Standards Based Adaptation of Clinical Documents for Interoperability of e-Health Services

Evgeniy Krastev¹, Dimitar Tcharaktchiev², Kalinka Kaloyanova¹, Lyubomir Kirov³, Petko Kovatchev¹, Simeon Abanos¹ and Nonka Mateva⁴

¹ Faculty of Mathematics and Informatics, University of Sofia St. Kliment Ohridsky, 5 James Bourchier Blvd., 1164 Sofia, Bulgaria
² Department of Medical Informatics, Medical University, Sofia, Bulgaria
³ Faculty of Medicine, University of Sofia St. Kliment Ohridsky, Sofia, Bulgaria
⁴ Department of Medical Informatics, Medical University of Plovdiv, Plovdiv, Bulgaria
eck@fmi.uni-sofia.bg

Abstract. The area of eHealth is constantly evolving in the last years. Health service providers generate and record large volumes of clinical data in electronic or paper format. Most of the time such data is transformed or duplicated in clinical documents manually. In other cases, clinical data is inaccessible due to incompatibility of data types and structures. Therefore, it is a challenge to transform existing models of clinical documents into internationally accepted standards that enable semantic interoperability among the participants in the clinical process. The objective of this paper is to present a model for transformation of such legacy models into a standard reference model that supports semantic interoperability in accordance with the EU approved standard CEN 13606. A realistic case study of information services in Bulgarian healthcare is used to build an UML model and design a corresponding CEN 13606 Archetype model. The results are obtained as part of an eHealth scientific program with the combined research efforts of a team of researchers having professional experience both in medicine and computer science. It allows building a prototype of an information system that enables semantic interoperability among clients in terms of RESTful web services provided by a NoSQL database.

Keywords: eHealth, semantic interoperability, CEN 13606 reference model, CEN 13606 archetype.

1 Introduction

Nowadays patients widely use modern information technology to consume services offered by the participants in the healthcare system—general practitioners (GP), labs, hospitals, etc. These services generate and record large volumes of clinical data in electronic or paper format. For example, in accordance with the best worldwide practices Bulgarian GPs record in electronic format data for more
than 25,000,000 medical exams annually [1]. This medical information is very comprehensive and contains data about medical history, diagnoses, medications, vaccination, allergies, lab results, results from consultant’s referrals and hospital stay, GP’s notes, etc. In a similar way the rest of the healthcare services providers (HSP) generate and record huge amounts of clinical data in electronic format. Undoubtedly, this data could be used in a very meaningful way for improving the quality of the health care services in Bulgaria. At the same time, the level of electronic exchange of clinical data and interoperability between HSPs remain quite limited. The establishment of direct connectivity between separate subjects of the healthcare system as well as the implementation of EU approved standards for representation of clinical data structures are one of the greatest challenges for the national healthcare system.

The main objective of this paper is to investigate the most frequently used definitions of clinical data structures and to propose a model for adapting them to CEN 13606 [3]. Harmonizing existing technologies with approved interoperability standards appears to be a better approach than replacing them with very new solutions. The proposed UML model is a result of the combined research efforts of a team of researchers combining professional experience both in medicine and computer science. The conclusions are based on original documents used in the clinical practice.

The paper is structured in seven sections. In the following section, we briefly describe the limitations in the information flow of XML documents employed to manage healthcare activities in Bulgaria. This realistic case study makes use of original information sources and provides evidence that the information flow is based on document definitions inconsistent with the CEN 13606 standard for semantic interoperability. It serves as a motivation to formulate in section 3 the problem statement in this paper. Section 4 establishes a common ground for harmonizing the existing document definitions in the national healthcare with the standard reference model of CEN 13606. Section 5 presents the proposed transition from W3C XML schemas used in legacy document definitions into the CEN 13606 reference model. Discussion about the applicability of the obtained results and concluding remarks are provided in sections 6 and 7.

2 Motivation Case Study

Management of health insurance funds and payments to healthcare services providers (HSP) involve gathering and processing a huge amount of information. These tasks are delegated to a specialized institution in Bulgaria – the National Health Insurance Fund (NHIF). NHIF collects account reports in electronic format from HSP like hospitals, GPs, dentists, and medical laboratories for the services they have provided to self-insured patients. The reports are submitted as
XML documents prepared in accordance with XML Schema definitions (XSD) designed and officially published by NHIF.

For example, let us consider the structure of such an XSD document employed to generate the account report in the GP practice [4]. This document has a set of global elements among which most important are elements Practice, Doctor and AmbList (see Fig. 1). Here AmbList represents a sheet of the Ambulatory paper.

**Fig. 1.** Relations between the Global elements in the XSD of a GP report.

**Fig. 2.** Global simple types in the XSD of a GP report.
book possessed by each patient, where for each visit of the patient the GP records medical status details such as prescribed pharmaceutical drugs, and directions for medical treatment, hospitalization, or specialized laboratory examinations. Recording of this kind of medical data is common for GP practices [1]. The XSD employs a set of 17 Global simple types to define the complex types of the Global elements to represent the data structure of the AmbList (see Fig. 2).

It is noteworthy that these simple types are direct references to standard W3C XML Schema datatypes such as string, date, boolean or named containers for a value of such datatypes like CodeType10 and Str150. The named containers include validation rules for the contained value in terms of regular expressions (see Fig. 3). These validation rules can be applied just to the value in the named container. For instance, it is impossible to create a validation rule in the XSD to ensure that DateFrom occurs before DateTo.

Moreover, the GP report as an instance of this XSD contains no trace of the semantic context accompanying the generation of the entries in the AmbList of this to report to the NHIF. It makes it difficult to trace the results of the prescribed treatment or even to relate together different sets of medical checks. For example, it is important to link the prescribed drug to a problem code, illness history or record the drug code with personalized dosage instructions and codes of risk factors [5]. A serious limitation for the efficient usage of this information is the one-way communication between the GPs and NHIF. It allows the NHIF just to manage the financial aspects in healthcare like accounting and distribution of total payments due to GPs and other HSPs for the healthcare activities they have completed. This substantially complicates the exchange of electronic data between different HSPs mainly because of using non-standard and incomplete data structures for representing clinical.

In these circumstances, a patient may face the following problems. Assume the patient has made some expensive and painful medical examinations ordered by his GP. Then, while on vacation or on a duty trip, the patient loses accidentally his Ambulatory paper book and he needs to visit another GP. The newly visited GP has no access to the AmbList records of the patient in the NHIF. In the general case, this GP cannot exchange information with the primary GP of the patient about the reasons he has prescribed the medical examinations or about certain discrepancies in the medical data provided by the patient. Most often in such case, the patient must repeat the medical examinations.

This model of communication in the healthcare system entails duplication of data and usage of different technologies for data management by each one of the HSPs. The XSD provided by the NHIF appear to be the only standard for healthcare data interchange in the national healthcare system. Finally, the absence of semantic context in the collection of medical data does not allow the application of modern means for computer-aided support for decision making at all the levels of the healthcare system [6].
3 Problem Statement

This case study exposes several information problems in the national healthcare system that can be discovered in many other countries as well. First, we note that the exchange of medical information is centralized in a single institution (NHIF) and the information flows unidirectionally to this institution. The scope and content of the electronic records are reduced to serve merely the purposes of an activity-based funding business model for health insurance management.

The W3C XML Schema definitions of these records do not satisfy any internationally recognized standard for semantic interoperability of health records like CEN 13606 [7]. These data structures do not refer to formally defined domain concepts. Therefore, clinical data represented this way cannot be validated and its meaning interpreted correctly by a software product in relation to a medical domain. The absence of semantic context in the XSD does not allow the information requester and the information provider to have a common understanding for the “content meaning” embedded in the exchanged information [8, 9]. As an immediate consequence, a patient usually communicates with different HSPs by means of paper documents such as their Ambulatory paper book. Besides, the content of paper documents is rarely complete and accurate [10]. Therefore, in the general case the patient must communicate the same way with multiple HSPs. This kind of communication causes many inconveniences for the end users of the healthcare system and it is a major source of dissatisfaction from the quality of the healthcare services.

```
<xsd:schema xmlns="http://nhif.bg/xsd/amblist"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema"
    xmlns:xdb="http://xmlns.oracle.com/xdb"
    targetNamespace="http://nhif.bg/xsd/amblist"
    elementFormDefault="qualified" version="1.5">
  <xsd:element name="Practice">
    <xsd:complexType>
      <xsd:sequence>
        <xsd:element name="PracticeCode" type="CodeType10"/>
        <xsd:element name="PracticeName" type="Str150" minOccurs="0"/>
        <xsd:element name="NHIFCode" type="CodeType10" minOccurs="0"/>
        <xsd:element name="ContractNo" type="CodeType6"/>
        <xsd:element name="ContractDate" type="xsd:date"/>
        <xsd:element name="FromDate" type="xsd:date"/>
        <xsd:element name="ToDate" type="xsd:date"/>
        <xsd:element name="ContrHA" type="xsd:boolean" default="0"/>
        <xsd:element ref="Doctor"/>
      </xsd:sequence>
    </xsd:complexType>
  </xsd:element>
</xsd:schema>
```

Fig. 3. Header of the XML Schema definition for a GP report.

The absence of direct connectivity between HSPs isolates the large volumes
of data they accumulate in executing their everyday activities in the healthcare domain. Modern eHealth approaches can help in finding reusable, scalable and knowledge-driven solution based on the implementation of standards and best practices in healthcare [11] [12]. The implementation of new approaches in a complex system as healthcare is a great challenge especially when it goes about a national healthcare system without system integrated medical databases and no direct connectivity between HSPs.

Therefore, in this paper we focus our efforts on the content of the XML documents submitted to NHIF on a regular basis by GPs in Bulgaria. Our main research hypothesis is that it is realistic to adapt these documents into the reference model of CEN 13606. In this paper, we prove that this task can be resolved in a typical case study.

Existing research papers provide evidence that the reference model of this standard can represent all the medical information contained within an electronic health record such as an Ambulatory list [13]. It enables different HSPs to exchange information in a semantically interoperable manner [14]. The CEN 13606 is compliant with the XML technologies employed to define the structure of documents submitted to the NHIF [15, 16]. Moreover, the relationships of CEN 13606 with HL7 and openEHR are well investigated [17, 18]. Finally, a successful harmonization of GP reports to NHIF with the reference model of CEN 13606 will have a great social effect because patients visit a GP before getting access to any other healthcare service. On the other side, semantic interoperability between GPs will improve the quality of the healthcare system.

4 CEN 13606 Standard for Semantic Interoperability

Interoperability refers to the ability two or more communicating parties to exchange information and use the exchanged information. There exist three levels of interoperability, syntactic, functional, and semantic. The levels of interoperability distinguish the degree a machine language can understanding the “meaning” of the exchanged information. Semantic interoperability provides a formal model allowing the receiving system to interpret the shared information by processing of formally defined concepts with a machine language.

The exchange of electronic health records (EHR) is an essential requirement for delivering eHealth services and products. In the general case EHR include patient demographics, his health status data, prescribed therapy, immunizations, laboratory data and other documentation required by the healthcare system. These records contain clinical data such as the GP records discussed in Section 2. Semantic interoperability in exchanging EHR plays an important role in healthcare as an instrument for the implementation of key functional and non-functional requirements that influence the effective management of the healthcare system and the quality of healthcare services.
The CEN 13606 standard is a five-part International standard approved by the EU for semantic interoperability of EHR. In this paper, we employ the first part of the standard, namely CEN13606-1:2019, where the EHR reference model is specified [3]. The reference model presents UML class diagrams and the technical details of the core classes that are required by this standard for delivering an extract from a clinical system to a recipient. The datatypes in the first edition of the standard CEN 13606 make use of CEN/TS 14796:2008 that is currently deprecated. Currently ISO 21090:2011 [19] is the International Standard that provides a comprehensive set of models for data types needed by all health IT systems including those represented in the reference model of CEN13606. This standard supports UML 2.0, specifies XML based representations of datatypes, and declares the semantics of the datatypes extending the ISO/IEC 11404 standard for terminology, notations, and datatypes [20]. All of it makes ISO 21090:2011 appropriate for adapting the datatypes in the XSD documents of the here considered case study to the datatypes in the reference model in CEN 13606-1:2019.

The reference model of CEN 13606 follows the hierarchical structure typical for the representation and storage organization of clinical data. Class EXTRACT is the root of a hierarchical structure of an EHR or a part of an EHR. This class identifies the owner of the information extracted from the EHR in terms of demographic data, access policy and version control. For shortness, we will not consider here these properties. Class EXTRACT references a hierarchically structured clinical data defined by classes all of which inherit class RECORD_COMPONENT. It ensures secure exchange of extract components in this hierarchy even in case of duplication of information. Class FOLDER is the container at the top level of the clinical data hierarchy. Its purpose is to divide the clinical data into compartments related to a single condition like individual ambulatory lists of patients as our case study. Class COMPOSITION is model of the information produced by a single data provider. This class describes the data provider (person responsible for executing the healthcare activity or healthcare service structure). Details about clinical action, observation, interpretation, or intention are hierarchically grouped in sections (class SECTION) and entries (class ENTRY), where each component of the hierarchy is referred to by the abstract class CONTENT. Typical examples for instances of ENTRY are “Diagnose”, “Prescribed drug”, while “Subjective symptoms”, “Documents” and “Procedure” are examples for instances of class SECTION. At the lowest level instances of class ENTRY by a hierarchy of clusters (class CLUSTER) and elements (class ELEMENT), where components of this hierarchy are referred to as items (class ITEM). Class ELEMENT is the leaf node in the CEN13606 reference model, and its instances contain a single data value (“Height”, “Body weight”, “Drug name” etc.).
UML Model for Transformation

CEN 13606 enables semantic interoperability by separating the representation of information and knowledge. At the first level of the proposed methodology, information is structured in terms of the core classes of the reference model introduced in the previous section. Knowledge is described at an upper level in terms of a language for defining archetypes specified in the second part of this standard known as CEN 13606-2:2019 [21]. This paper considers the transformation of information of existing XML documents into the CEN 13606 reference model, where the documents belong to the healthcare domain and are valid instances with respect to a given XSD document of a W3C XML Schema definition.

The first stage in transforming the information into the reference model of CEN 13606 is to convert the XSD document into a set of classes. Nowadays there exist many software tools to complete this task. For example, Java Architecture for XML Binding (JAXB) is one such software framework that allows mapping Java classes to XSD documents [22]. This way the Global elements displayed in Figure 1 are converted in Java classes with the same names. As a result, we get flexibility to manipulate freely separate pieces of semantic meaning that are otherwise embedded in the monolithic structure of an XML document.

The challenge is to fit the obtained classes at appropriate places into the framework of the reference model of CEN 13606. First, the transformation of information must preserve the semantics of the core classes and their relationships in the reference model. Second, the classes from the here considered healthcare domain must be mapped to classes in the reference model satisfying best practices for information management in healthcare [1]. In other words, a synergy of professional experience in both medicine and information technology is essential for the implementation of such transformation of information structures.

Let us consider the clinical data contained in an Ambulatory list (AmbList) displayed in Fig. 1. This figure shows the major entities that buildup the information content of an Ambulatory list. Among these entities, there are entities that are of administrative origin and others carry important clinical context. Entities with administrative content such as Practice (identification properties of the GP), Patient (identification properties of the patient), Doctor (identification properties of the GP) and the set of datamembers of the AmbList (the id code of the AmbList, date of issue, payment source etc.) map to the categories of FOLDER and COMPOSITION classes of entities.

It is common entities with semantic context related to healthcare procedures and services to serve as roots of hierarchical structures of documents with specialized content of information. Such documents like Docs (documents for specialized healthcare activities), SIMPLList (directions for specialized non-hospital medical care), Procedure (highly specialized healthcare services, laboratory...
probes) or *Profilact* (prophylactically administered activities) from Fig. 1 may contain observations for physical, neuro-psychological and general development, temporary labor inability and so on. Therefore, it is appropriate to map these entities to the semantic context of the *SECTION* class in the reference model. On the other side, there are also entities that carry summary of clinical experience, such as *MainDiag* and *Diag* (the Main Diagnose and secondary Diagnosis coded using ICD-10) from Fig. 1. It is common to discover this kind of entities as instances of *ENTRY* classes of *SECTION* type in the CEN 13606 reference model. These objects provide content support for the previously considered group of *SECTION* entities.

This way we have demonstrated our approach for transforming an existing clinical document into the CEN 13606 standard reference model. According to this approach:

- entities of administrative origin having the same values in each document instance are mapped to *FOLDER* and *COMPOSITION* classes of entities;
- entities of clinical semantic origin that require more detailed classification of content with different values for each instance are mapped to *SECTION* types; and
- entities that appear with different values for each instance of a *SECTION* type are mapped to *ENTRY* classes.

This transformation process obligatory requires practical experience in healthcare information management and consultation with a clinical professional.
The final step but not the least important in the transformation process is the mapping of ISO 21090:2011 compatible datatypes to the datatypes used in the XSD representation of the existing clinical document. In the here considered use case we discover user-defined datatypes like CodeType10 and Str150 (see Fig. 2) as well as simple types in XML such as string, date or boolean. ISO 21090:2011 supports directly simple XML types like string, date or boolean. Fig. 4 shows that these types can be wrapped in classes that are more complex, respectively ST, TS, and BL, if desired. The same refers to value containers such as CodeType10 and Str150.

The thus discovered information entities in XSD must be structured in a reference model that is compatible with the CEN13606 standard. Therefore, the second stage of the proposed information transformation involves extending core classes in the reference model with classes discovered in the XSD document. This way we to preserve the standard content in the core classes and the relationships among them as per the standard specifications. It ensures that the obtained extended reference model is compatible in terms of object-oriented design with the base implementation of that model. A similar approach is followed in [13] for publishing CEN 13606 compatible EHR extracts.
The extended reference model of a GP report is displayed in Fig. 5, where the shaded blocks represent the objects inheriting some core class in the base definition of CEN 13606. The state of the objects is determined by the values of the data members of the respective class. For simplicity of presentation core classes CLUSTER and ELEMENT as well as some, the discovered information entities in the XSD are not shown in Fig. 5. The proposed inheritance relationships allow these objects to be referenced in any software implementation that makes use of the core classes in this standard. The greatest challenge in developing the extended reference model is the proper selection of the base class among the standard core classes for each one of the classes obtained from the XSD model. ISO 21090:2011 compatible datatypes.

6 Archetype Design

The Reference Model shown in Fig. 5 satisfies the CEN 13606; however, it is not ready for practical implementations. One of its drawbacks is that it does not reflect well the business process followed by GPs to collect clinical information, evaluate, and prescribe a therapy to a patient in need. Another drawback is that this model is difficult to share and implement readily in software applications.
For this purpose, we convert the Information model shown in Fig. 5 into the Archetype model of CEN 13606 [21, 13]. This model allows us to structure in a meaningful way the CONTENT of the Reference Model of CEN 13606-1:2019 by means of the software tool LinkEHR Studio [23]. The obtained archetype CEN-EN13606-COMPOSITION.AmbulatoryList.v1 is shown in Fig. 6.

Unlike the Reference model shown in Fig. 4 the Archetype model transforms the data entities in the XML Schema definition of a GP report (see Fig. 1) into a set of four major SECTIONs that are intuitively understood by a GP. These SECTIONs represent the four major stages of a visit to a GP (Observation, Evaluation, Instruction and Action).

Initially a GP performs an OBSERVATION of the vital signs of the patient and records the measured values as well as gets familiar with the history of the illness or other clinical evidence. Next, the GP makes an EVALUATION of the clinical information in the semantic context of the overall health status of the patient. This allows the GP to assign a therapy for the illness in terms of a set of INSTRUCTIONs and propose a set of ACTIONs the patient must execute.

It is noteworthy that the proposed grouping of concepts based on their semantic context in the business process a GP follows in executing his activities is not clearly defined in the CEN 13606 standard. Similarly, to the openEHR open source specification [11] it helps research workers both from information and medical sciences to understand the information model of the clinical document.

This archetype model itself is expressed in an Archetype Description Language (ADL) that is based on XML. This model is inherently extensible and scalable with respect to the semantic context we want to include in clinical data exchange. It also allows representing any clinical document in an EU standard-based model that can be shared among application developers as well as used by GPs to exchange clinical data.

The Archetype model of CEN 13606 provides several features that are required for achieving semantic interoperability. For example, it allows imposing constraints on the range of attributes of primitive types, constraints on the cardinalities of these attributes as well as introducing constraints on complex objects. The ontology section (see Fig. 6) of that model allows describing and binding information entities to clinical terminologies like SNOMED CT. For example, the entries of MainDiag and Diag in AmbList are bound to the ICD-10 terminology. It helps to embed semantic context in easy to discover patterns related to Procedures ENTRY concepts of health-related activities (SECTION-PROCEDURE) or to Observable ENTRY concepts (SECTION-ACTION), where specific values identify a health-related finding of the GP.

The Archetype model of the clinical documents has been exported in a W3C XML schema. It allowed managing instances of this model in a uniform way on a noSQL database like eXist-db [24]. For testing purposes, the database was installed locally, however, without any limitations it can be managed remotely or in
a cloud infrastructure [24, 25, 26]. Clients execute XQuery requests on instances of the Archetype model by means of RESTful web services. This allows exchanging the semantic context belonging to heterogeneous repositories of clinical data for example in cross-border exchange of personal health records [27].

Fig. 6. CEN 13606 Archetype model of an Ambulatory List.
7 Conclusion

Computer systems in Bulgarian healthcare generate and record large volumes of clinical data in electronic format. This data is a valuable resource of information for improving the quality of the health care services. The objective of this paper is to propose a common methodology for transforming existing information from XML documents exchanged in the healthcare domain making use of the Archetype model of CEN 13606. As a motivation case study, we consider the clinical documents generated in a GP practice. First, the global elements in the XSD document are converted to Java classes. Next, the Reference model of CEN 13606 is extended with these classes by preserving the original structure of core classes and their relationships in the international standard. Further, on we build an Archetype model of the Ambulatory list prepared by GPs. The architecture of concepts in this model is aligned with the logic of the business process followed by a GP in his practice and the semantic context is enriched by binding the model entries to clinical terminologies. The proposed methodology aims to overcome the currently limited scope of centralized exchange of clinical records and provide means for introducing modern approaches for computer assisted knowledge management in healthcare through semantic interoperability. The UML model for transformation of information structures is the result of the combined research efforts of a team of researchers combining professional experience both in medicine and computer science. The conclusions are based on original documents used in the clinical practice and best practices reported in the existing literature sources. The results are obtained as part of an eHealth scientific program with the combined research efforts of a team of researchers having professional experience both in medicine and computer science. It allows building a prototype of an information system that enables semantic interoperability among clients in terms of RESTful web services provided by a NoSQL database.

Acknowledgment

This research is supported by the National Scientific Program eHealth in Bulgaria.

References


