FIS 2005 **http://www.mdpi.org/fis2005**

Identification and Composition of Metadata for Cooperative Information Systems
Illustration on the health information systems

Gabriella SALZANO * **and Christian BOURRET**

Université de Marne-la-Vallée, Cité Descartes - 5 Bd Descartes
Champs-sur-Marne – 77454 Marne la Vallée Cedex 2, France

Author to whom correspondence should be addressed; E-mail: gabriella.salzano@univ-mlv.fr

Abstract

Networked health organizations require cooperative Information Systems (IS) to provide different information services. These services contribute to improving the quality of care and to reducing costs. Furthermore, they improve processes for both care delivery and public health. These services require the merging of knowledge from multiple health sub-domains (pathologies and drugs, for instance) or various other domains. Correlating health with geography, enables analyses in interaction with many other dimensions (social, economical, environmental …).
Information and communication chains are particularly complex for these IS, regardless of their geographical scope (local, regional, national or international).
We will adopt an IS perspective to focus on some problems where geographical aspects are essential. We then bring several elements to build an interoperability approach based on advanced methodologies and technologies.
We begin by recalling some dimensions of the heterogeneity problems and interoperability issues for complex systems (§1). Then, we analyze some examples of IS for health networked organizations, where the geographical dimension is structuring for different purposes (§2). Our approach to interoperability between heterogeneous IS is based on the identification and composition of appropriate content descriptive and domain dependent metadata, in accordance with the standardization frameworks on the involved domains (§3). It is supported by an interoperability infrastructure based on those technologies recommended by the W3C consortium (§4). In particular, Namespaces, Resource Description Framework (RDF), XML/RDF Schema and Web Ontology Language improve levels of knowledge representation and exchange across heterogeneous domains, and help generate information even if corresponding data are not explicitly stored.

Keywords: interoperability, semantic web, metadata, health information systems

1. Interoperability Dimensions

Sheth and Larson [1] analyze interoperability by using three fundamental dimensions, which are distribution, autonomy and heterogeneity:

- Distribution identifies the interdependencies and interactions between components; it is possible to decompose these relationships with respect to multiple axes (who, why, what, when, how, where) [2]
- Autonomy may exhibit several forms:
    - its design includes the choice of the domain[1] being managed and the conceptualization of the context
    - its association with other systems refers to the ability of a system or component to choose the components with which it shares its resources, as well the manner in which it does so
    - communication refers to the ability of a node (system or component) to communicate with the others
    - execution concerns the ability to execute local operations independently of the external components
- Heterogeneity arises at multiple layers ([3], [4]):
    - the system (for instance, the technical platforms, the database management systems, their capabilities)
    - the information itself, at different levels: syntax, structure and semantic

As these dimensions have to be considered simultaneously to build interoperability solutions for a set of heterogeneous distributed systems, interoperability is a very complex challenge.
Moreover, according to the IEEE Std. Dictionary [5], interoperability between two or more systems is defined as the ability of these systems to exchange information and to share the information that has been exchanged, through the use of a single set of rules. Two levels of interoperability emerge: on the one hand, the technical level concerns the communication and exchange of data; on the other hand, the semantic level is related to shared use of knowledge, and information exchanged from disparate systems.

In the following section, we analyze several examples of cooperative information systems where the health and geographical dimensions are strongly coupled, for different purposes.

2. The Coupling of Health and Geography Dimensions: Heterogeneity and Potential

A cooperative IS is confronted with the need to guarantee semantic interoperability between its components. Schematically, two larges classes of IS may be identified as transactional or decisional. They are so defined in accordance with their main objectives: to support the organizations in the care delivery process or in the decisional process.
We will illustrate how the heterogeneity between IS of the first class generates several conflicts that must be resolved in order to build IS in the second class. We also wish to demonstrate the potential of the decisional IS when health and geography are coupled.
Hospitals possess IS belonging to the transactional class of IS, just like other organizations in the health domain known as “health networks” [6].

---

[1] From now, the domain represents the data or information which has to be managed (i.e. the Universe of Discourse).

In France, the health networks have been recognized by the Law of march 2002 [7] as a privileged means to group and coordinate professional actors belonging to different healthcare structures (like general practitioners or primary care doctors, medical or social practices, specialized centers, peripheral hospitals and clinics), in specific geographical areas. To date, there are between 1000 and 2000 health networks, with different financial supports, administrative structures, institutional dependencies and territorial scopes (often local, with a potential departmental or regional extent). They can present inclusion criteria, with respect to pathologies (aids, diabetics, C hepatitis, asthma, cancer, …), populations (drug addicted, persons in precarious situations, palliative care, gerontology …), geographical areas (suburbs or rural areas, with very scattered populations) or intervention scenarios (general or specific: accident, emergency, …).
To reduce errors in high risk procedures, delays, duplicative or unnecessary acts, these IS aim at supporting an improved professionals' coordination, based on protocols and sharing information (health records). The Electronic Health Care Record (EHR) is a fundamental, structuring artifact to store, transmit, share and access information about individual patients.
A HER is made of several components and collects information which can be [8]:

- subjective, referring to the reason of the contact between patient and healthcare professional,
- objective, as clinical or laboratory exams,
- an assessment, like diagnosis,
- or a plan referring to the clinical actions that must be taken.

Therefore, data interpretation depends on the context (history, symptoms, diagnostic, treatments …).
If we except some compulsory elements, EHR content, form and support are free: many thousand of medical concepts are covered by different classifications, related to different areas, with overlapping domains, and having different structures.
HER components are structured or not, containing text, still or moving images produced by multiple systems.
Moreover, the regulation frame (legal, ethical, deontological…) for the whole management of personal and critical data (confidentiality, security, integrity...) differs strongly from a country to another.

In all developed countries, the evolution towards electronic support of patient records (starting from the paper support) is carried out gradually The French Hospitext project [9] was among the first to adopt a documentary approach to carrying out a computerised medical record. It made use of a hypertext documentary representation to navigate among its elements and produce a set of synthesis documents, in accordance with the medically standardized readings of the record.

At the international level, for instance, Canada placed the HNO among its national priorities, along with the *Réseau Canadien de la Santé* (Canadian Health Network) [10]. The United Kingdom' National Health Service (United Kingdom), has been deploying the project *Information for Health: An Information Strategy for the Modern NHS* since 1998, especially at the level of Primary Care Trusts [11].

The first addressed difficulty concerns the support of the health information exchanges. The idea is to avoid a situation in which healthcare parties manage an enormous number of different interfaces. Thus, important harmonization efforts are accomplished in the communication area to build a synthesis between the internet standards recommended by the W3C consortium, the European pre-norm 13606, based on the exchanges models (established by the Technical Committee for the European Standardization of Health Informatics, CEN TC 251) and the American or international health domain de facto standards (HPRIM, DICOM and HL7) [12].
Certain problems and requirements were geography-related. This was the case for great decisional systems, especially those supporting multiple interactions between hospitals, health agencies, and various institutions and organisms, which may be public or private. In fact, on

several dimensions and at different aggregation levels, these IS need to correlate information that belong to these two domains, whether they are coupled explicitly or not.

- The SNIIR-AM (*Système National d'Information Inter-Régimes de l'Assurance Maladie*) Data Warehouse project [13] aims at supporting the French national health department to select the intervention domains while taking into account numerous problems (financial, medical, social, accounts, public health and political). It covers the entire population (62 million persons in 2004) and has been considered the largest Data Warehouse in the world. The geography is a structuring dimension for it, at each level (from the design to the global and technical architecture). For many resources, there is no single way to identify it: for instance, doctors are identified with a regional number, which changes if the doctor changes of region.
- GENNERE (*a Generic Epidemiological Network for Nephrology and Rheumatology*) [14] is a networked information system designed to answer epidemiological needs. Based on a French experiment in the field of End-Stage Renal Diseases (ESRD), it has been adapted to Chinese medical needs and administrative rules. This project has raised various questions concerning how to take into account several kinds of specifities to build a general system, multidomain and multi countries.
- The SCALE program [15], launched by the European Commission, associates multiple information sources from different domains, such as health, environment and law. It develops a systematic approach to improving a European Environment and Health Strategy, to assess and minimize adverse health effects due to environmental pollution. The SCALE term sums up its objectives: it is "based on **S**cientific evidence, focused on **C**hildren, meant to raise **A**wareness, improve the situation by use of **L**egal instruments and ensure a continual **E**valuation of the progress made".

The more the level of the IS services increases, the more we have to simultaneously cross several domains and build multiple aggregations for different dimensions. So, in the next section, we will focus on the metadata to address this challenge.

3. Metadata and Interoperability

Faced with the proliferation of the information of various types, the interoperability solutions have evolved. This evolution occurred thanks to the development of standardization frameworks and the advanced information and communication technologies, especially Internet. A. Sheth [16] identifies three generations for the interoperability solutions: respectively before 1985, until 1995 and since 1996. From the first to the third generations, efforts to solve heterogeneity problems shifted from the database management systems, to the data (syntactic and structural), and then towards the information and knowledge.
Metadata (defined as "information about information") [17] allow human agents and automatic systems to discover whether there exist resources related to the application profiles. Since the second generation of interoperability solutions, metadata participate in the interoperability architectures. Their use develops alongside the emergence of approaches and tools to manage and correlate the contexts (ontology).
A general criterion to classify the metadata concerns their capacity in "capturing the (data and information) content of the information asset. …The semantic content (i.e. a level of abstraction closer to that of humans) is important for the metadata to model application domain-specific information" [16]. More precisely, cooperative health information systems require metadata which simultaneously use knowledge or human perception or cognition

(content descriptive metadata) and which are described by employing the terms specific to the domain of information (domain specific metadata).
It is a hard task to specify, manage and correlate such metadata because they require human expertise for specifying the domain (ontology) and for generating information correlation [18], [19], [20].
For the decisional IS, Connolly and Begg [21] introduce another classification of the metadata, which refers to the dataflows. Five flows are identified

- Inflow, for the extraction, cleaning and loading of source data
- Upflow, for adding value to the data (summarizing, packaging, ...)
- Downflow, for archiving and backing-up the data
- Outflow, for making data available to end-users
- Metaflow, for managing the metadata

In the next section we will consider content descriptive and domain dependent metadata for inflow, where heterogeneity problems are crucial. We will illustrate how some technologies developed within the framework of the consortium of WWW (W3C) [22] can be used to specify these metadata in order to support several standardization frameworks and to merge knowledge from multiple medical sub-domains (pathologies and drugs ...) or domains (health and environment, for instance).

4. Technical Infrastructure

The technical infrastructure used to represent content descriptive and domain dependent metadata for an HNO repository is necessarily modularized over multiple interoperable components, and based on XML technologies. By separating the content of the data from its structure and presentation, XML enables interchange and inference facilities over the internet.
Figure 1, which uses the famous "semantic web cake/tower" [23], points out the technologies involved at different contexts (specific domains and networked domain).

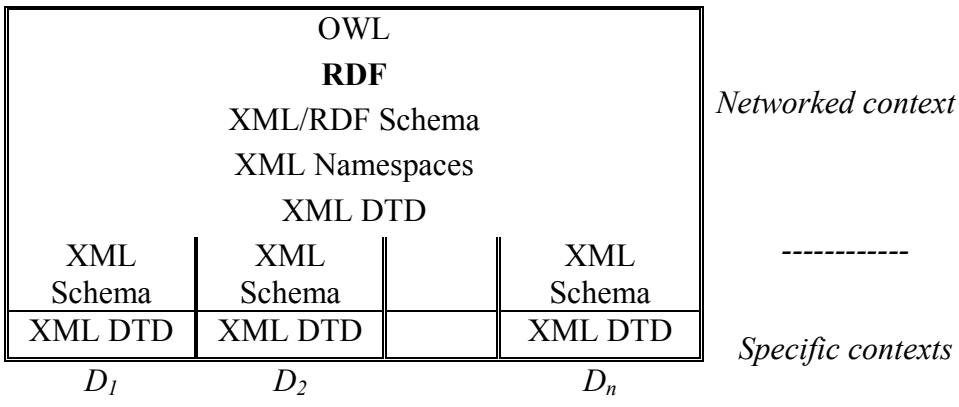

Figure 1: *XML technologies for the interoperability between specific contexts*

The Resource Description Framework (RDF) plays a central role. RDF statements are based on particular triples (resource, predicate, value) which can be developed in independent ways from different organizations. As anything (property, values and statements on the triples) may be manipulated as a resource, RDF is analogous to a very general relational data model, and metadata descriptions can be progressively enriched and combined. The other related technologies are:

- XML namespace facility, which allows the automatic combination of multiple meta-models. It offers a simple and modular approach to qualify groups or significant sub-groups of metadata elements, by associating them to the specific namespaces referenced by a unique Uniform Resource Identifiers (URI).
  This approach is adopted by the Medical Markup Language (MML) [24]. MML has been developed as a medical information exchange standard so that medical information documents described in MML comply with the rules of the Clinical Document Architecture (CDA) developed by the HL7.
  A MML document format is divided in several modules, which may be generated autonomously by different actors. For each instance of a module, the format and version of the module are controlled, by specifying these elements. Some examples are given in figure 2.

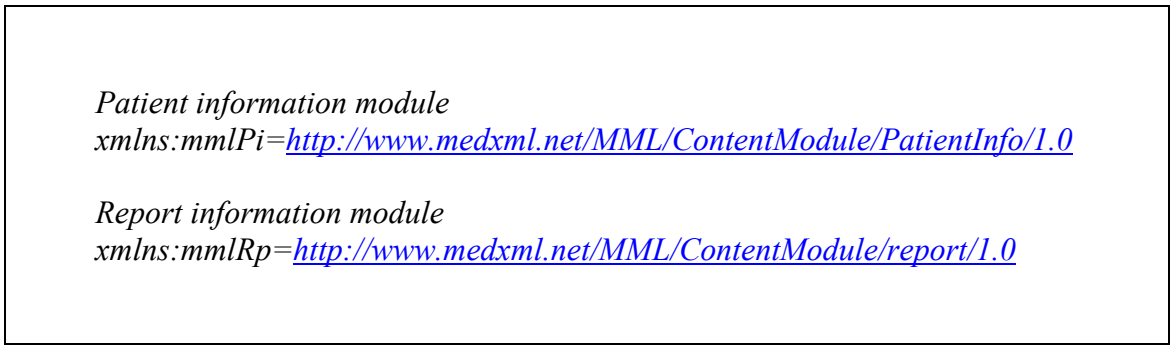

Figure 2: *Examples of namespaces declarations for MML CDA modules [24]*

- The XML/RDF schemas, to give the way of describing the constraints on the metadata (the objects, their attributes and relationships) within a specific area of interest (domain or subdomain).
- The Web Ontology Language (OWL) over the RDF/XML data, which is comparable to the business domain applications. By using inference facilities on some correspondences, OWL can discover new information, that was not stored explicitly in the RDF/XML data, and detect possible conflicts between multiple representations, before storing it in the datawarehouse. This implies building ontology for the network context, where conflicts between local contexts are solved and the correspondences are explicit [4].

Enhancing the standardization frameworks renders this technical architecture more utilized, practical and attractive to the medical informatics community. But much work remains to be done.

If the heterogeneity problems have been solved between the legacy systems and the correspondences have been specified, then it is technically possible to analyze health information by socio-economical or environmental criteria.
If we look to the localization aspects, many problems, pertaining to granularity discrepancies and structural conflicts, are not trivial. On the one hand, for instance, the addresses are expressed in MML compliant databases with a full address expression, or subdivided by using 4 elements (figure 2). On the other hand, ISO19115 [25], recommended for the geographical domain, defines the health as a very large topic category, to cover several sub-domains (health, health services, human ecology, and safety) to represent disease and illness as well as factors affecting health, or health services.

```
<mmlAd:Address mmlAd:repCode = "A" mmlAd:addressClass = "business"
mmlAd:tableId ="MML0025">
  <mmlAd:full>5200 Kihara, Kiyotake-cho, Miyazaki-gun, Miyazaki-
prefecture</mmlAd:full>
  <mmlAd:zip>889-1692</mmlAd:zip>
  <mmlAd:countryCode>JPN</mmlAd:countryCode>
</mmlAd:Address>
```

Figure 3: *Example of a MML full address [24]*

5. Conclusions

In complex domains such as health, the information systems of networked organizations are faced with interoperability challenges. We have analyzed some interactions between health and geographic information sources, their heterogeneity, and their potential. Our approach to interoperability is based on the identification and composition of content descriptive and domain dependent metadata. RDF, with its related technologies, appears at the center of a flexible technical infrastructure to support them.
In further research we shall consider the analysis of several theoretical aspects (semistructured data) and the assessment of the "distance" between the real world databases and the metadata recommended for the application domains.

6. References

[1] Sheth A, Larson J. A.: "Federated database systems for managing distributed, heterogeneous, and autonomous databases", ACM Computing Surveys 22(3): 183–236

[2] Salzano G., Bourret C.: "Health Networks and Global Health Services: An Information System Analysis", ICSSHC2004 (The 8th International Conference on System Science in Health Care) - Health Care Systems: Public and Private Management, Université de Genève, 1-4 septembre 2004

[3] Elmagarmid A., Rusinkiewicz M. and Sheth A. (eds): "Management of Heterogeneous and Autonomous Database Systems", Morgan Kaufmann Publishers, Inc., San Francisco, California, 1999.

[4] Salzano G.: "Integration Methodology for Heterogeneous Databases", in Heterogeneous Information Exchange and Organizational Hubs, edited by H. Bestougeff, J.E. Dubois, B. Thurasingsham, Kluwer Academic Publishers, Netherlands, pages 1-16, 2002

[5] IEEE P802.15, "Coexistence, Interoperability, and Other Terms", November, 1999 IEEE P802.15-99/114r1, http://w3.antd.nist.gov/IEEE/99-114.pdf

[6] Bourret C.: "Data Concerns and Challenges in Health: Networks, Information & Communication Systems and Electronic Records", Data Science Journal, volume 3, september 2004, pp. 96 – 113.

[7] Loi n° 2002-303 sur les Droits des patients et la qualité du système de santé, France, 4 mars 2002 http://www.legifrance.gouv.fr/WAspad/UnTexteDeJorf?numjo=mesx01000921

[8] Salzano G., Bourret C.: "Interoperability among medical applications", Proceedings IEEE HealthCom 2002 4th International Workshop on Enterprise Networking and Computing in Health Care Industry, Nancy, France, June 6th –7th 2002, pp. 117-120.

[9] Brunie V., Morizet-Mahoudeaux P. and Bachimont B.: "Separating Textual Contents from Structures for Reading Hypertext Structured Medical Records", in HYPERTEXT '98: Pittsburgh, PA, USA, June 20-24, 1998

[10] Réseau Canadien de la Santé: www.canadian-health-network.ca

[11] National Health Service: Information for Health http://www.nhsia.nhs.uk/def/pages/info4health/contents.asp

[12] Groupement pour la Modernisation du Système d'Information Hospitalier, http://www.gmsih.fr/tiki-index.php

[13] Nakashe D.: "Problems in Designing Huge Datawarehouses and Datamarts", American Conference on Information System, 2003, http://cedric.cnam.fr/PUBLIS/RC559.doc

[14] Simonet A., Landais P. and al.: "GENNERE: a Generic Epidemiological Network for Nephrology and Rheumatology", Conceptual Modeling - ER 2004, 23rd International Conference on Conceptual Modeling, Shanghai, China, November 2004, Proceedings, Lecture Notes in Computer Science 3288,2004

[15] SCALE project, http://www.brussels-conference.org/project.htm

[16] Sheth A.: "Changing Focus on Interoperability in Information Systems: From System, Syntax, Structure to Semantics", in Interoperating Geographic Information Systems M F Goodchild, M J Egenhofer, R Fegeas and C A Kottman (eds), Kluwer Publishers, 1999

[17] Metadata and Resource Description. W3C, Technology and Society Domain. www.w3.org/Metadata/

[18] Malet G., Munoz F., Appleyard R., Hersh W.: "A Model for Enhancing Internet Medical Document Retrieval with Medical Core Metadata", J Am Med Inform Assoc. 1999 Mar–Apr; 6(2): 163–172.

[19] Catley C., Frize M.: "Design of a health care architecture for medical data interoperability and application integration", Proc. Joint BMES/EMBS Conference, Houston, 2002 http://www.sce.carleton.ca/~ccatley/embs-bmes2002_CatleyFrize.pdf

[20] Crichton D., Hughes J. S., Downing G. J., Kincaid H., Srivastava S.: "An Interoperable Data Architecture for Data Exchange in a Biomedical Research Network", Fourteenth IEEE Symposium on Computer-Based Medical Systems (CMBS'01) March 26 - 27, 2001, Bethesda, Maryland http://csdl.computer.org/comp/proceedings/cbms/2001/1004/00/1004toc.htm

[21] Connolly T., Begg C. E.: "DataBase Systems: A Practical Approach to Design, Implementation and Management", Fourth Edition, 2005

[22] World Wide Web Consortium (W3C), http://www.w3.org/

[23] Berners-Lee T.: "The semantic Web and Research Challenges", 2003, http://web-services.gov/The%20Semantic%20Web-TBL203.ppt

[24] MML: Medical Markup Language Specifications, Version 3.0 http://www.medxml.net/E_mml30/MMLV3Spec.pdf

[25] UK Gemini: A Geo-spatial Metadata Interoperability Initiative - ISO 19115: Metadata Standard – Proposed Element Set, 21 December 2003 http://www.gigateway.org.uk/metadata/pdf/ISO19115ProposedElements.pdf