Knowledge Management of Controlled Vocabularies for Semantic Interoperability of Healthcare Applications

Joyce Rocha de Matos Nogueira
MLHIM Technological Development Unit
Rio de Janeiro State University
Rio de Janeiro, Brazil
joyce.nogueira@mlhim.org

Nathália Cristina Laurindo de Oliveira Ahiadzro
Department of Healthcare Information Technology
Rio de Janeiro State University
Rio de Janeiro, Brazil
nathcris87@yahoo.com.br

Luciana Tricai Cavalini
Department of Healthcare Information Technology
Rio de Janeiro State University
Rio de Janeiro, Brazil
luciana.cavalini@uerj.br

Timothy Wayne Cook
Emergent Group for Research and Innovation on Healthcare Information Technology
Rio de Janeiro State University
Rio de Janeiro, Brazil
tim@mlhim.org

***Abstract*— Controlled vocabularies such as terminologies and ontologies are regarded as a solution to the achievement of semantic interoperability across healthcare applications. This is yet to be attained. The semantics of current healthcare applications are directly in the source code and database structure and therefore cannot be readily shared. The multilevel model-driven approach is a method to produce semantically interoperable healthcare applications by providing sharable concept models. At this point, the methodology is proven, but implementations are still required. This study presents the fundamentals of knowledge modeling of controlled vocabularies for a XML-based multilevel model specification. The term subset 'Tuberculosis' of the 10th Revision of the International Classification of Diseases (ICD-10) is modeled as constraints to the XML Schema that defines the Reference Model. The Brazilian Mortality and Hospital Information Systems provided sample data for demonstration. The demonstration of the semantic validation of the XML data instances that include the ICD-10 'Tuberculosis' term set, converted to the multilevel model generated, is presented.**

Keywords*— *health information systems; information exchange; vocabulary, controlled.

I. INTRODUCTION

The complexity of healthcare information is due to the dynamic nature of healthcare systems and biomedical science. This creates challenges for the achievement of semantic interoperability and the maintenance of healthcare information systems [1]. An additional challenge is the approach used by healthcare domain experts to interpret and express healthcare concepts. These expressions can vary based on culture, geographical location and educational background [2].

The variation in expression, interpretation and application of healthcare concepts creates an environment where software is not easily deployed across multiple locations, as is the case with other industrialized applications. The evidence for this is seen as user dissatisfaction, typically expressed in surveys as complaints about inadequate workflow or user interface design, although this same application could be better rated in one or more particular installations [3].

Thus, every healthcare IT installation appears to be unique in order to satisfy the needs of the sub-domain and the user specific workflow. This naturally creates interoperability issues between these applications [4].

The enormous amount of concepts in biomedical sciences (something in excess of 3,000,000 terms) and the difficulty of reaching consensus among experts about the representation of medical knowledge results in extreme variability of the system requirements. This is true even within the same level of complexity of care. Some of the reasons for this extreme complexity lay in the fact that the route of each patient through various healthcare settings is not consistent with every other patient, and the outcomes of any individual encounter may affect the outcome of the following. This creates the need for a layer for semantic interoperability between the various information systems; a necessity for the health informatics field [5] and especially for automated clinical decision support (CDS).

One of the most common solutions proposed in the scientific literature for obtaining semantic interoperability among healthcare applications is the adoption of controlled vocabularies, primarily terminologies, classifications and ontologies [6]. Controlled vocabularies play an essential role in

This study was partially funded by the Carlos Chagas Foundation for Research Support in Rio de Janeiro (Faperj), Grants no. E-26 010.001825/2014, E-26 110.144/2014, E-26 200.679/2014 and E-26 201.527/2014) and by the Rio de Janeiro State University – Subpresidence of Research (Proatec Program) and Innovation Agency (Qualitec Program).

978-1-4673-9548-9/15 $31.00 © 2015 IEEE
DOI 10.1109/ICHI.2015.72

the standardization of medical knowledge representation models and many projects have demonstrated some success with this approach [7].

There is evidence that the isolated adoption of terminologies, classifications or ontologies has not been effective in delivering the desired ability to communicate semantically valid extracts of information between independently developed, distributed applications [8]. In order to effectively communicate, both structure and semantics must be decidable between the communicating parties.

Uptake of controlled vocabularies has not been followed by enough reports on the successful exchange of semantically coherent extracts of information between different applications. The few reports of successful implementation were in extremely controlled situations. This has not led to any significant implementations for situations that are typically found in the reality of the healthcare systems [9].

The World Health Organization (WHO) has been developing and maintaining the International Classification of Diseases (ICD) for almost a century [10]. ICD is now in its 10th Revision, released in 1990. All WHO Member States use the ICD and it has been translated into 43 languages. Most countries (117) use the system to report mortality data, a primary health indicator [11].

The ICD is essentially a nested hierarchical listing of coded terms. The codes allow for concise representation of specific concepts agreed upon by consensus of a group of international experts. This structure is very useful for computerization of healthcare data stored in SQL databases [12]. The negative aspect of this structure is that it does not easily expand and encompass new concepts, which results in several issues related to the validity and reliability of diagnostic information recorded according to the ICD terms [13].

In order to provide a sustainable solution for semantic interoperability, the concept of multilevel model-driven (MMD) approach introduced in the late 20th century by the openEHR Foundation [14] has been adapted for compliance with Semantic Web technologies and the Internet of Things by the implementation of the MMD principles in XML technologies [15].

The XML-based MMD approach is proven effective to provide semantic interoperability [15], but implementations are needed to increase its uptake. This study has the objective to demonstrate the implementation of a controlled vocabulary and the semantic validation of its data extracts according to this technology.

II. Methods

The Multilevel Healthcare Information Modeling (MLHIM) specifications were adopted, since they are the existing XML-based MMD implementation. The MLHIM Reference and Domain Models are formalized in XML Schema Definition files. The Concept Constraint Definitions (CCDs) are the expression of the Domain Models, being fully transportable across healthcare applications, so any instance data created according to a CCD will always be interpreted correctly [17]. The Reference Model schema provides a common structural reference for all concepts for all applications.

Two anonymized public databases managed by the Brazilian Ministry of Health, the Mortality Information System (MIS) and the Hospital Information System (HIS) were selected for the development of the use case presented in this study because both rely on ICD-10 for diagnostic record coding. The latest year of available mortality data was 2012, so this year was used to migrate the original data models to CCDs.

The CCDs for "Brazilian Death Certificate" and "Brazilian Public Hospital Discharge Summary" were modeled by the use of the Concept Constraint Definition Generator (CCD-Gen) (www.ccdgen.com), a web-based CCD editor. Each variable of both systems were modeled as a Pluggable complexType (PcT). A PcT is the fragment of XML Schema that represents each specific data component of a given medical concept. According to the specifications, the 'dv' element of a PcT has to be restricted to one of the defined data types; for the representation of terms from a controlled vocabulary, the DvCodedStringType is the proper choice. DvCodedStringType is constructed of required and optional elements. The required elements define name of the data component; the source vocabulary name, abbreviation and version; and the list of codes and strings (data values) that compose a given term subset.

For MIS, ICD-10 was defined as the reference terminology for the PcTs 'Cause of Death: Line A', 'Cause of Death: Line B', 'Cause of Death: Line C', 'Cause of Death: Line D', 'Cause of Death: Line II (a)', 'Cause of Death: Line II (b)', 'Underlying Cause of Death (Pre-Investigation)' and 'Underlying Cause of Death (Post-Investigation)'. For HIS, the same procedure was adopted for the PcTs 'Primary Diagnosis' and 'Secondary Diagnosis'.

For demonstration purposes, the CCD instances produced by each CCD were used as a model to convert the original MIS and HIS databases to XML files, persisted in an eXist-DB XML database. The ICD-10 'Tuberculosis' diagnostic group was chosen for implementation of the instances, since it is a well-defined, complete ICD-10 term subset, widely used in morbidity and mortality information systems. The English version of the ICD-10 code database was the source of data.

Two independent XML parsers and validators, Xerces and SaxonEE, via the oXygen version 14.2 editor, were used to execute the proof of validation chain of the simulated data instances: from the XML data instances to the CCD and from the CCD to the MLHIM RM Schema. Finally, the MLHIM RM Schema was validated against the W3C XML Schema 1.1 specifications.

III. Results

The CCDs for "Brazilian Death Certificate" and "Brazilian Public Hospital Discharge Summary" produced two separate CCDs. The MIS CCD can be downloaded at https://github.com/mlhim/tb/tree/master/src/CCD_45b4e463-9efe-4724-a5e5-0c3bc7143991, and the HIS CCD at

https://github.com/mlhim/tb/tree/master/src/CCD_6e5d7f04-4071-494c-a308-b07304b15f11. Figures 1 and 2 show respectively the Brazilian Death Certificate and the Brazilian Public Hospital Discharge Summary CCD headers in XML Schema.

Fig. 1. Brazilian Death Certificade CCD header in XML Schema.

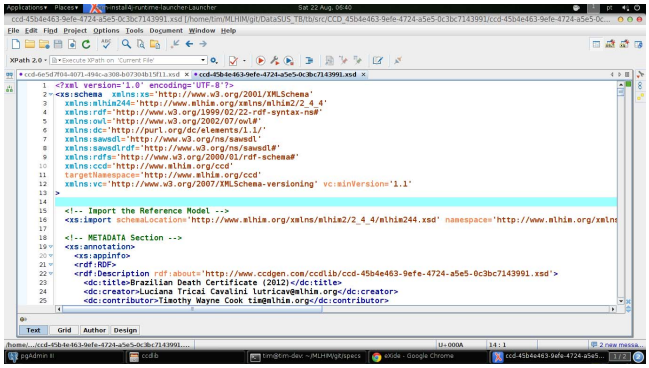
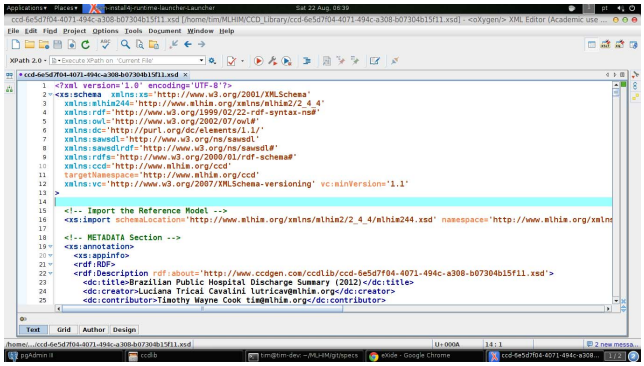


Fig. 2. Brazilian Public Health Discharge Summary CCD header in XML Schema.

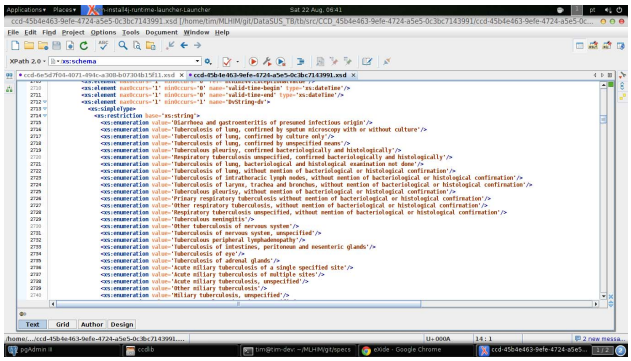


For both CCDs, the values of the 'terminology_name', 'terminology_abbrev' and 'terminology_version' elements were respectively fixed as 'International Statistical Classification of Diseases and Related Health Problems 10th Revision', 'ICD-10' and '2010'. The 'terminology_code' and 'dv' element values of the selected subset were defined according to the 4-digit ICD-10 codes and their respective term definitions. Table I shows the Names and functions of the required elements of a DvCodedStringType, and Figure 3 shows the representation of the 4-digit ICD-10 term definitions of the tuberculosis codes as enumerations of the DvCodedStringType 'dv' elements.

TABLE I. DVCODEDSTRING TYPE REQUIRED ELEMENTS

Element name	Element function
terminology_name	Full name of the controlled vocabulary
terminology_abbrev	Abbreviation of the controlled vocabulary
terminology_version	Release or version identifier
terminology_code	Uniquely identifiable code string from the controlled vocabulary
dv	Term description associated to a given code string from

Fig. 3. Brazilian Death Certificade CCD header in XML Schema.



All data instances produced by this method were semantically validated according to the correspondent CCD validation chain described above. This was repeated for all data instances, with a success rate of 100%. The resulting simulated databases can be downloaded at https://github.com/mlhim/tb. Figure 4 shows a validated data instance (green square on the upper left of the oXygen interface generated by the application.

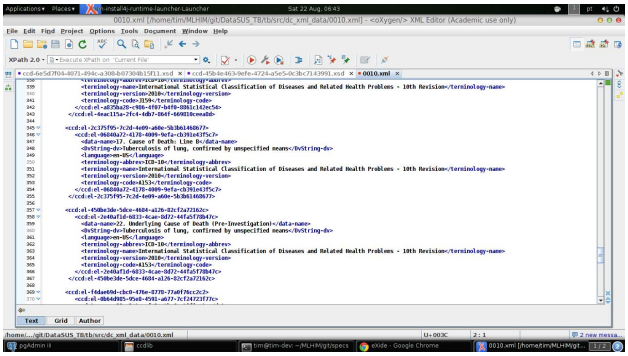
IV. DISCUSSION

This paper shows, as one example of implementation of MMD, the conversion of an ICD-10 term subset into the XML-based MMD technologies. The technical procedure adopted for this process was relatively easy to implement, since the CCD-Gen automates the code generation of the CCDs.

For implementation purposes, the selected term subsets should be modeled into specific PcTs that will be combined into higher-level concept structures called Clusters. The procedure requires the inclusion of the contents of the selected PcTs as a complexType of the ClusterType, each PcT being identified by a Universal Unique Identifier (UUID). Finally, the Cluster is defined as a complexType of one of the children EntryTypes (DemographicEntryType, AdminEntryType or

CareEntryType), which constrains the ‘definition’ element of the CCDType.

Fig. 4. Data instance validated according to the CCD and the MLHIM Reference Model schemas.



The CCD modelling is concluded with the definition of the Resource Description Framework (RDF) annotation of each complexType, when applicable, as well as the required and optional Dublin Core Metadata Initiative (DCMI) information that semantically defines the healthcare concept. The RDF provides additional semantics to each element based on the modeler knowledge and intent in building the models. This allows the end user of the data to understand the meaning of the data models and therefore the compliant data.

The representation of term subsets of the ICD-10 classification or any other controlled vocabulary in XML MMD-based applications can be achieved by the generalization of the procedure here described for the ‘Tuberculosis’ ICD-10 term subset. The CCD provides for the possibility of defining data models of any size or complexity. Since the MLHIM eco-system allows for any number of CCDs per healthcare concept, this enables the implementation of CCDs for the representation of specific subsets of all types of controlled vocabularies, without the need for the acquisition of costly and complex terminology services. CCDs also allow for embedded business rules that can be applied to one PcT or across PcTs in a CCD. This provides for any level of complexity to be contained in a semantically coherent, shareable module to be used in any size healthcare application from small devices to enterprise EHRs.

V. Conclusion

The adoption of controlled vocabularies has been an important asset towards the standardization of healthcare information systems. However, the adoption of such knowledge representation models has not been enough yet to provide full semantic interoperability between healthcare information systems. The innovative approach proposed by multilevel modeling technologies is promising, and it is friendly to controlled vocabularies of all types. The case study presented in this study can be generalized to larger and multiple terminology sets, according to the specific application requirements, without losing the semantics of healthcare information from the source data. It is expected that, in a near future, the wider adoption of multilevel modeling technologies as a basis for healthcare applications, in combination with controlled vocabularies whenever required, will become an important asset for the health informatics field to overcome its current challenges, thus enabling the development of patient-centered, longitudinal healthcare information ecosystems. The primary challenge is that MMD modeling does require a new approach to modeling concepts.

Acknowledgment

The authors thank Liana Poggian Braz e Alípio Neto do Nascimento Carvalho for operating the CCD-Gen in part of the modeling process.

References

[1] M. Martínez-García, and E. Hernández-Lemus, “Health systems as complex systems,” Am. J. Operat. Res., vol 3, pp. 113–126, January 2013.

[2] S. de Lusignan, C. Pearce, N. T. Shaw, S. T. Liaw, G. Michalakidis, M. T. Vicente, et al., “What are the barriers to conducting international research using routinely collected primary care data?,” Stud. Health Technol. Inform., vol. 165, pp. 135 – 140, 2011.

[3] D. B. Meinert, “Resistance to Electronic Medical Records (EMRs): a barrier to improved quality of care,” Issues Inform. Sci. Inform. Technol., vol. 2, pp. 494–504, 2005.

[4] J. S., Ash, and D. W. Bates, “Factors and forces affecting EHR system adoption: report of a 2004 ACMI discussion,” J. Am. Med. Inform. Assoc., vol. 12, pp. 8–12, january-February 2005.

[5] B. Kaplan, and K. D. Harris-Salamone, “Health IT success and failure: recommendations from literature and an AMIA workshop,” J. Am. Med. Inform. Assoc. vol. 16, pp. 291–299, May-June 2009.

[6] P. Ganguly, P. Ray, and N. Parameswaran, “Semantic interoperability in telemedicine through ontology-driven services,” Telemed J. E-Health, vol. 11, pp. 405–412, June 2005.

[7] P. Knaup, O. Bott, C. Kohl, C. Lovis, and S. Garde, “Electronic patient records: moving from islands and bridges towards electronic health records for continuity of care,” Yearb. Med. Inform., vol. 16, pp. 34–46, 2007.

[8] D. Kalra, and B. G. Blobel, “Semantic interoperability of EHR systems,” Stud. Health Technol. Inform., vol. 127, pp. 231–245, 2007.

[9] G. A. Lewis, E. Morris, S. Simanta, and L. Wrage, “Why standards are not enough to guarantee end-to-end interoperability,” 7th Intern. Conf. Composit. Bas. Softw. Syst., pp. 164–173, February 2008.

[10] R. Laurenti, “Analysis of information on health data: 1893-1993, a hundred years of the International Classification of Diseases,” Rev. Saúde Pública, vol. 25, pp. 407–417, December 1991.

[11] G. Rey, A. Aouba, G. Pavillon, R. Hoffman, I. Plug, R. Westerling, et al., “Cause-specific mortality time series analysis: a general method to detect and correct for abrupt data production changes,” Popul. Health Metr., vol. 9, pp. 1–11, September 2011.

[12] World Health Organization, International statistical classification of diseases and related health problems - 10th revision, Geneva: World Health Organization, 2010.

[13] G. Surján, “Questions on validity of International Classification of Diseases-coded diagnoses,” Int. J. Med. Inform., vol. 54, pp. 77–95, May 1999.

[14] D. Kalra, T. Beale, and S. Heard, “The openEHR Foundation,” Stud. Health Technol. Inform., vol. 115, pp. 153–173, 2005.

[15] L. T. Cavalini, and T. W. Cook, “Use of XML Schema Definition for the development of semantically interoperable healthcare applications,” Lect. Notes Comput. Sci., vol. 8315, pp. 125–145, 2014.