Towards An XML-Based Normalization for Healthcare Data Exchanges

I. El Azami, M.M.O. Cherkaoui and C. Tahon

1Department of Math and Informatics, Sidi Mohammed Ben Abdellah University, FSDM, BP 1796, Fez, Morocco
2TEMPO Laboratory, University of Valenciennes and Hainaux-Combrésis, PSI, Le Mont Houy 59313 Valenciennes Cedex 9, France

Abstract: Problem statement: Most of healthcare data exchanges are textual, poorly structured and often not accessible by clinical professionals. Otherwise, the variety of medical applications and medical standards makes difficult sharing and communicating healthcare data in a highly heterogeneous environment. Approach: XML and related standards (XML schema, XSL) provided an infrastructure that might change the situation. Our aim in this study was to define an exchange model providing a common structure of shared healthcare data to allow a better, easier and structured communication within and between hospital information systems. Results: We realized an XML-based model that we detailed the content and the structure. Thus, seen the confidential character of healthcare data, we described an approach to secure the data transfer. We were situated regarding existing models and standards such as HL7, DICOM and the PMSI and we took into account critics made for them. Conclusion/Recommendations: The model we proposed provide a practical solution allowing a secure and structured healthcare data exchange and will serve as a summary version of the common computerized patient record.

Key words: Hospital information system, healthcare data, standards, XML

INTRODUCTION

The patient care process involves many actors, with very different statuses and contains numerous actions potentially realized on a larger area. Otherwise, the ethical and legal constraints associated to the processing and dissemination of patient’s data are important. The needs for communication among health professionals who work cooperatively to support the patient, within the health establishment but also between different health organizations, makes necessary the opening of hospital information systems and require the definition of an exchange model of healthcare data. The definition of a common structure of shared healthcare data is a concept that will allow easier opening of hospital information systems and provide an infrastructure for interoperability and standardization.

A Hospital Information System can be defined as a system to facilitate the management of all medical and administrative data of a hospital and care quality. This system summarizes, among others, the essential data needed for a medico-economic evaluation. Management of healthcare data includes all functions related to the care of a given patient and medical activities in general. These data is obviously very complex, very eclectic and therefore very difficult to model. The patient record is therefore the physical memory which registers all necessary data to support and monitor a patient. Data stored in the patient records are increasingly complex, due to the emergence of new investigations: Radiological, biological. Now, a reflection on healthcare data and its management in hospitals is urgent. Like modern hospitals in industrialized countries, a HIS will find its place and it would be a valuable tool for better management and real benefit to the healthcare data.

This study discusses medical and administrative data which are essential for communication within hospital establishments. The aim is to develop a solution to make this information accessible and shareable in large scale. We are providing an XML-based model describing, at once, the content and structure of data. This will reduce ambiguity and enhance knowledge transfer while preserving the semantics and optimizing the flow of data.

Corresponding Author: I. El Azami, Department of Math and Informatics, Sidi Mohammed Ben Abdellah University, FSDM, BP 1796, Fez, Morocco
A “canonical” (Cherkaoui et al., 2008) and standard representation of the model is essential. The definition of this representation should be simple, secure and complete. Simplicity: The representation is as simple as possible reading and use. Secure: Healthcare data must preserve its secret and confidential character. Completeness: No loss of information must occur between healthcare source activity and its canonical representation; all the basic information and summary information, presenting a criterion of relevance, must be preserved.

MATERIALS AND METHODS

Data exchange via XML: In the hospital sector as in many areas, many applications already exist and have difficulty sharing their data. For this, we can in simple cases work base to base. The concept of Extract, Load and Transform (ETL) has been historically introduced. Using XML as the exchange model is of great interest (George, 2001), to facilitate processing, validate data, integrate the textual aspect and manage metadata.

Beyond ETL, the Enterprise Application Integrator (EAI) emerged as a more general and more attractive oriented direct exchange or through a database. The heart of EAI is based on XML, a coordinator performing routing and an XSL transformation engine often based on XSL (Fig. 1). Application data (messages, files, table) are first converted to XML by adapters (El Azami et al., 2007), also called connector. They are then processed and stored in database (exchange asynchronous) or sent to the target (synchronous exchange). These systems are currently in full development.

The integration of XML as a medium of exchange in the heart of EAI reduces the number of connectors required (with N sizes, just N connectors and not N² as with changes in format-format), to systematize the supervision tools, to better secure messages and finally based on recognized standards (W3C, 2010). These advantages are crucial.

More than a metalanguage, XML is actually a galaxy with technological languages derived for schemas (XML Schema), links (XLink), stylesheets (XSL), interfaces of object-oriented programming (DOM) or event (SAX), generators forms (XForms), coding object model (XMI). Parsers and XSL processors are fundamental tools that make XML more than a tool for representing data, a basic development environment.

The sharing of data between distributed application modules via XML will allow a proper independence of modules and rigorous structuring of communications. Thus, the marriage of XML and databases is quickly becoming the key to success and openness to information systems (Gardarin, 2002).

Towards an XML-based model to represent and exchange healthcare data:

Which data to exchange? Data contained in the healthcare information system is used by many different units within the hospital. It covers often patient data, practices and information on available resources. They are also useful for administrators who analyze the performance of services and establishment as a whole. Of course, these data may be used for research purposes or social development.

Healthcare data has no medical sense by itself but according to a context (Charlet et al., 2002). Consequently, only data classified “relevant” is intended to be exchanged and communicated (Cherkaoui et al., 2008), we mean by relevant data, a data:

- Associated to its production context
- Deemed useful
- And potentially reusable for the patient, for the hospital, for management purposes, research or social development

Overview of the existing:

XML in healthcare: During the past few years, XML has been introduced into the healthcare industry and is now being widely used. Applications of XML in healthcare have a wide range including academic studies that contain descriptions of the contents of heart sound components (Modegi, 2001), representations of models
of biochemical reaction networks (Hucka et al., 2003), ADE reports (Kataoka et al., 2002) and pharmaceutical inquiries (Dugas et al., 2003). XML is also entering the medical domain through standards activities that include inter-system messaging (HL7 3.0), the Clinical Document Architecture (CDA), as well as knowledge representations, including clinical guidelines (Ganslandt et al., 2002; Dart et al., 2008) and the Arden Syntax.

**PMSI:** After the adoption of “Programme de Medicalisation des Systems d’Information” (PMSI) in France, the public and private health institutions need to analyze their business and provide to medical services of the state and Health Insurance data relating to their means of work and their activities. For this end, they must implement information systems that take account of care conditions and modes.

For hospital stays in acute care-Medicine, Surgery, Obstetrics and dentistry (MCO)-this analysis is based on the systematic collection of a limited number of administrative and medical information (Table 1), which constitute «le Resume de Sortie Standardise (RSS)».

Any hospital stay in the MCO of public or private health institution should lead to the production of a Standardized Summary Release (RSS), consisting of one or several Summaries of Medical Unit (RUM).

The exchange of these relevant data between health establishment and the authority will evaluate the healthcare activity of these institutions and thus help the authority to make strategic decisions at the level of funding hospitals on one hand and provide key factors for improving quality of care on other.

However, this exchange is done traditionally by unstructured text file and makes complicated the automated processing of shared data. A shift to XML format seems very beneficial. On one hand, to allow better integration with existing and modern HIS and secondly, to allow standardization of representation and an automated processing of data exchanged exploiting the galaxy of XML standards (XML, XSL, XPath, XForm).

**HL7 standard:** There were several alternative solutions and standards to enable data exchange between different healthcare information systems. HL7 (Health Level Seven) seems the most famous standard for the exchange of medical and administrative data; it defined a reference model (RIM) that the structure of XML messages exchanged is based on.

However, HL7, in its third version (HL7 v.3), was widely criticized by professionals and health practitioners. On one hand, its XML messages are super complicated and introduce a large number of codes, their understanding is far from obvious by uninitiated users (Aerts, 2008). On the other hand, the lack of clarity in the definition of ‘Act’ Class of RIM presents an ambiguity of interpretation and resulting semantic and ontological confusion (Browne, 2008). Thus, the lack of representation of some basic concepts in medical field namely diseases, drug interactions, injuries, organs and others signals a remarkable failure and relies on questions about the scope of its reference model (Werner, 2009; Barry and Werner, 2006). Moreover, it is accused to HL7, the difficulty of learning and its implementation in information systems of healthcare organizations which affects its marketing (Barry and Werner, 2006).

**DICOM standard:** The field of Digital Imaging and Communications in Medicine (DICOM) is medical imaging. Today, medical image means the image and its environment grouped together in what we call the act of medical imaging. The life cycle of this act begins with drafting the request for review by the clinician and ends with one hand shipment of the results to the clinician and storage on other. Throughout this life cycle, there should be interaction between the information system of health and equipment acquisition and image processing.

| Table 1: The RUM content | Medical data |
|--------------------------|-------------|
| Administrative data      | Medical data |
| Identifiers corresponding to the entire stay in the MCO units of establishment: | Diagnosis: primary diagnosis, diagnostic related, significant diagnoses associated |
| Number of local administrative stay and number of RSS | Medical activities |
| Number of the institution in the National register of health and social institutions (Finess) | Weight at entry into the medical unit for newborns |
| Number of medical unit | Gestational age for the mother and newborn |
| Type approval of the medical unit | Simplified Acute Physiology Score (SAPS II) |
| Sex | Documentary data associated |
| Date of birth | |
| Zip code of residence | |
| Dates and modes of entry and exit, from and destination | |
| Number of meetings | |

802
DICOM’s approach is that of managing a real case of medical imaging seen as a subfolder of the medical record itself subset of patient records. This need to retrieve the information system work lists and administrative data about the patient and subsequently leads to:

- Communicate to the information system all the facts concerning the realization of the examination (examination started, finished, type of examination performed, delivered dose, contrast injection)
- Engage with the archiving and information system to ensure proper archiving of images and informs location to the information system

The question of the role of XML is very open in the DICOM Committee, in particular about structured reports. In the DICOM Committee proponents of the specific syntax of origin will be difficult to stay on “hard” line to preserve the exclusivity of the original syntax in DICOM, especially for the representation of documents as records to circulate throughout the hospital (Gmsih, 2001). The definition of an XML model is therefore of great importance for the representation and communication of DICOM records.

Summary: The emergence of these standards and many others comes to fill a need for standardization of healthcare data representation and its sharing. But, when we have a multitude of views and approaches, we inherit the initial problem of heterogeneity.

Here we propose an alternative for normalization, trying to align the made efforts in the prospect of having a common pivot model. The model inherits of common data of most used standards in hospital sector and additional data necessary for communication, while respecting our approach to include only information deemed relevant.

Thus, we took into account security of data, the future evolution of the model, its readability and ease of use by health practitioners and information systems in order to avoid disadvantages similar to those alleged in the HL7 standard for example. The model is based on XML recognized as the international standard of exchange.

RESULTS AND DISCUSSION

Elements of the model: We have identified five major classes of information, namely information relating to the patient, the history, documents, medications and medical practices. Below is a list of relevant information of each of these classes.

**Patient:** Identifiers, first name and last name, sex, date of birth, address, Weight at entry into the medical unit for newborns, gestational age for mother and newborn.

**History:** Medical history, surgical history, Gynecological and obstetric history, Family history.

**Practices:** Institution number, Number of medical unit, Type approval of the medical unit, Dates and modes of entry and exit, from and destination, Number of meetings, Simplified Acute Physiology Score (SAPS II), Diagnosis: Primary Diagnosis (PD), significant Associated Diagnosis (DA), Medical activities.

**Medications:** Name of drug, description, dosage, administration, Drug interactions.

**Documents:** Medical images, results of diagnostic and biological tests, reports.

The XML model: Here we discuss our XML representation related to the listed data.

**The patient:** A unique identification is provided for each patient. Work has been initiated under the IHE (Integrating the healthcare enterprise) to harmonize the identification of patients between HL7 and DICOM (Fig. 2). Thus, HL7 collaborates with the ISO/TC 215 for the definition of an international policy for identifying patients (Gmsih, 2001; 2002a; 2002b). The combination of several items (id, name, date of birth) is increasingly recommended by the standards in relation (Gmsih, 2002b). However, we acknowledge the possibility of having multiple identifiers for a patient (local id, RSS id, approximation id (Gmsih, 2001a).

**Histories:** In every medical intervention, consultation of the patient’s history is inevitable. Decisions made by the practitioners are, in most cases, directed by medical, surgical, gynecological and obstetric, psychiatric or family prior of patient (Fig. 3). Patient’s histories have been sufficiently discussed in the efforts of the Working Groupe de travail C HL7 France-HPRIM (2005). We adopt a structure that contains 3 items: “Pathology”, “Date of occurrence” and “Comment”. In the case of family history, the “relationship” should be mentioned.

```
<patient>
  <identifiants>
    <id type="" valeur=""> id </id>
    <nom><nom>
    <prenom><prenom>
    <sex><sex>
    <dateNaissance><dateNaissance>
    <adresse><adresse>
    <poidsNouveauNé><poidsNouveauNé>
    <AgeGestational><AgeGestational>
  </patient>
```

Fig. 2: XML section describing patient data
Practices: Information structuring on medical practices is based on our approach describing the medical activity in the form of a tree (Cherkaoui et al., 2008), we defined a design pattern (Gamma, 1995), which we named “Medical Activity Pattern”, dedicated to this assumption, the root of the tree is the input activity (Fig. 4), nodes are related sub activities and leaves of the tree are the medical actions that can’t be subdivided into sub activities (Cherkaoui et al., 2008). We recommend the CCAM coding (Classification Commune des Actes Medicaux) for activities and the ICD-10 coding (International Classification of Diseases 10) for diagnosis.

Medications Knowing exactly which medications and regimens patients use can help physicians to avoid drug interactions, manage side effects and more effectively direct the patient’s treatment (Staroselsky et al., 2008).

Fig. 4: XML section describing practices data

Fig. 5: XML section describing medications data

Medication lists represent one of the most important components of the electronic health records since they are used for filling refill requests, assessing quality, performing research and for informing computerized clinical decision support. As Wagner and Hogan (1996) point out, it is especially important to maintain accurate structured lists in the presence of automated decision support because medication information presented in free-text format or in any other non-standard part of the medical record would be unreadable and unusable by the automated decision support system, resulting in loss of many potential benefits of the system (Wagner and Hogan, 1996). A section describing structured information about patient’s medications is included in our model (Fig. 5).

Documents: The advantage of XML is that it allows representation of structured information including both the content (text, image) and description.
Most medical documents have some structure (Fig. 6), XML allows to represent them in a standard way (Leventhal et al., 1998; Harold, 1998). Besides, this is the reason why the HL7 CDA standard adopted XML in its work on medical documents. Similarly, the DICOM standard is moving towards this direction especially for its purpose DICOM Structured Reporting (SR) (Dart et al., 2008). The data to be communicated about medical documents is described below.

The definition of our data model and its structure is given by a DTD. Here we show the Patient and Practices parts of the DTD (Fig. 7).

**Data security:** Seen the secret character of the healthcare data and to protect all which is communicated within the framework of the medical relation, a security approach of the XML data exchanged is indispensable.

On this matter, several studies were led to give to the developer a granular control on the XML contents; eXtensible Access Control Markup Language (XACML) is a specification of OASIS, it supplies a means to standardize the decisions of access control for XML documents. The XACML specification allows determining if it is necessary to authorize the access asked for a resource, if it is about a part of a document, about a document in its entirety, or about several documents. We find also The XML signatures (XML-SIG) which are strictly connected to the encoding (W3C, 2002a).

Besides being able to use standard methods of encoding during the transmission of documents XML, the W3C and the IETF propose a standard of encoding XML data and tags within a document (Xenc) (W3C, 2002b). It would so allow coding various parts of a document, the idea being that only the sensitive information must be protected. So, the encoding of certain parts of a document by means of various keys would allow passing on the same XML document to diverse addresses, these last ones can decipher only the parts concerning them.

```
<documents>
<document>
<code><code>
<type><type>
<auteur><auteur>
<dateProduction><dateProduction>
<descriptionContent><descriptionContent>
<observations>
<observation><observation>
<observations>
</document>
</documents>
```

Fig. 6: XML section describing documents data

Fig. 7: Patient and practices DTD parts of the XML model

```
<patient>
<identifiants
<id type="IdLocal" value="20934235">"
<identifiants
<nom>Dupent</nom>
<prenom>Eric<prenom>
<sex>M<sex>
<dateNaissance>13.05.1955<dateNaissance>
<adresse>35 Place du parc<adresse>
</patient>
```

Fig. 8: Example of patient’s data

When an XML document is coded by this method, a tag indicating the beginning and the end of the coded data appears in the document. It is defined by the element <EncryptedData>, which relates to the domain name of the W3C encoding. The names of tags themselves are replaced by tags <CipherData> and <CipherValue>; the data are shown in the form of the encrypted channel which results from it. This proposed standard offers a level of granular control which allows the person who supplies the XML data to control their visibility according to the public target. Furthermore, as the data itself-but not the file-are coded, this one can be always recognized by XML parsers and treated consequently.

**Example:** Let’s take the following data about the patient ‘Dupent’ (Fig. 8).
Fig. 9: Patient's data after encoding (Xenc) of the "name"

If we realize that the "name" of the patient has to remain confidential, we are going to code the element <name> replacing it by the tag <EncryptedData> as the following way (Fig. 9).

**CONCLUSION**

The work presented in this study is intended to provide an alternative normalization of healthcare data exchange between hospital information systems. A model using the structure and the flexibility offered by XML has been given.

The proposed model combines the medical and administrative information relevant and necessary for communication. Useful data for decisions on public health but also for the evaluation of hospital activity as in the case of PMSI can be easily obtained by a simple XPath query on the XML document tree, which allows optimization research. Thus, the automated processing of these data using SAX and DOM will provide a significant improvement to PMSI.

In term of security, the encoding approach of XML tags (Xenc) which we adopted, in association with the use of XSL which will give the possibility of having customizable views according to the access privileges to data, shall allow to respect the ethical and legal constraints associated to treatment and secure broadcasting of healthcare data.

The fact of relying on the XML standard in our model, supports its integration into various information systems (existing and/or new) on one hand and facilitates its web implementation on the other. XML is a simple, very flexible text format which is playing an increasingly important role in the exchange of a wide variety of data on the Web and elsewhere. It shows promise as an interchange format for health information exchange.

Finally, this model will serve as a summary version of the common computerized patient record, implementing the portability of patient records and its sharing.

**REFERENCES**

Aerts, J., 2008. Ten good reasons why an HL7-XML message is not always the best solution as a format for a CDISC standard (and especially not for submission data), XML4Pharma, Version 0.5. http://www.xml4pharma.com/HL7-XML/HL7-XML_for_CDISC_Standards.pdf

Barry, S. and C. Werner, 2006. HL7 RIM: An incoherent standard. Stud. Health Technol. Inform., 2006; 124: 133-138. http://ontology.buffalo.edu/HL7/doublestandards.pdf

Browne, E., 2008. Shortcomings of HL7 Reference Information Model (RIM). http://www.hisa.org.au/system/files/u2233/Eric_Browne.pdf

Charlet, J., E. Cordonnier and B. Gibaud, 2002. Interopérabilité en médecine : quand le contenu interroge le contenant et l'organisation, Information-Interaction-intelligence, Vol. 2, n°2, Cepadues-Edition. http://www.revue-i3.org/volume02/numero02/article02_02_02.pdf

Cherkaoui, M.M.O., D. Doukkali, I. El Azami and C. Tahon, 2008. Communication de l’activité médicale entre SIH: Une représentation canonique et un modèle de patron. La Conférence Internationale GI$EH’08 Lausanne en Suisse, Sept. 4-6, EPFL, pp: 1-7. http://giseh08.epfl.ch/webdav/site/giseh08/users/112541/public/programme_detaille.pdf

Dart, T., Y. Xu, G. Chatellier and P. Degoulet, 2008. Computerization of guidelines: Towards a "guideline markup language. Stud. Health Technol. Inform., 84: 186-190. PMID: 11604730

Dugas, M., S. Weinzierl, A. Pecar, S. Endres and J. Hasford, 2003. Design and implementation of a common drug information database for a university hospital. Pharm. World Sci., 25: 156-161.

El Azami, I., D. Doukkali and O. Cherkaoui, 2007. Approach based on patterns for mediation between hospital information systems. Proceeding of the 33rd International Conference on Operational Research Applied to Health Services, July 15-20, Saint-Etienne, France, pp: 1-69. http://www.emse.fr/orahs/ORAHs_Program.pdf

Gamma, J.R., 1995. Design Patterns, Elements of Reusable Object-Oriented Software. Addison-Wesley, ISBN: 0201633612, pp: 416.
Ganslandt, T., M.L. Mueller, C.F. Kriegstein, N. Senninger and H.U. Prokosch, 2002. An XML-based system for the flexible classification and retrieval of clinical practice guidelines. Proc. AMIA Symp., 2002: 280-284. PMID: 12463831

Gardarin, G., 2002. XML: Des Bases de Données Aux Services Web. Dunod, Paris, ISBN: 13: 978-2100069330, pp: 536.

George, S., 2001. Medical support system for continuation of care based on XML web technology. Int. J. Med. Inform., 64: 385-400.

Gmsih, 2001. GMSIH: Les standards en Imagerie Médicale: (Série documents d’information), 1- DICOM: Digital Image Communication in Medicine, Version 1.0. http://www.gmsih.fr/fre/content/download/4397/44040/file/11.vi21dicominfo1.pdf

Gmsih, 2002a. Identification du patient: Principes de l’identification du patient, Tom 1: Domaine d’identification et Domaine de rapprochement, Tom 2: Modèle de rapprochement, version 1.1. http://www.gmsih.fr/fre/content/download/439/2810/file/30.principe1V1.1web.pdf

Gmsih, 2002b. Projet Si 1.1 Principes et Processus d’identification du patient. Synthèse de l’analyse des Normes et Standards relatifs à l’identification du patient, Version 1. http://www.gmsih.fr/fre/content/download/442/2822/file/34.si1.1norv1web.pdf

Groupe de travail C HL7 France-HPRIM, 2005. Guide d’implémentation du Volet Médical format CDA Release 2-Niveau 3. http://guepard.contingences.com/IMG/pdf/Guide_I mplementation_Volet_Medical_CDA_Release_2_ Niveau_3_v0-7.pdf

Harold, E.R., 1998. XML: Extensible Markup Language. Hungry Minds, Foster City, CA., ISBN: 0764531999, pp: 426.

Hucka, M., A. Finney, H.M. Sauro, H. Bolouri and J.C. Doyle et al., 2003. The systems biology markup language (SBML): A medium for representation and exchange of biochemical network models. Bioinformatics, 19: 524-531.

Kataoka, S., K. Ohe, M. Mochizuki, S. Ueda, 2002. Developing and integrating an adverse drug reaction reporting system with the hospital information system. Yakugaku Zasshi, 122: 113-116.

Leventhal M., D. Lewis and M. Fuchs, 1998. Designing XML Internet Applications. Prentice Hall, PTR Upper Saddle River, NJ., USA., ISBN: 0-13-616822-1, pp: 624.

Modegi, T., 2001. XML transcription method for biomedical acoustic signals. Stud. Health Technol. Inform., 84: 366-370. PMID: 11604765

Staroselsky, M., L.A. Volk, R. Tsurikova, L.P. Newmark and M. Lippincott et al., 2008. An effort to improve electronic health record medication list accuracy between visits: Patients’ and physicians’ response. Int. J. Med. Inform., 77: 153-160.

W3C, 2002a. XML-signature syntax and processing. http://www.w3.org/TR/2002/REC-xmldsig-core-20020212/

W3C, 2002b. XML encryption syntax and processing. http://www.w3.org/TR/2002/CR-xmlenc-core-20020304/

W3C, 2010. RIF standard supports data integration, enterprise agility. http://www.w3.org/

Wagner, M.M. and W.R. Hogan, 1996. The accuracy of medication data in an outpatient electronic medical record. J. Am. Med. Inform. Assoc., 3: 234-244.

Werner C., 2009. HL7 RIM: Still no coherent way to track concerns. http://hl7-watch.blogspot.com/2009/05/ hl7-rim-still-has-no-coherent-way-to.html