Semantic validation of standard based electronic health record documents with W3C XML Schema

C. Rinner, S. Janzek-Hawlat, S. Sibinovic, G. Duftschmid

Section of Medical Information and Retrieval Systems

Core Unit for Medical Statistics and Informatics

Medical University of Vienna, Austria

Corresponding author:

Christoph Rinner

Section of Medical Information and Retrieval Systems

Core Unit for Medical Statistics and Informatics

Medical University of Vienna

Spitalgasse 23

1090 Vienna, Austria

++43 1 40400 6693 (Phone)

christoph.rinner@meduniwien.ac.at

Summary

Objectives. The goal of this article is to examine whether W3C XML Schema provides a practicable solution for the semantic validation of standard based electronic health record (EHR) documents. With semantic validation we mean that the EHR documents are checked for conformance with the underlying archetypes and reference model.

Methods. We describe an approach that allows XML Schemas to be derived from archetypes based on a specific naming convention. The archetype constraints are augmented with additional components of the reference model within the XML Schema representation. A copy of the EHR document that is transformed according to the before-mentioned naming convention is used for the actual validation against the XML Schema.

Results. We tested our approach by semantically validating EHR documents conformant to three different ISO/EN 13606 archetypes respective to three sections of the CDA implementation guide “Continuity of Care Document (CCD)” and an implementation guide for diabetes therapy data. We further developed a tool to automate the different steps of our semantic validation approach.

Conclusions. For two particular kinds of archetype prescriptions, individual transformations are required for the corresponding EHR documents. Otherwise, a fully generic validation is possible. In general, we consider W3C XML Schema as a practicable solution for the semantic validation of standard based EHR documents.

Keywords: Medical Records; Medical Records System, Computerized; Reference Standards;

1 Introduction

The integration of patient data stored in different currently mostly isolated institutional electronic health record (EHR) systems is widely seen as a starting point for various expected improvements of health care. The European Union acts as a political driving force behind this vision, and underlines its corresponding commitment by naming “interoperability of EHRs” in its *eHealth action plan* as one of the goals to be strived for by member states [1].

A major prerequisite for an efficient realization of interoperable EHRs is the application of EHR standards. In this domain the ISO/EN 13606 EHR architecture standard [2, 3] and the HL7 Clinical Document Architecture Release 2 (CDA) standard [4, 5] build the most important representatives. The architecture of the openEHR foundation [6] builds a third remarkable source even though it does not have the status of an official standard. The openEHR and ISO/EN 13606 architectures incorporate the so-called *dual model approach*. This technique is characterized by the fact that it separates knowledge from information [7]. As described in the following, the CDA also follows this idea. Information is stored conformant to a static reference model (RM) which builds the first part of the dual model approach. ISO/EN 13606 EHR extracts and CDA documents are instances of such RMs. As also indicated by the “D” in CDA, documents represent a common context unit for the storage and exchange of EHR data[1]. Clinical documents conforming to the ISO/EN 13606 respective the CDA standard – which we will call *EHR documents* in the following – are the basis on which we build our work on semantic validation. EHR documents are represented in XML format and may be checked against the corresponding RM for syntactic validity. The knowledge is represented as ISO/EN 13606 archetypes (part 2 of [2, 3]) respective as CDA

[1] In the ISO/EN 13606 standard a document corresponds to an EHR extract comprising a single instance of the RM’s class COMPOSITION. In the context of this work we do not consider EHR extracts consisting of multiple documents.

implementation guides. They serve to define the semantics of clinical concepts which are part of EHR documents. This is done by specifying constraints on the predefined generic data structures of the RM. CDA implementation guides define the semantics of complete EHR documents and prescribe a set of corresponding constraints on RM classes, starting with the root class of the RM. ISO/EN 13606 archetypes may either represent complete EHR documents by constraining the RM class COMPOSITION or they may represent only parts of a document by constraining finer-grained classes of the RM.

ISO/EN 13606 archetypes and CDA implementation guides may be used to check the semantic validity of EHR documents. ISO/EN 13606 archetypes specify the semantic definitions in computer-processable form, CDA implementation guides just consist of textual instructions. HL7 is currently working on the so-called “templates” specification [8], which will also allow a computer-processable representation of the semantic definitions within CDA implementation guides. Currently, however, the authors are not aware of any existing CDA templates that are based on the model defined in [8]. Thus, only textual CDA implementation guides and ISO/EN 13606 archetypes were considered in this paper. The general term *archetype* is defined in [9] as “a model of a clinical or other domain-specific concept which defines the structure and business rules of the concept”. In the following we will use this general term to subsume the specific implementations *ISO/EN 13606 archetype* and *CDA implementation guide.*

When communicating EHR documents it is advisable to validate them on receipt to ensure a faultless processing and integration in the receiving system [10]. As stated in part 4 of [2, 3], the integrity of the EHR information that is stored, processed, and communicated is an essential security requirement, which can only be achieved by validating this information.

For EHR systems based on the dual model approach, two types of validation must be distinguished: (1) The syntactic validation against the RM only, without considering archetype-based prescriptions, and (2) the semantic validation against the RM *and* against the semantic definitions within archetypes. A semantic interoperable exchange of EHR documents can only be guaranteed if the documents are semantically validated. In particular, if EHR documents are confirmed to satisfy the prescriptions of certain archetypes, it will become possible to reliably query these documents respective parts thereof, based on the underlying archetypes' structure.

For the syntactic validation of EHR documents, W3C XML Schema [11], in the following just called XML Schema, is commonly used. Official XML Schemas are available for the RMs of CDA [5] and openEHR [6]. In the case of ISO/EN 13606 an unofficial XML Schema is available from [12]. Within our work we used the schemas from [5] and [12], the latter with minor adaptations in the definition of the data types. Further, an Eclipse based tool [13] for the syntactic validation of CDA documents has recently been published. We used this tool to double check the results of the XML Schema based validation of CDA documents.

In the present article we propose a method for the semantic validation of archetype-based EHR documents originating from a dual-model EHR architecture. The paper is organized as follows: Section 2 presents an overview of related work on the validation of EHR documents. Section 3 summarizes the objectives of our work. Section 4 introduces a method for the semantic validation of EHR documents based on XML Schema, which was tested for the ISO/EN 13606 and CDA standards. The corresponding results are presented in section 5. After a discussion of our method in section 6 we conclude the paper in section 7 with a short outlook on future work.

2 Related work

The constraints defined in CDA implementation guides are currently typically validated using manually created Schematron scripts [14, 15]. Corresponding Schematron scripts for several CDA implementation guides have been integrated into web services [16, 17]. For a full semantic validation these services offer an additional syntactic validation against the CDA RM using XML Schema.

In [18, 19] Maldonado and colleagues describe an archetype-based tool named LinkEHR-ED, which provides different services in the transformation of proprietary health data into data that conforms to a given EHR RM (the tool was tested with the RMs of ISO/EN 13606, openEHR and the CDA) and one or more archetypes. Amongst others, it contains a validation module that tests for a given archetype whether its constraints correctly narrow the corresponding RM classes respective the parent archetype from which the current archetype is derived. The validation of EHR documents is not particularly addressed in [18, 19].

Martinez-Costa and colleagues represent ISO/EN 13606 archetypes in a way that allows their semantic management and processing comparable to functionalities known from the Semantic Web domain [20]. They propose to transform ISO/EN 13606 archetypes represented in the archetype definition language (ADL) (part 2 of [2, 3]) into an Ontology Web Language (OWL) representation, as this would allow a more efficient implementation of semantic activities such as comparison, classification, selection and consistency checking of ISO/EN 13606 archetypes. Validation of EHR extracts by means of an OWL-based representation of ISO/EN 13606 archetypes is not particularly addressed.

Munoz and colleagues demonstrate in [10] how they developed a server that supports the storage and exchange of EHR extracts conformant to ISO/EN 13606 archetypes. They focus

on the development of a central server that mediates between communicating EHR systems. For this purpose it receives and stores archetyped EHR extracts and delivers them on request. It is mentioned that XML Schema is used to validate the correctness of EHR extracts. However, it is not addressed whether EHR extracts are validated semantically or just syntactically. As there is no indication on how the design of the XML Schemas was complicated by the “unique particle attribution constraint rule” (see section 4.2) it seems that only conformance with the RM was checked, where the unique particle attribution constraint rule does not become relevant.

In [21] Martinez and colleagues describe a patient monitoring system that communicates data from an Intensive Care Unit to an EHR server as ISO/EN 13606 EHR extracts. The data originate from medical devices and are transformed from an XML-based ISO/IEEE 11073-conformant format to ISO/EN 13606 EHR extracts using XSLT. XML Schema is used to validate the EHR extracts although it is again not transparent whether they are semantically validated or syntactically only. As analogous to [10] no problems concerning the “unique particle attribution constraint rule” are reported in the design of the XML Schema we again assume that only conformance with the RM was checked.

For the semantic validation of openEHR electronic health records, an early work [22] describes a corresponding architecture that is based on XML Schema and XSLT scripts. Recently the openEHR foundation has begun working on the implementation of a validation application programming interface (API) in Java [23]. It will allow EHR data conformant to the openEHR RM to be instantiated within a Java environment and to be validated against Java instances of openEHR archetypes obtained via the openEHR archetype parser [6].

To sum up, currently the most common approach for the semantic validation of CDA documents is to use XML Schema for checking conformance with the RM and the additional validation language Schematron to cover the prescriptions of implementation guides. Although several publications exist that describe implementations of the ISO/EN 13606 standard, none of them addresses the semantic validation of ISO/EN 13606 EHR extracts. For the openEHR architecture a semantic validation Java API is under development and an initial version is available from [23].

3 Objectives

Our goal is to examine whether dual model based EHR documents can be semantically validated in a practicable way using XML Schema. Besides XML Schema, other constraint languages such as Schematron or RELAX NG [24] would also be candidates for this purpose. In this work, however, we will focus on XML Schema as we expect several benefits from using this technology. We will address these benefits in the following.

XML Schema would be an obvious solution since it is commonly used to validate XML documents. As mentioned in section 2, XML Schema is already regularly applied in the EHR domain for the syntactic validation of EHR documents. The complete validation process could be simplified if XML Schema could also be used for semantic validation instead of using a different constraint language such as Schematron or RELAX NG for this purpose.

Applying XML Schema for the semantic validation of EHR documents also has the advantage that the hierarchical structure of the data that is prescribed by the archetype would become obvious. This should be particularly helpful for users who lack prior knowledge of archetypes and EHR standards but are familiar with the XML technology. The hierarchical structure of

the data does not become obvious with Schematron, which uses isolated rules to represent individual archetype prescriptions.

In [25] we described how the transformation of proprietary formatted EHR documents to standardized ISO/EN 13606 EHR extracts can be alleviated by expressing the required transformations as mappings between XML Schemas. The same XML Schemas could serve an additional purpose if they were also applicable for semantic validation of EHR extracts.

As XML Schema represents an official W3C standard, numerous tools for all kinds of operating systems as well as programming languages exist to create, edit, and validate XML documents with XML Schema. Many of these tools are also available under open-source licenses. The same is true for XSLT, which can be automatically derived from Schematron scripts. XSLT tools, however, primarily aim to support the transformation of XML documents, not their validation. Besides, when using Schematron as the constraint language for semantic validation one would rather look for tools supporting Schematron (which are not offered in the same variety as for XML Schema) rather than the derived XSLT scripts. We therefore aim for an easy to implement solution that makes use of XML Schema based tools as much as possible. Without requiring a complex framework, a corresponding "light-weight" solution could also help to lower the barrier for using dual model approach standards and thus contribute to their propagation.

4 Methods

As mentioned in section 3 our goal is to implement an approach for the semantic validation of EHR documents using XML Schema. This means we have to transform the archetype[2]

[2] Actually EHR documents will frequently be described by more than one archetype. In the following we assume that there is always one "root" archetype, which may include other archetypes via slots.

describing the EHR document into an XML Schema. As we will show in section 4.1, a transformation of just an archetype into an XML Schema is not sufficient, as it does not allow EHR documents to be completely checked for conformance with the RM. We will explain how the constraints of an archetype have to be augmented with additional components of the RM within the XML Schema representation.

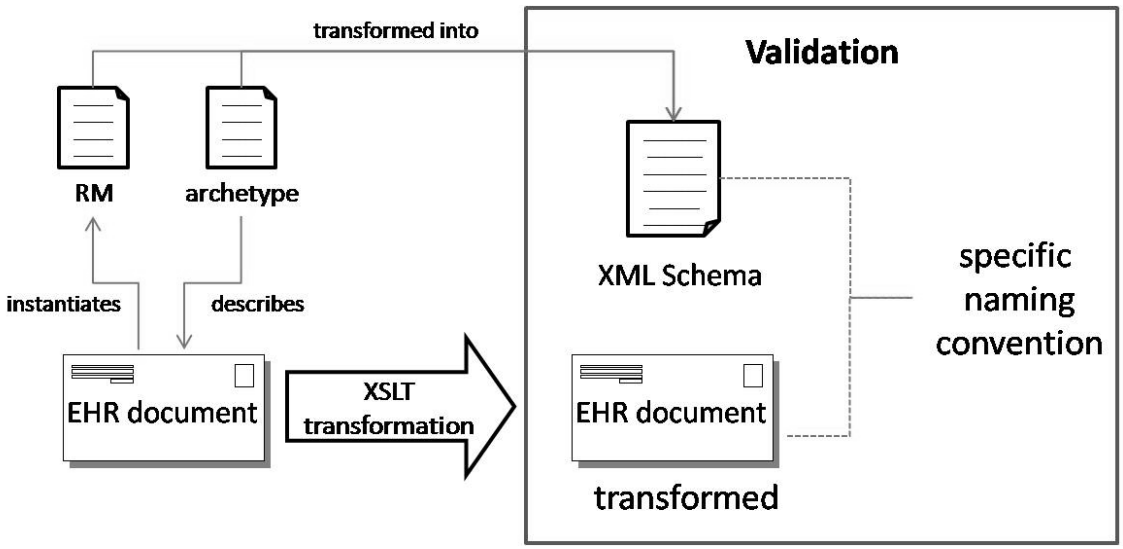

Figure 1: Overview of our approach for the semantic validation of an EHR document that is described by an archetype. An EHR document may be either a CDA document or an ISO/EN 13606 EHR extract comprising a single instance of the RM class COMPOSITION. Under the general term "archetype" we subsume ISO/EN 13606 archetypes and CDA implementation guides. The archetype and the corresponding RM are merged and transformed into an XML Schema using a specific naming convention. By means of ISO/EN 13606-specific respective CDA-specific XSLT scripts, a copy of the EHR document is transformed according to the before-mentioned naming convention. If the transformed copy of the EHR document is conformant to the XML Schema, the original EHR document has passed the validation process.

Unfortunately, a direct transformation of an archetype into an XML Schema is not possible due to XML Schema's *unique particle attribution constraint rule*. We will explain in section 4.2 how we solve this problem by using a specific naming convention for Schema elements.

For the semantic validation of EHR documents a copy of the EHR document is generated, the contents of which are transformed according to the before-mentioned naming convention. As will be described in section 4.3 we use XSLT scripts for this transformation which are specific to the underlying standards. The transformed copy of the EHR document is then

validated against the XML Schema. If this validation succeeds the original EHR document is clear for any further processing (e.g., import into a receiving EHR system). A summary of the complete validation process is depicted in Figure 1.

4.1 Merging the archetype and the reference model within the XML Schema

For a semantic validation of EHR documents the prescriptions of the archetype and the RM have to be united. In [18] a method is presented where these two different types of prescriptions are merged within a so-called "comprehensive archetype" for the purpose of mapping legacy data to EHR documents. This merging means that the comprehensive archetype contains the complete class hierarchy of the RM with all its attributes and relations, even though only a fragment of them is constrained by the archetype. The dynamic derivation of a comprehensive archetype from an archetype may be seen as a temporary switch to a single model approach to simplify a particular task (e.g., mapping or validating EHR documents) within a dual model approach EHR environment.

We use the same concept of comprehensive archetypes represented as XML Schemas for the purpose of semantically validating EHR documents. For each archetype a corresponding XML Schema is created that unites the prescriptions of the archetype and the RM. These XML Schemas are conceptually similar to openEHR's *operational templates* [26] insofar, as both represent standalone fully populated artefacts. This is in contrast to archetypes which represent differential artefacts, i.e. they only include those attributes and relations of RM classes they constrain. In particular, also the archetype slots are filled with concrete instances in our XML Schemas as well as in the operational templates.

4.2 The unique particle attribution constraint rule problem

In the dual model approach the RM classes are constrained using archetypes. In archetype-conformant EHR documents this frequently results in multiple instances of the same RM class at the same hierarchical level but with different contents as depicted in Figure 2.

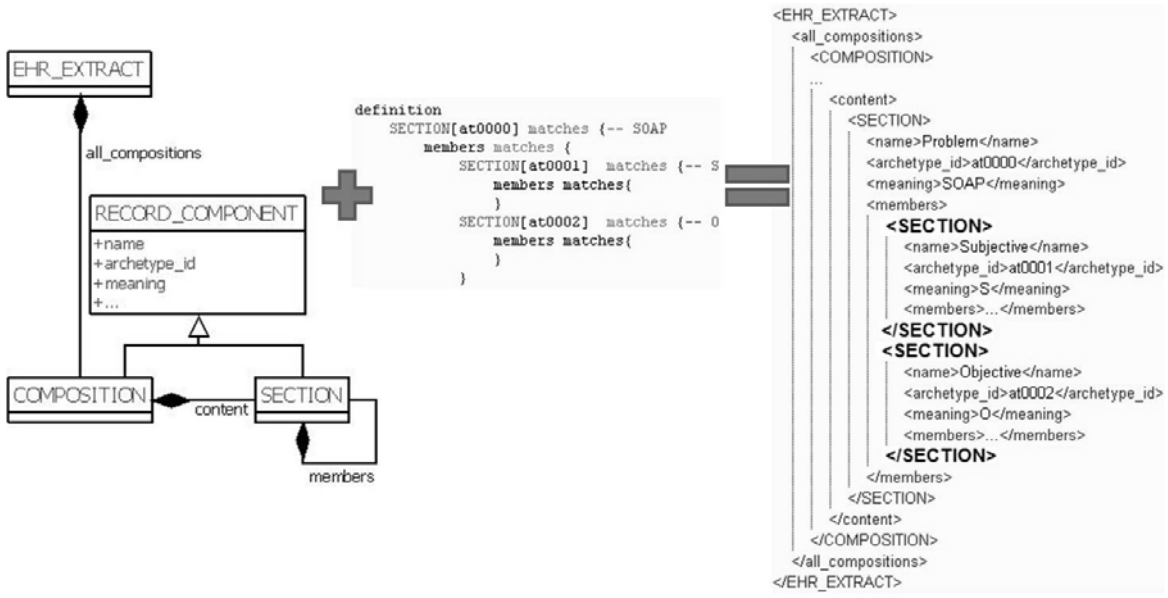

Figure 2: The RM (left, fragment of the ISO/EN 13606 RM in UML format) and the archetype (centre, in ADL format) are the blueprint for an EHR document (right, in XML format) with two SECTION instances on the same hierarchical level with different contents. Note that the EHR document also contains RM components that are not constrained by the archetype.

As Tun and colleagues demonstrate in [22], an XML Schema that is straightforwardly derived from an archetype such as the one depicted in Figure 2 would violate the *unique particle attribution constraint rule*. This rule requires an XML Schema to be defined in such a way that it is always possible to unambiguously associate an instance element in an XML document to a schema element without having to consider the instance element's content or structure. In other words, schema elements prescribing different contents or structures must not be named identically. An XML Schema straightforwardly derived from the archetype in Figure 2 would violate this rule by containing two schema elements named SECTION at the same hierarchical level with differing prescriptions for their contents.

To solve the before mentioned naming problem, we use a specific naming convention for designating schema elements. The goal is to generate unique names for the schema elements by incorporating a specific content of the archetype node[3] which is represented by the schema element within the XML Schema. An obvious choice is to use identifying attributes for this purpose, which exist in the form of the archetype node identifier for ISO/EN 13606 and openEHR respective the template identifier[4] for CDA implementation guides. As an example, the archetype depicted in Figure 2 contains an archetype node with identifier "at0000", which constrains the RM class SECTION. A CDA implementation guide may contain an archetype node with template identifier "1.2.3.4", which constrains the RM class Observation. Within the XML Schema we would name the corresponding schema elements "SECTION_at0000" respective "Observation_1_2_3_4"[5].

In [22] violations of the unique particle attribution constraint rule were also detected in cases, where archetypes restrict the values of a data field to being from a set of predefined instances of complex types. This problem can also be solved as described above, except that a different attribute of the archetype node has to be incorporated in the schema element name. If the value of a data field is prescribed to be one of several predefined complex type instances, the latter must obviously be distinguishable through at least one of their attributes. This attribute can then be incorporated in the schema elements corresponding to the complex type instances to achieve unique schema element names. In the example depicted in [22] a data field *subject* is prescribed to hold one of three predefined instances of complex type CV (Coded Value*)*,

[3] Archetypes are composed of nodes, where each archetype node constrains a single RM class.

[4] Instead of using template identifiers some CDA implementation guides require the different components of a CDA document to hold a predefined value within their *code* attribute to express the component's semantics. In this case the code value can be incorporated in the name of the schema element instead of the template identifier.

[5] Dots are not allowed within XML tag names therefore we replace them with underscores.

namely {codeValue=”701”, displayName=”self”}, {codeValue=”702”, displayName=”fetus”}, {codeValue=”703”, displayName=”new-born”}. For instances of CV the attribute *codeValue* is a good choice to be incorporated in the corresponding schema element name. This would result in three schema elements named “subject_701”, “subject_702”, and “subject_703” contained within a choice element.

The unique particle attribution constraint rule is a prescription that is unique to XML Schema. In Schematron a corresponding naming problem does not occur. A Schematron script consists of a sequence of patterns each containing rules, naming restrictions are not imposed. Rules are triggered via their *context* attribute for XML document elements holding a particular content. In this case the assertions contained in the rule, which would be used to represent the archetype constraints, are checked for the triggering document element.

A third problem was reported in [22] when archetype slots have to be represented within an XML Schema, which allow the dynamic inclusion of smaller-grained archetypes via wildcards. This problem is not relevant for us as wildcard-based references to archetypes do not appear in our XML Schemas. Two scenarios can be distinguished:

- Institutions that exchange EHR documents agree in advance to which archetypes the data are conformant to. This may be the case for example in a national EHR system, where only selected types of EHR documents – each corresponding to predefined archetypes – are exchanged between health care providers. If the archetypes used in this course contain smaller-grained archetypes via slots, also these sub-archetypes have to be chosen in advance. This means that the XML Schema representing the archetype does not contain wildcards within its references to schemas representing the

smaller-grained archetypes. Instead the wildcards are replaced by references to concrete schemas representing the smaller-grained archetypes.

- If communicating institutions do not agree in advance to which archetypes the EHR documents are conformant to, received EHR documents have to be scanned for the archetype identifiers indicated in the documents themselves[6]. EHR documents refer to concrete archetypes only, wildcards do not occur in the attribute holding the archetype identifier. The XML Schema can then be derived[7] from the detected archetypes, wildcards again do not appear in the schema.

4.3 Transformation of the EHR documents to be validated against the XML Schema

In the previous section we explained how we use a specific naming convention within the XML Schema derived from the archetype and RM to avoid violations of the unique particle attribution constraint rule. As the original EHR document uses the naming convention given by the original XML Schema pendant of the RM, it obviously does not validate against the XML Schema which we derive from the archetype and RM. As a prerequisite for validation, the original EHR document first has to be adapted to the naming convention used within our derived XML Schema. As mentioned in section 4.2 the naming convention is based on the idea of creating unique names for schema elements by incorporating identifying attributes of the corresponding archetype nodes in the schema element names. Within the EHR document we now have to rename each XML element analogously to its corresponding schema element. As shown in Figure 1 a copy of the original EHR document is used for this purpose. For any

[6] Components of an EHR document that comply with an archetype hold the identifier of the archetype within a corresponding attribute.

[7] In this scenario only an automatic derivation of the XML Schema from the archetypes is feasible. A corresponding tool is presented in section 5.1.

further processing of the EHR document after a successful validation, the original EHR document is used. The transformation of the "validation copy" is done via XSLT scripts [27].

As ISO/EN 13606 and CDA are based on different RMs, specific XSLT scripts have to be used. These scripts implement the before mentioned renaming procedure in a generic way. The renaming procedure is implemented generically insofar, as it can be applied to any archetype without considering its contents (see section 4.3.1). For archetypes containing two particular kinds of prescriptions, *xsl:templates* have to be provided in the scripts, which have to be parameterized in an archetype-specific way (see section 4.3.2).

4.3.1 Generic transformations

According to part 1 of [2, 3] each component of an EHR extract that corresponds to an archetype node has to hold the archetype's node identifier within its attribute *archetype_id*. In the course of the transformation it is checked for each XML element corresponding to an ISO/EN 13606 RM class instance within the EHR extract whether it contains a subelement *archetype_id* holding an archetype node identifier. If this is the case the XML element is renamed by appending the archetype node identifier to the original element name separated with an underscore (see Figure 3). If the archetype node identifier contains characters (e.g. dots) that are not allowed in XML tag names, they are also replaced with underscores. If the XML element does not contain a subelement *archetype_id*, no archetype-based prescriptions exist for the corresponding class instance within the EHR extract, i.e. the class instance hierarchically resides above or below the archetype-conformant part of the EHR extract. In this case the XML element does not have to be renamed, as the corresponding schema element of our XML Schema also remains identical to its pendent in the original XML Schema of the RM.

Figure 3: An instance of class SECTION within an ISO/EN 13606 EHR extract is transformed by appending the archetype node identifier to the original name of the corresponding XML element.

According to [28] each instance of a HL7 Reference Information Model (RIM) class – all class instances within a CDA document are instances of RIM classes – that corresponds to a set of template-defined constraints has to hold the template's identifier within its attribute *templateId*. Analogous to ISO/EN 13606 EHR extracts we therefore check for each XML element corresponding to a class instance of the CDA document whether it contains a subelement *templateId*. If this is the case the corresponding XML element is renamed by appending the template identifier to the original element name separated with an underscore (see Figure 4).

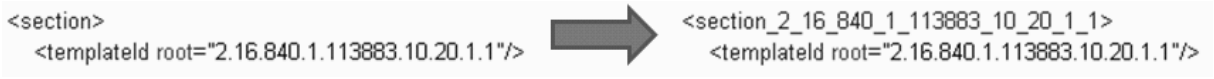

Figure 4: An instance of class *section* within a CDA document is transformed by appending the template identifier to the original name of the corresponding XML element. Dots contained in the template identifier, which is itself contained in the *root* attribute of element *templateId* are replaced with underscores.

The generic XSLT scripts defining the transformation of EHR extracts respective CDA documents consists of 22 respective 23 lines of code.

4.3.2 Archetype-specific transformations

Archetype-specific transformations of the EHR document are required in two cases:

- Archetypes may prescribe choices of complex types, such as coded values, for the value of a data field. In this case an identifying attribute of the complex type has to be used for the renaming process instead of the *archetype_id* respective the *templateId* (compare section 4.2). Within the generic XSLT scripts a corresponding renaming

xsl:template has to be provided which has to be parameterized with the paths of the data fields within the archetype, for which a choice of complex types is prescribed.

- Archetypes may contain an unordered set of archetype nodes with individual nodes occurring more than once within the set. XML Schema allows the definition of a set of unordered XML elements but limits the occurrence of each element in the set to zero or one. If elements within the set occur more than once, XML Schema requires the set to be ordered. One solution to this problem is to prescribe an "artificial" order for an unordered set of archetype nodes within the XML Schema, e.g. by sorting them by the archetype identifiers. The same order then also has to be established within the EHR document. Thus, a corresponding sort *xsl:template* has to be provided within the generic XSLT scripts. This *xsl:template* has to be parameterized with the paths of unordered sets within the archetype, for which individual elements occur more than once.

5 Results

Using a corresponding tool (see section 5.1) we tested the method proposed in section 4 by semantically validating different instances of EHR documents that were either conformant to an ISO/EN 13606 archetype or to a CDA implementation guide. For this purpose three different ISO/EN 13606 archetypes (see section 5.2), three sections of the CDA implementation guide "Continuity of Care Document (CCD)" [29] and an implementation guide for diabetes therapy data (see section 5.3) were used.

5.1 Semantic validation tool

We developed a tool that automates the different steps of our semantic validation method. Using the naming convention described in section 4.2 it automatically derives an XML Schema from a given archetype and a RM [30]. Based on the XSLT scripts (see section 4.3) it transforms a given EHR document and checks it for conformance with the generated XML Schema. If the EHR document is not valid, the full list of incompatibilities with the XML Schema is yielded. The transformation of an EHR document and its validation requires less than a second for our test examples. In contrast to Schematron, this includes the semantic as well as the syntactic validation within a single step.

The tool presumes that a computer-processable representation of the archetype exists. This is the case for ISO/EN 13606 archetypes represented in ADL. To receive a processable object tree of the ISO/EN 13606 archetypes, we used openEHR's archetype parser, which is freely available [6]. Concerning the transformation of constraints from ADL to XML Schema we used the internal XML Schema mechanisms to represent fixed values, enumerations, intervals and cardinality constraints. Internal references within the archetype were replaced by the target object nodes in the XML Schema. We did not yet implement transformations of regular expressions for which ADL and XML Schema use different dialects.

CDA implementation guides are represented as free text and are therefore only human-readable. Consequently, our tool is not able to derive an XML Schema from a CDA implementation guide, instead the schema has to be designed manually. The remaining validation process is covered by the tool.

5.2 Semantic validation of ISO/EN 13606 EHR extracts

To apply our method to ISO/EN 13606 EHR extracts, we manually derived three ISO/EN 13606 archetypes[8] from three existing openEHR archetypes "openEHR-EHR-OBSERVATION.dimensions.v1", "openEHR-EHR-OBSERVATION. body_weight.v1", and "openEHR-EHR-OBSERVATION.heart_rate.v1" according to the corresponding conversion provisions specified in part 3 of [2, 3].

Using the tool described in section 5.1 we automatically derived XML schemas from the three archetypes and the ISO/EN 13606 RM. We then generated sample EHR extracts from test data within an existing health information system using the method described in [25]. Again using our tool, we validated the EHR extracts against the generated XML schemas. This last step can also be done with standard XML tools.

5.3 Semantic validation of CDA documents

To apply our method to CDA documents we used three sections of the implementation guide "Continuity of Care Document (CCD)" [29] and an implementation guide for diabetes therapy data developed in the course of a project financed by the Austrian National Bank (OeNB), which focused on the documentation of diabetes data by patients and their transfer to care providers. As implementation guides are human-readable only we could not rely on our tool for deriving the XML schema but had to do this manually. To facilitate the manual creation process we used a standard XML tool[9] to automatically derive a draft version of the XML

[8] http://www.meduniwien.ac.at/msi/mias/models/CEN-EHR-ENTRY.[dimensions, body_weight, heart_rate].v1.adl

[9] The automatic generation of a schema from an XML document is typically supported by XML tools, such as XML-Spy (http://www.altova.com/) for example.

Schema from sample CDA documents[10] that were transformed as described in **Fehler! Verweisquelle konnte nicht gefunden werden.**section 4.3.1. This initial schema draft was then manually edited (e.g. cardinalities were changed, prescribed values were added, etc.) to integrate the semantic definitions specified in the corresponding implementation guide. Using this approach we transformed three sections of the CCD implementation and the private diabetes therapy data implementation guide into XML Schema. The CDA documents were validated against the XML Schemas using our tool.

6 Discussion

We pointed out in section 3 that the application of XML Schema for the semantic validation of EHR documents could deliver several benefits. Our results confirm that it allows EHR documents to be checked for conformance with the RM and archetypes, no additional validation language is required. Also as expected, the hierarchical structure of the EHR document that is prescribed by the archetype is made obvious by the XML Schema. While this structure is already recognizable in the ADL representation of ISO/EN 13606 archetypes, it is not observable from the Schematron representation of CDA implementation guides. We could further confirm that the same XML Schema can be used for the transformation of proprietary formatted EHR documents to standardized ISO/EN 13606 EHR extracts as well as for the semantic validation of the EHR extracts. As we expected, we were well supported in our work with existing XML tools. In particular, we used open source Java libraries based on SAX and Xalan XSLT for XML transformation and validation, and Altova's commercial tool XML Spy for editing XML Schemas.

[10] A sample CCD document is available from http://www.hl7.org.

However, also some shortcomings of XML Schema in our context have to be considered. We will report on these shortcomings in the following and provide recommendations how they may be overcome:

- The XSLT scripts performing the transformation of the EHR documents have to be specifically parameterized in case of archetypes which prescribe (a) choices of complex types for data fields, respective (b) unordered sets with elements occurring more than once (compare section 4.3.2). For the first case, an archetype-specific parameterization of the scripts could be avoided if the validation was weakened insofar as only one attribute of the complex types were checked (e.g. in the example in section 4.2, attribute *codeValue* of complex type CV is checked whereas attribute *displayName* is ignored). This is for example also partially done in the CCD Schematron scripts provided by HL7. For the second case a solution could be that communicating institutions agree upon a particular order for sets that are allowed to be unordered in an archetype. Both are only partial solutions but could still be acceptable in many cases. Besides, the before mentioned two kind of prescriptions only occur in some archetypes. For many archetypes, including our test cases, a transformation of the corresponding EHR documents is possible without any parameterization.
- Inheritance can only be used to a limited extent in the design of the XML Schema, resulting in code duplication. If a RM class' attribute is constrained differently within two or more nodes of an archetype, the attribute has to be multiply defined within different schema elements, where each definition covers one of the different archetype nodes' constraints. As an example, the *meaning* attributes of the "subjective" and "objective" SECTIONs within the archetype depicted in Figure 2 are constrained to

the values “S” respective “O”[11]. This means the *meaning* attribute has to be defined in both schema elements representing the two SECTIONs. Consequently the well-known problem arises that if an adaptation becomes necessary it may have to be conducted on numerous places within the XML Schema. To overcome this problem, we developed a tool (see section 5.1) which automatically derives an XML Schema from an ISO/EN 13606 archetype. Although it is less convenient, a manual design of the XML Schema is nevertheless still possible. For CDA implementation guides, which are human-readable only, this is in fact the only option.

- XML Schema does not support definition of “inter-element” constraints, such as a blood pressure archetype prescribing that its systolic blood pressure component must always contain a higher value than its diastolic blood pressure component. This shortcoming will be solved in the upcoming version 1.1 of XML Schema, which will support assertion components[12] for this purpose.
- CDA implementation guides distinguish between mandatory (indicated by keyword SHALL) and optional (indicated by keywords SHOULD and MAY) constraints. This suggests that a semantic validation should yield errors for violations of the former and warnings for violations of the latter. XML Schema does not support the differentiation of errors and warnings. This limitation is not relevant for EN/ISO 13606 archetypes, as the ADL does not make a corresponding distinction. A potential solution (we have not evaluated yet) could be to create two separate XML Schemas, one holding the mandatory constraints and the other holding the optional constraints.

[11] According to the ISO/EN 13606 RM the *meaning* attribute, which is of complex data type Coded Value, holds the archetype node name in archetyped systems. For reasons of simplicity only meaning’s *displayName* attribute is shown in the EHR document in Figure 2.

[12] See http://www.w3.org/TR/2009/CR-xmlschema11-1-20090430/#cAssertions

- Archetypes may restrict a value to hold an arbitrary code from a given coding system. While it is possible to check via XML Schema whether the *codingScheme* attribute of a complex type Coded Data (CD) for example holds the correct identifier of the desired coding scheme, it is not possible to check whether the *codeValue* attribute actually holds a valid code of this coding scheme. In our opinion, however, XML Schema should not be blamed for this "shortcoming" as it is not the obvious medium for this kind of checking anyway. Instead, a network of communicating EHR systems should include a terminology server which might be queried for this purpose.

A current limitation of our semantic validation tool (compare section 5.1) is that it assumes one "root" archetype that describes the EHR document. The tool does not consider EHR documents adhering to different archetypes which are not united via a superordinate archetype. Concerning the problem of transforming choices of complex types instances in valid schema elements (see section 4.2) our semantic validation tool currently only supports choices of type Coded Value. Corresponding extensions of our tool are planned as future work.

During the transformation of openEHR archetypes into ISO/EN 13606 archetypes we discovered a problem caused by the transformation rules defined in part 3 of [2, 3]. These rules prescribe that the name of the original openEHR RM class is codified and written to the *meaning* attribute of the ISO/EN 13606 RM class to which the former is mapped. According to part 1 of [2, 3] the *meaning* attribute has to store the archetype node name in archetyped systems. Since the *meaning* attribute has cardinality 0..1, however, it cannot contain both at the same time.

7 Conclusions and Future Work

We have presented an approach which allows the semantic validation of EHR documents by means of XML Schema. The XML Schema unites prescriptions of the RM and archetypes. It uses a specific naming convention to avoid violations of the unique particle attribution constraint rule. By means of XSLT scripts a copy of the EHR document is transformed according to the before-mentioned naming convention. The transformed copy of the EHR document is then validated against the XML Schema. We tested our approach with EHR documents conformant to three different ISO/EN 13606 archetypes and two different CDA implementation guides.

Advantages of our approach are that using XML Schema for semantic validation (1) EHR documents may be checked for conformance with the RM and with archetypes using one single technology once the XML Schema has been created, (2) the hierarchical structure of EHR documents prescribed by an archetype becomes obvious, (3) the same XML Schema can be used for the transformation of proprietary formatted EHR documents to standardized EHR documents as well as for the latters' semantic validation, (4) an extensive suite of existing, in many cases open source tools can be used, and (5) the complexity of the dual model approach is reduced as the prescriptions of the archetype and the RM are united within one XML Schema.

Limitations are that (1) for archetypes containing two particular kinds of prescriptions, the XSLT scripts performing the transformation of EHR documents have to be parameterized in an archetype-specific way, (2) in the design of the XML Schema inheritance can only be used to a limited extent which suggests an automatic derivation of the XML Schema from an archetype, (3) XML Schema currently does not support the definition of "inter-element"

constraints, even though this will be solved in version 1.1, (4) XML Schema does not support the differentiation of errors and warnings.

As the transformation of EHR documents has to be parameterized for archetypes containing two particular kinds of prescriptions, a fully-generic XML Schema based semantic validation is not always possible. However, as these prescriptions only occur sporadically within archetypes and as the resulting parameterization may be avoided with additional agreements between communicating institutions, we see XML Schema as a practicable technology for the semantic validation of standardised EHR documents. When it comes to the semantic validation of large EHR extracts, which may in the extreme case contain a patient's complete EHR consisting of numerous documents, the XML Schema derived from the underlying archetypes will tend to become too complex to allow a reasonable handling, however.

As one of our next steps we plan to extend our semantic validation tool to consider also EHR extracts consisting of multiple documents. This will allow us to examine up to what size of an EHR extract our approach is still reasonably applicable. Further, we intend to implement the automatic derivation of XML Schemas from computer-processable CDA templates, when sample CDA templates become available that instantiate the model specified in [8]. To make the manual declaration of the archetype to which an EHR document complies obsolete, we plan to implement a procedure that scans an EHR document and automatically identifies the underlying archetypes.

8 Acknowledgements

This work was partially financed by the Austrian National Bank (OeNB) project #12683 "Telediab: Ein telemedizinisches Tagebuch für Patienten mit Diabetes mellitus Typ 2".

9 References

1. Commission of the European Communities. e-Health - making healthcare better for European citizens: An action plan for a European e-Health Area, available at: http://ec.europa.eu/information_society/doc/qualif/health/COM_2004_0356_F_EN_ACTE.pdf. 2004.
2. European Committee for Standardization. EN 13606 Electronic healthcare record communication. 2007.
3. International Organization for Standardization. ISO 13606 Electronic health record communication. 2008.
4. Dolin RH, Alschuler L, Boyer S, Beebe C, Beilen FM, Biron PV, et al. HL7 Clinical Document Architecture, Release 2. J Am Med Inform Assoc 2006;13(1):30-9.
5. Health Level Seven (HL7). Clinical Document Architecture, Release 2.0 http://www.hl7.org/v3ballot/html/infrastructure/cda/cda.htm. Last accessed July 17, 2009.
6. openEHR foundation. http://www.openehr.org. Last accessed July 17, 2009.
7. Beale T. Archetypes, Constraint-based Domain Models for Futureproof Information Systems, available at: http://www.openehr.org/publications/archetypes/archetypes_beale_web_2000.pdf. 2001.
8. Health Level Seven (HL7). Specification and Use of Reusable Constraint Templates. http://www.hl7.org/v3ballot/html/infrastructure/templates/templates.htm. Last accessed July 17, 2009.
9. International Organization for Standardization. ISO/TR 20514:2005 Health informatics -- Electronic health record -- Definition, scope and context. 2005.
10. Munoz A, Somolinos R, Pascual M, Fragua JA, Gonzalez MA, Monteagudo JL, et al. Proof-of-concept design and development of an EN13606-based electronic health care record service. J Am Med Inform Assoc 2007 Jan-Feb;14(1):118-29.
11. World Wide Web Consortium (W3C). XML Schema. http://www.w3.org/XML/Schema. Last accessed July 17, 2009.
12. Biomedical Informatics Group, Universidad Politécnica de Valencia, Spain. LinkEHR Normalization Platform. http://www.linkehr.net. Last accessed July 17, 2009.
13. HL7Book. Eclipse Instance Editor. http://hl7book.net/index.php?title=Eclipse_Instance_Editor. Last accessed July 17, 2009.
14. ISO/IEC International Standard for Schematron. http://www.schematron.com. Last accessed July 17, 2009.
15. Ferranti JM, Musser RC, Kawamoto K, Hammond WE. The clinical document architecture and the continuity of care record: a critical analysis. J Am Med Inform Assoc 2006 May-Jun;13(3):245-52.
16. National Institute of Standards and Technology (NIST). http://xreg2.nist.gov/cda-validation/validation.html. Last accessed July 17, 2009.
17. Alschuler Associates LLC. CDA Validator. http://www.alschulerassociates.com/validator/. Last accessed July 17, 2009.
18. Maldonado JA, Moner D, Boscá D, Fernández-Breis JT, Angulo C, Robles M. LinkEHR-Ed: A multi-reference model archetype editor based on formal semantics. Int J Med Inform 2009;78(8):559-70.

19. Maldonado JA, Moner D, Tomas D, Angulo C, Robles M, Fernandez JT. Framework for clinical data standardization based on archetypes. Stud Health Technol Inform 2007;12(Pt 1):454-8.
20. Martinez-Costa C, Menarguez-Tortosa M, Fernandez-Breis JT, Maldonado JA. A model-driven approach for representing clinical archetypes for Semantic Web environments. J Biomed Inform 2009;42(1):150-64.
21. Martinez I, Fernandez J, Galarraga M, Serrano L, de Toledo P, Escayola J, et al. Implementation experience of a patient monitoring solution based on end-to-end standards. Conf Proc IEEE Eng Med Biol Soc 2007;2007:6426-9.
22. Tun Z, Bird LJ, Goodchild A. Validating Electronic Health Records Using Archetypes and XML: CRC for Enterprise Distributed Systems: University of Queensland , available at: http://citeseer.ist.psu.edu/tun02validating.html. 2002.
23. openEHR foundation. Data Validation. http://www.openehr.org/wiki/display/dev/Data+Validation. Last accessed July 17, 2009.
24. International Organization for Standardization. Information technology -- Document Schema Definition Language (DSDL) -- Part 2: Regular-grammar-based validation -- RELAX NG. http://www.relaxng.org/. Last accessed July 17, 2009.
25. Rinner C, Wrba T, Duftschmid G. Publishing relational medical data as prEN 13606 Archetype compliant EHR extracts using XML technologies. Tagungsband der eHealth 2007 – Medical Informatics meets eHealth. 2007.
26. Beale T, Heard S. *open*EHR Templates Release 1.0.2, available at: http://www.openehr.org/releases/1.0.2/. 2009.
27. World Wide Web Consortium (W3C). XSL Transformations (XSLT) Version 2.0. http://www.w3.org/TR/xslt20/. Last accessed July 17, 2009.
28. Health Level Seven (HL7). Reference Information Model (RIM). http://www.hl7.org/v3ballot/html/infrastructure/rim/rim.htm. Last accessed October 9, 2009.
29. Health Level Seven (HL7) ASTM. Implementation Guide: CDA Release 2 - Continuity of Care Document (CCD). http://www.hl7.org/Library/Committees/structure/CCD.NearFinal.zip. Last accessed July 17, 2009.
30. Janzek-Hawlat S, Kuttin O, Sibinovic S, Duftschmid G. Automatisierte Generierung von XML-Schemata aus EN/ISO 13606 Archetypen. In: Schreier G, Hayn D, Ammenwerth E, editors. eHealth2009 & eHealth Benchmarking 2009 - Medical Informatics meets eHealth; Vienna 2009. p. 69-75.