats, secondary use of healthcare data in research projects like SMITH is possible.

Each of the partner hospitals within the SMITH consortium is currently in the process of establishing a Data Integration Center, where routinely collected clinical data from within a hospital will be prepared for more comprehensive research projects. Each Data Integration Center will establish its own Transformation Pipeline, since each hospitals' IT landscape differs. This is where the modular concept of our approach should fit, as the modularity should support a fast development of new components. Each consortium member is able to develop the components according to their own needs and ideally shares them to foster wider useage.

The extensive reusability through the modular approach allows later incorporation into existing tools. Frameworks for data integration such as the Integrated Data Repository Toolkit (IDRT) [9] and tools including ETL services or communication servers, are considered for future review and use within our project. Additionally we intend to enhance the Transformation Pipeline with modules for data extraction from free text [10]. However we do not commit to incorporating these tools initially.

At the current stage of development the Transformation Pipeline does not contain a fixed concept for the inclusion of process and provenance definitions in order to curate data. For example, how artefacts such as measurement errors should be eliminated. In the future, it will be necessary to determine which additional sources (in addition to the data source itself) are necessary to obtain such definitions. In the USA, the eMERGE Network [8] is an example of such an endeavor which is currently not available in Germany. Here, additional work is necessary that includes a detailed review of clinical guidelines, workflows and transactions and close cooperation with the involved clinicians and other health professionals.

5. Conclusion

Most routinely collected medical data in Germany's healthcare is still stored in proprietary data repositories. Hence, in this paper we present a way to transfer data from a healthcare IT system to a standardised data format using the Transformation Pipeline.

Modules of the Transformation Pipeline apply definitions for technical standards and terminologies for semantic interoperability. There are additional modules planned for further data extraction and to curate data with process and provenance definitions.

A prototypical conduction of the transformation process has been performed. The pipeline transformed data from the local PDMS into HL7 FHIR resources. Furthermore, the Transformation Engine adds LOINC codes by retrieving these from an external terminology server based on the designation of the extracted data.

An increased use of transformation frameworks like the one outlined here, requires extensive collaboration between medical societies, SDOs and medical informatics specialists in healthcare IT departments. This cooperation aims at refining information models for interoperability standards to be applied for data storage and transfer in healthcare and clinical research but should also contribute to many practical improvements of digital medical documentation of both routinely collected medical data as well as medical data that is collected explicitly for clinical studies within clinical data management systems.

Acknowledgement

This work was supported by German Ministry of Education and Research (BMBF No. 01ZZ1803C) within the project SMITH.

References

[1] Blobel, B. (2018). Interoperable EHR Systems – Challenges, Standards and Solutions. *EJBI*, **14(2)**, 10–19.

[2] Gehring S, Eulenfeld R (2018) German Medical Informatics Initiative: Unlocking Data for Research and Health Care. *Methods Inf Med* 2018; **57(S 01)**, e46–e49.

[3] Winter A, Stäubert S, Ammon D et al. (2018) Smart Medical Information Technology for Healthcare: Data Integration based on Interoperability Standards. *Methods Inf Med* 2018; **57(S 01)**, e92–e105.

[4] Health Level Seven International. CDA© Release 2. http://www.hl7.org/implement/standards/product_brief.cfm?product_id=7 [cited 2018 Oct 31].

[5] HL7 FHIR Resourcelist. http://hl7.org/fhir/resourcelist.html [cited 2018 Oct 31]

[6] HL7 Implementation Guidance for Unique Object Identifiers (OIDs), Release 1, 2011. http://www.hl7.org/implement/standards/product_brief.cfm?product_id=210 [cited 2018 Oct 31]

[7] HAPI FHIR Documentation. http://hapifhir.io/apidocs-dstu3 [cited 2018 Oct 31]

[8] Gottesman O, Kuivaniemi H, Tromp G et al. (2013) The Electronic Medical Records and Genomics (eMERGE) Network: Past, Present, and Future. *Genet Med* 2013;**15(10)**, 761–71.

[9] Baum B, Christoph J, Engel I et al. (2016) Integrated Data Repository Toolkit (IDRT). *Methods Inf Med* 2016; **55(2)**, 125–35.

[10] Wang Y, Wang L, Rastegar-Mojarad M et al. (2018) Clinical information extraction applications: A literature review. *Journal of Biom Inf* 2018; **77**, 34–49.