TermiCron – Bridging the Gap between FHIR Terminology Servers and Metadata Repositories

Joshua Wiedekopf [a], Hannes Ulrich [a], Cora Drenkhahn [a], Ann-Kristin Kock-Schoppenhauer [a], Josef Ingenerf [a,b]

[a] IT Center for Clinical Research (ITCR-L), University of Lübeck, Germany
[b] Institute of Medical Informatics, University of Lübeck, Germany

Abstract

The large variability of data models, specifications, and interpretations of data elements is particular to the healthcare domain. Achieving semantic interoperability is the first step to enable reuse of healthcare data. To ensure interoperability, metadata repositories (MDR) are increasingly used to manage data elements on a structural level, while terminology servers (TS) manage the ontologies, terminologies, coding systems and value sets on a semantic level. In practice, however, this strict separation is not always followed; instead, semantical information is stored and maintained directly in the MDR, as a link between both system is missing. This may be reasonable up to a certain level of complexity, but it quickly reaches its limitations with increasing complexity. The goal of this approach is to combine both components in a compatible manner. We present TermiCron, a synchronization engine that provides synchronized value sets from TS in MDRs, including versioning and annotations. Prototypical results were shown for the terminology server Ontoserver and two established MDR systems. Bridging the semantic and structural gap between the two infrastructure components, this approach enables shared use of metadata and reuse of corresponding health information by establishing a clear separation of the two systems and thus serves to strengthen reuse as well as to increase quality.

Keywords:

Metadata Repository, HL7 FHIR, Secondary Use, Terminology Servers

Introduction

The increasing proliferation of electronic health records and mobile applications is driving a rapid growth in clinical data. This development offers great potential for secondary use of clinical data, which is essential to improve healthcare, reduce healthcare costs, and enhance clinical research [1]. The field of healthcare is broad and comprehensive, encompassing many different areas with respective specific characteristics, naming and requirements. This reflects in the diversity of data models, value sets and interpretations of data elements. To enable these diverse data sets for secondary use results in the urgent need of automatic understanding of previously unknown information and structures to integrate and enable this information for further (re)use [2].

To gain semantic interoperability of the distributed software systems involved, the reuse and joint evaluation of distributed data and use of the semantic associations and relationships among shared entities is essential [3]. Among many other informatics approaches, two infrastructure components play an important role in this environment, one being metadata repositories (MDR) and the other terminology servers (TS). According to the authoritative definitions, an MDR manages elementary and complex data elements in terms of metadata [4]. The data elements in an MDR provide those metadata at the schema level (e.g., attributes, data types) to allow the instance data described by the metadata to be interpreted in the required context (e.g., address schemas versus real addresses). A TS is a server specialized in terminologies, coding systems and value sets, which provides the complex content interactively and primarily via standardized interfaces. Both infrastructure components partly overlap in their functionality in real implementations, but MDR systems are supposed to be one of many application systems that can make use of terminology services. Structurally simple terminologies and especially value sets can in principle be offered integrated within the MDR systems, which also explains the partial overlap of both functionalities in actual MDR implementations. The MDR requires value sets to adequately describe the data structures. Especially simple sets of terms are provided in value lists like gender (*m*, *w*, *u*) and are quite easy to handle without assignment to concepts, with the limitations of not being able to interprete them in a comprehensive way. The real benefit of a terminology server arises with the quantitative and qualitative complexity of terminologies and value sets.

The management of conceptual terminologies with partly complex internal data structures, their versioning as well as mappings between terminologies/value sets is a task not in the scope of MDR systems. Outsourcing the terminological information to a TS also serves the purpose of reusing the data, which may include merging different primary sources with potentially different MDR systems. In practice, the situation is unfortunately different, as all possible information is documented directly in the MDR, due to the fact of a missing link between the TS and the MDR system.

The aim of this study is to close the gap between the two systems, TS and MDR, in order to establish a clear

separation of the two systems and thus to strengthen reuse as well as to increase quality.

Methods

The healthcare system is driven by communication standards which shall foster the exchange and the cooperation of all parties involved. HL7 FHIR has been introduced in 2014 [5, 6] and has seen increasing adoption for cross-institutional data exchange [7], including in situations where data exchange is legally required, such as the transmission of digital certificates for incapability to work in Germany [8]. The use of FHIR can facilitate international exchange of medical data for clinical research [9], where the requirement for harmonization and dataset descriptions in MDRs is especially indicated or can be applied to implementations of terminology services. Nevertheless, MDRs will also be of increasing importance for the clinical routines, if data is to be exchanged with other institutions, to document and verify the schema-level information represented by the units of information interchange.

Datasets in FHIR

The HL7 FHIR standard is centered around the use of harmonized resource definitions. These machine-readable artefacts specify the data model and content formats on a high level. However, since the international FHIR specification does not seek to model every use case in healthcare for every jurisdiction, profiling mechanisms are foundational to the design of the standard. In this way, use-case- and jurisdiction-specific data exchange using harmonized profiles and extensions is possible and practicable [6]. Using these mechanisms, the existing resource definitons are constrained to disable unneeded elements or enforce their population, extensions are added to capture data elements required for this use case, and data elements are bound against terminology artefacts to ensure semantic interoperability. Since profiles are specified using FHIR resources, they can be exchanged between FHIR servers using the same mechanism as instance resources. Aditionally, there are registries of known FHIR extensions and profiles to foster re-use and cross-profile compatibility. A popular registry and collaboration profile is provided by the *Simplifier [10]* software, which serves as a governance platform for many specification projects [7, 11].

Meaningful Terminology Binding

The aforementioned binding against terminology is the fundamental aspect this work adresses. In FHIR resource definitions, codeable elements are prevalent. While some of those are bound against coding schemes provided by the FHIR specification (such as the gender of a patient, which has to be populated using codes from this standard coding scheme), many fields leave the binding to the profiling mechanism. For example, since diagnoses are coded differently in most jurisdictions, a US-specific profile on *Condition* would specify a binding against a different terminology than a profile specific for use in Germany [6].

Because of the importance of standardized medical terminology, FHIR also provides means of specifying these terminologies and coding systems as FHIR resources themselves. There are two central FHIR resources in this regard: *CodeSystem* (CS) and *ValueSet (VS)*. Using the first of these, concepts are defined. Each concept is assigned a code that is unique within the CS. It can have rich properties which link concepts to each other, or provide additional information such as deprecation status or semantic annotation using reference terminologies such as SNOMED CT [6]. A CS itself is uniquely identified using an assigned canonical URI, and whenever a code from this code system is used in another FHIR resource, the canonical URI has to be provided. Hence, codes in FHIR can be more appropriately thought of as a two-tuple of canonical URI and the code, rather than only a code by itself.

Because many CS define a rather large amount of concepts, the second resource type, *ValueSet*, is used to pick codes from these CS for a specific use case. For example, in a diagnosis CS, there may be a hierachy of codes associated with Diabetes Mellitus, which can be provided in such a VS. Like CS, VS are also uniquely identified using a canonical URI. It has to be stressed that these VS are not restricted to picking codes only from a single CS, but that they may refer to multiple CS or VS. Also, if the included CS define properties linking concepts, they may be specified using an implicit specification ("all sub-codes of Diabetes Mellitus") rather than cherry-picking the appropriate codes.

Resource definitions and profiles always refer to VS when specifying allowable codes for a codeable element, rather than to a CS directly. Hence, if all codes from a CS are required, the CS resource can provide an implicit VS with all available codes by specifying an additional canonial URI for that purpose. The relationships between the resource types CS and VS, and their use within profiles, are illustrated in Figure 1.

To work with these resources, FHIR not only specifies the resources, but also a HTTP-based interface that can be used to consume these resources from a FHIR TS that implements these operations, such as CSIRO Ontoserver [12]. This is also useful for representing terminology independent of other parts of the FHIR specification. These interfaces are also particularly important for providing the *expansion*, i.e. the list of all contained codes of a VS that makes use of implicit specifications, since this operation may require significant computing power for complex CS with large polyhierachical relationships, such as SNOMED CT.

Since profiles depend on the specification of terminological artefacts, they are imperative to distribute together with profile and extension definitions. Hence, collaboration platforms such as Simplifier also support the distribution of these resources.

Concept Representation in Metadata Repositories

In the field of metadata representation, the ISO/TS 21526 Health informatics — Metadata repository requirements (MetaRep) is the emerging successor of the commonly

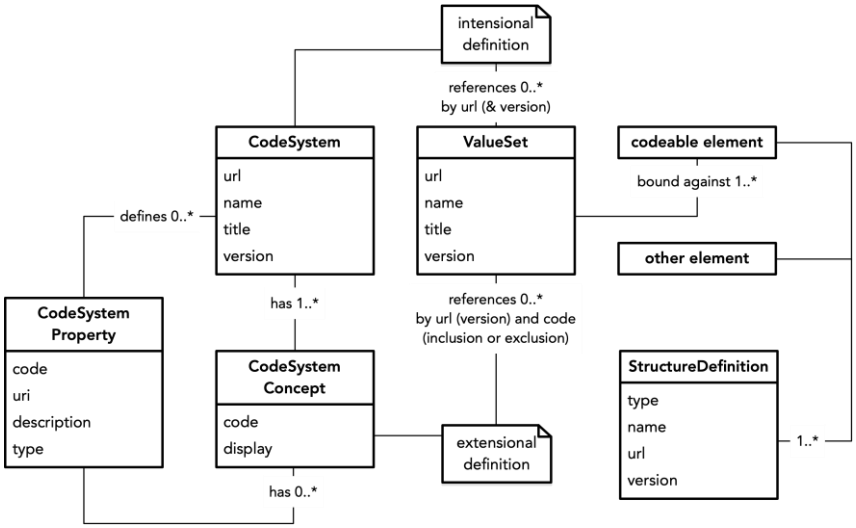

Figure 1: Relationships between CodeSystem, ValueSet and profiles/StructureDefinition in HL7 FHIR R4

used ISO 11179 [4, 13]. The designated successor offers several improvements with respect to its predecessor, as could be shown in a previous work [14]. The improved ISO 21526 concept package is of great interest for working with synchronized concepts. The standard describes (in a simplified way) how concept systems, the included concepts and their connections should be represented in MDRs. In addition to the elementary code, a concept also has a URI that refers to external definitions. Thus, ISO 21526 directly anticipates the external definition and management of concepts.

However, existing metadata repository implementations so far lack the support for accessing externally-provided terminology or VS. As they are often conceptually inspired by the ISO 11179 standard, they require the definition of *concepts within concept systems* which are comparable to VS within FHIR [15]. This information has to be created within the MDR implementation itself.

Since FHIR modelling has become prevalent in the Medical Informatics research community, the use and availability of terminological artefacts referenced in these definitions has also drastically increased, with many use-case specific ValueSets being shared in registries such as Simplifier. However, any dataset description of these profiles in an MDR will also necessitate the definition of the referenced terminologies to verify conformance on the schema level. To our current knowledge, no MDR available on the market can make use of these existing terminology resource, and hence requires the creation of the respective catalogs in manual, labor-intensive and error-prone processes. We strongly advocate for a separation of concerns, where the MDR is not responsible for the maintenance of terminological artefacts, which is currently not adopted by the MDR implementations on the market.

To close this perceived gap, we present a flexible and powerful approach that consumes VS as specified in HL7 FHIR from a variety of sources, in order to convert them to the concept system format required by the respective MDR. In this way, maintenance of terminological artefefacts is delegated to terminology servers and FHIR governance frameworks, which greatly facilitates the adoption of MDRs.

Results

TermiCron is centered around an adaptable pipeline that can be customized to both support a variety of input formats, and to enable the output of catalogs in a number of representations for different MDRs. The concept is illustrated in Figure 2.

Framework

The user interface of our application consists of a Command Line Interface (CLI) with a variety of commands and options to configure the end-to-end pipeline. This implementation is easy to adapt to other MDR implementations and requirements, and the CLI can be used for automated provisioning of MDR concepts with appropriate configuration. In this fashion, the user can use any of the input methods with any of the output methods implemented in our system by providing the approriate commands to the CLI.

Input

Since our concept revolves around the FHIR Terminology Module specification, we support the input of resources (VS and CS with implicit VS) in FHIR R4 format via four different routes. First, a FHIR TS can be used to host the resource definitions that are required for conversion. Using an instance of the FHIR resource type *Bundle*, a collection of resources for conversion can be referenced on the same server. This abstraction is required since not every resource available on a TS may be appropriate to provide to an MDR, especially with regards to reference terminologies such as SNOMED CT, which provide a very large number of concepts.

We have chosen the Bundle (with type “Collection”) in part due to the fact that many FHIR-based terminology servers, in particular Ontoserver [12], support the storage of these resources. Furthermore, Ontoserver offers facilities for syndicating resources across a network of connected instances, including Bundles. Hence, the bundle specified by our approach can be distributed via this syndication process in addition to the resources

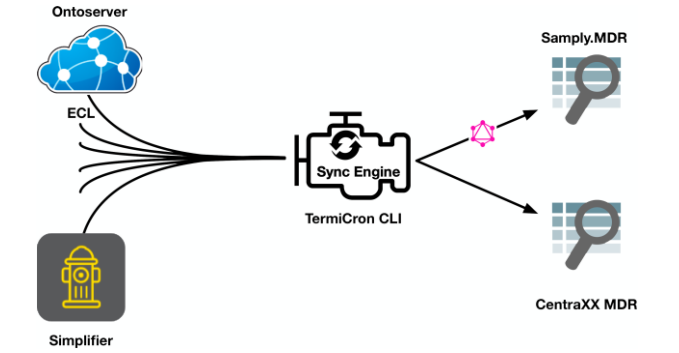

Figure 2: TermiCron Concept

themselves. We have specified a publicly-available profile to computationally validate bundle instances against the requirements of our impementation, and have developed a user interface for the creation of conformant bundles.

Another route for ValueSet input into our system is the file-based ingest. This process accesses files from a given directory on internal storage to facilitate ad-hoc conversion of a small selection of resources. However, this method may require access to a supporting FHIR Terminology Server that is used for querying the expansion of VS with implicit definitions, as this is not generally provided by the serialized resource.

Because Simplifier is increasingly used as a collaboration platform for specifying use-case-specific packages of profiles and extensions, with associated VS, we also provide support for consuming those packages directly from the Simplifier Package Registry, by specifying only the name of the package, and optionally the version number. However, this implementation is internally based on the file-based input route described above, and hence requires access to a supporting FHIR TS for VS expansion.

A fourth use case was implemented in order to convert expressions in the SNOMED CT Expression Constraint Langugage (ECL) ad-hoc into VS, and consequently into an MDR-compatible representation. ECL is a query language that can be used to select a subset of SNOMED CT codes in a specific hierarchy, or using properties. Using this powerful grammar, queries such as "all conditions/disorders related to Diabetes Mellitus" can be very easily created, as shown in Listing 1.

Such queries can be evaluated by FHIR TS implementations to retrieve the list of all concepts represented by this query. Especially with increasing data exchange across institutions using SNOMED CT within large initiatives such as the Medical Informatics Initiative in Germany [16], the need to represent these kinds of code catalogs in MDRs will increase as well. To facilitate conversion of the ECL expansion into adequate MDR catalogs, attributes such as the computer- and human-readable names of the catalog, as well as the business version, have to be specified by the user.

The next step of the process converts the FHIR representation to a simple model that serves as the common denominator for the output methods. This

```
< 64572001 |Disease| : 42752001 |Due to| = <<
73211009 |Diabetes mellitus|
```

Listing 1: Example ECL Expression

models is very similar to the FHIR model, but simpler with regards to computational complexity.

Output

There are a number of different methods available for transferring the semantic information contained in this model to the MDRs: File-based or a direct transfer via existing APIs. As the transfer format, either an ISO 21526-compliant representation, or the proprietary formats of the MDR can be used. The uniform metadata interface, called QL^{4}MDR [17], is used for standards-compliant transmission. The interface is based on GraphQL and is designed to allow ISO-compliant queries to MDRs, even if these and their underlying data storage are not standard-compliant. QL^{4}MDR was originally designed for ISO 11179 but extended for 21526 [14]. The file-based transfer is meant for systems that do not allow automatic requests or even changes via their proprietary APIs. For this, the file format must additionally be adapted to the local idioms of the systems. To facilitate the development against other MDR implementations, TermiCron was developed with generic interfaces to minimize the adaptation effort.

Availability of Results

For prototypical development, two MDR systems were integrated and all synchronization methods were successfully tested: the open-source Samply.MDR [15] and the commercial Kairos MDR [18].

The source code of the application was implemented in Kotlin and is freely available under the terms of the GNU Affero General Public License at https://github.com/itcr-uni-luebeck/TermiCron.

Discussion

A common semantic understanding is essential for harmonized data integration. MDRs, based on value sets and controlled vocabularies, are an important control and quality instance for the integration of healthcare-related data. However, in order to be useful, the data in the MDR must always be up-to-date. Yet, data maintenance is a time-consuming endeavor. Therefore, our study aims at improving and helping data maintenance. Through TermiCron, the simple and automatic use of consented VS was made directly available for data integration. The synchonization of ValueSets through a single source of truth can profitably facilitate the use of schema-level semantics in MDRs. The workload that data curators have in maintaining VS has been reduced. In addition, the use of FHIR has also made the internal versioning system and existing governance structures in Simpifier available to MDRs. By using the Ontoserver and its syndication in particular, an organized and audited distribution of resources, VS, CS and TermiCron bundles, can be enabled. This will enable synchronization on a large, or national level.

To enable TermiCron for use in MDRs, the system must be adapted to local requirements and formats if the system does not have a standard-compliant interface. The pipeline has been implemented as generically as possible to facilitate easy integration with existing systems. However, any new integration must take care of data inconsistencies. If a VS is dropped from the synchronization as obsolete, it must not be deleted to leave the legacy data consistent and valid. For this, connections between the ValueSets and versions must be created and made available. On the FHIR side this is possible by using ConceptMaps and on the MDR side concept relations are considered in the ISO standard. But not all MDR systems implement these entities – in our use case, the Kairos.MDR provided such functionality. The ISO-compliant QL^{4}MDR interface offers the entities, but the Samply.MDR does not. Therefore, change management must be considered separately for each MDR.

Conclusions

In summary, it should have become clear that metadata repositories (MDR) are primarily concerned with data structures and their description. With the goal of supporting semantic interoperability, the topic of "terminology" for annotating data structures as well as value lists also plays an important role in MDRs.

In view of the internal logic of complex terminologies, the provision of specific terminological services requires an effort that is not the core task of an MDR and should therefore be omitted. Structurally, terminologies also differ widely (e.g., LOINC, ICD-10, MedDRA), so that the maintenance of terminologies alone, as well as annotated lists of values, should be delegated to a TS via demanding services. An MDR then makes use of such services via interfaces. So, bridging the semantic and organizational gap between both import infrastructure components will enable a shared use and reuse of metadata and thus the corresponding healthcare information.

Acknowledgements

This work is funded by the German Federal Ministry of Education and Research (BMBF) as part of the Medical Informatics Initiative Germany, grant 01ZZ1802Z and was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft) DFG grant IN50/3-2. The authors would like to extend their sincere thanks to KAIROS GmbH of Bochum, Germany, developers of CentraXX MDR, for their support of this work.

References

[1] Prokosch H-U, and Ganslandt T, Perspectives for medical informatics, *Methods of Information in Medicine*. **48** (2009) 38–44.

[2] Krumm R, Semjonow A, Tio J, Duhme H, Bürkle T, Haier J, et al., The need for harmonized structured documentation and chances of secondary use–Results of a systematic analysis with automated form comparison for prostate and breast cancer, *Journal of Biomedical Informatics*. **51** (2014) 86–99.

[3] Galis A, and Gavras A, The Future Internet: Future Internet Assembly 2013: Validated Results and New Horizons, Springer Nature, 2013.

[4] ISO/IEC, ISO/TS 21526:2019 - Health informatics — Metadata repository requirements (MetaRep), *ISO*. (2019).

[5] Bender D, and Sartipi K, HL7 FHIR: An Agile and RESTful approach to healthcare information exchange, in: Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, IEEE, 2013: pp. 326–331.

[6] Benson T, and Grieve G, Principles of Health Interoperability: SNOMED CT, HL7 and FHIR, 3rd ed., Springer International Publishing, 2016.

[7] Sass J, Bartschke A, Lehne M, Essenwanger A, Rinaldi E, Rudolph S, et al., The German Corona Consensus Dataset (GECCO): a standardized dataset for COVID-19 research in university medicine and beyond, *BMC Medical Informatics and Decision Making*. **20** (2020) 341.

[8] Kassenärztliche Bundesvereinigung, Anwendungen der TI - Elektronische Arbeitsunfähigkeitsbescheinigung, (2021).

[9] Leroux H, Metke-Jimenez A, and Lawley MJ, Towards achieving semantic interoperability of clinical study data with FHIR, *J Biomed Semant*. **8** (2017) 41.

[10] Firely B.V., SIMPLIFIER.NET, (n.d.).

[11] Wulff A, Haarbrandt B, and Marschollek M, Clinical Knowledge Governance Framework for Nationwide Data Infrastructure Projects., in: EHealth, 2018: pp. 196–203.

[12] Metke-Jimenez A, Steel J, Hansen D, and Lawley M, Ontoserver: a syndicated terminology server, *Journal of Biomedical Semantics*. **9** (2018) 24.

[13] ISO/IEC, ISO/IEC 11179-3:2013 - Information technology -- Metadata registries (MDR) -- Part 3: Registry metamodel and basic attributes, (2013).

[14] Ulrich H, Kock-Schoppenhauer A-K, Drenkhahn C, Löbe M, and Ingenerf J, Analysis of ISO/TS 21526 Towards the Extension of a Standardized Query API, *Stud Health Technol Inform*. **275** (2020) 202–206.

[15] Kadioglu D, Breil B, Knell C, Lablans M, Mate S, Schlue D, et al., Samply.MDR - A Metadata Repository and Its Application in Various Research Networks, *Stud Health Technol Inform*. **253** (2018) 50–54.

[16] Semler SC, Wissing F, and Heyder R, German Medical Informatics Initiative, *Methods Inf Med*. **57** (2018) e50–e56.

[17] Ulrich H, Kern J, Tas D, Kock-Schoppenhauer A-K, Ückert F, Ingenerf J, et al., QL4MDR: a GraphQL query language for ISO 11179-based metadata repositories, *BMC Medical Informatics and Decision Making*. **19** (2019) 45.

[18] Link D, CentraXX MDR | KAIROS GmbH, *CentraXX MDR*. (n.d.).

Address for correspondence

Joshua Wiedekopf, Universität zu Lübeck, Ratzeburger Allee 160, 23562 Lübeck, Germany; Telephone: +4945131015646; E-Mail: j.wiedekopf@uni-luebeck.de