Designing and implementing health data and information providers

Michel Joubert\textsuperscript{a,b,*}, Jean-Charles Dufour\textsuperscript{a,b}, Sylvain Aymard\textsuperscript{a}, Laurent Falco\textsuperscript{b}, Marius Fieschi\textsuperscript{a,b}

\textsuperscript{a} Laboratoire d’Enseignement et de Recherche en Traitement de l’Information Médicale, Faculté de Médecine, Université de la Méditerranée, 27 bid Jean Moulin, 13005 Marseille, France
\textsuperscript{b} Service de l’Information Médicale, Hôpital de la Timone — Adultes, Assistance Publique — Hôpitaux de Marseille, France

Received 31 October 2003; received in revised form 17 March 2004; accepted 6 April 2004

\begin{abstract}
Objectives: To model and implement web portals providing access to certified and high-quality information in the domain of health. Material and methods: The Unified Medical Language System (UMLS) knowledge sources of the U.S. National Library of Medicine and principles of implementation resulting from the previous ARIANE project are described. The XML technology that allows files transformations by the means of XSLT is briefly presented. Results: The design and implementation of software modules that exploit knowledge sources, operate the translation of a user’s query to selected information sources, and wrap obtained results are detailed. Querying documentary and factual medical databases are presented. Discussion: Current implementation and wrapping perspectives are discussed in terms of integration and interoperability of health information and data resources. © 2004 Elsevier Ireland Ltd. All rights reserved.
\end{abstract}

1. Introduction

The aim of the project ARIANE was to develop a Web portal that stores content descriptions and access means to Web sites certified and indexed by hand, which queries them when possible, or provides access to them, with respect to previous works made to guarantee quality of health information over the Internet [1–6]. In the framework of the project ARIANE we designed and implemented a middleware architecture for a portal dedicated to healthcare professionals [7]. It is made of a user interface, a broker and mediators. The user interface assists end users to express their queries by means of the Unified Medical Language System (UMLS) knowledge sources of the U.S. National Library of Medicine [8]. The broker matches a query against descriptions
of information sources registered by hand. When a user selects a source the description of which matches a query, a mediator is then built dynamically and activated to query the target information source in its own query language, when it is possible, or to establish a connection to it.

The fundamental results of the semantic integration produced by the project ARIANE [9] are reused in the framework of the project CoMeDIAS (Computerized Medical Data through Internet Access Services). ARIANE was dedicated to documentary resources. The aim of the project CoMeDIAS is to include factual databases into the sphere of activity while exploiting the assets of the previous project. Its aim is to produce a platform for applications dedicated to gather information and data in various databases inside an information system and to wrap them into a unique user interface. Data may concern a patient: clinical or surgical data, medical imaging, etc. It may also concern a medical problem: clinical guidelines, operating or medical protocols, drugs prescriptions, etc. CoMeDIAS must not be seen as an application, but as a set of services thanks to which it is possible to build user interfaces that allow end-users to "intelligently" navigate inside databases and easily gather useful information in databases inside or outside their institution. Integration and interoperability of heterogeneous databases is operated in CoMeDIAS by a middleware made of components interfaced with existing communication protocols by the way of connectors able to physically communicate with databases according to the appropriate formats.

The IHE (Integrating the Healthcare Enterprise) approach started in the USA several years ago, and extended in Europe, with the intent to facilitate the communication of medical imaging data [10]. Based on the same idea which consists of making systems interoperable by means of structured information exchanges, the DICOM (Digital Imaging and Communication in Medicine) standard, in the medical imaging field again, stands out as a norm [11]. While Health Level Seven (HL7) evolves towards a universal medical data interchange format [12]. As soon as applications exchange or share data and documents, problems appear due to differences in coding rules. Therefore, shared coding rules and transcoding capabilities are crucial in making systems interoperable. Coding and transcoding processes need to use standard nomenclatures and thesauri. The project UMLS aims at unifying them into a coherent system that allows to translate a given concept denoted into a nomenclature into its representation in another one, among other features. Besides coding, structure of exchanged data is of great significance for computing systems interoperability since it provides systems with the needed basis to "understand" each other.

The increasing sophistication of World-Wide-Web software brings today a new challenge to software integration and/or interoperability. Several Web sites and prototypes presented as health portals share this objective, even if their prospective end-users are of different kinds. We present here a model and an implementation of the querying and answering processes according to the current eXtensible Markup Language (XML) technology [13,14]. In this article we present an implementation based on XML and eXtensible Stylesheet Language Transformation (XSLT) that goes beyond the status of prototype. It has been tested with the PubMed source of information and, due to its software architecture, it has been easily reused in the development of a tumor samples bank at the University hospitals of Marseille. The implementations are component-based with the intent to be modular and to make sure that developed components will be reusable [15,16].

2. Material and methods

2.1. The UMLS knowledge sources

The UMLS is a complex collection of medical concepts, terms and relationships issued from standard classifications. Among the components that constitute the UMLS knowledge sources we exploit principally the so-called Metathesaurus and the Semantic Network. The core concepts, which have been isolated in the Metathesaurus are connected to generic types of concepts in the Semantic Network. The types of concepts are interconnected by binary semantic relationships. Concepts potentially inherit of properties, especially of semantic relationships, from the types they are associated to. Interconnections between concepts by the means of semantic relationships constitute possible associations that must be validated by the means of co-occurrences of concepts isolated from the Medline database [17].

2.2. The XML technology

XML is a universal format for structured documents and data on the Web. The XML syntax uses matching start and end tags to mark up information. Its simple syntax is easy to process by machine, and has the attraction to support a wide variety of applications. The eXtensible Stylesheet Language (XSL) provides with the means to display XML files. The
Designing and implementing health data and information providers

2.3. Principles of implementation

Component-based programming allows a flexible and efficient mechanism for software development. It has proved its efficiency in medical informatics applications [16,23]. Following this approach, a Web service [24] is made of components, which cooperate with some other software modules for the achievement of a requested action, such as a query to a database. Interactions with a Web service are operated by the means of XML files exchanged by the way of Internet protocols. This global view on Web services is schematized by the diagram of Fig. 1, where they appear as mediators between users and databases.

The user interface and Web services communicate by the way of an Internet protocol. With the intent to communicate with a database, a Web service must use an appropriate protocol. At this level connectors are coming up: a connector is able to physically transmit a query to a data resource and to receive results according to appropriate protocols and formats. The place and function of such connectors are illustrated in Fig. 2.

3. Results

3.1. Querying documentary databases

By the means of a graphical user interface, end-users express queries to documentary resources as medical concepts connected by semantic relationships picked out in the UMLS Metathesaurus and Semantic Network. Then a user’s query is treated sequentially by four processes: (1) a first process translates concepts and semantic relationships into the corresponding terms in the vocabulary of the target source; (2) a second process translates the query according to a model of syntax and a data representation compatible with the source to be accessed; (3) a third process formats a query for a later transportation by a communication service with a selected source; and (4) a last process formats results obtained from the information source. These processes are detailed in what follows. An example will serve as a thread. The initial user query is “treatment of stomach ulcer by ranitidine”. Two concepts are isolated: “stomach ulcer” and “ranitidine”, and a semantic relationship: “treats”.

The aim of the first process is to translate the XML representation of a query (by means of UMLS concepts and semantic relationships) into another XML file involving the vocabulary of a target infor-
mation source. This is operated by means of a XSL file and UMLS knowledge sources. Knowing that the target source is PubMed and its vocabulary is MeSH, the sample query is then represented as in Fig. 3.

A second process translates the file produced by the above process into a XML file the content of which is formulated according to the syntax and grammar used to query the target source. This is operated by a software component that exploits UMLS knowledge sources. In the present example, the result obtained is a string representing the request in the query language of PubMed, as illustrated in Fig. 4. The diagram of Fig. 5 illustrates the transformation mechanism operated by the above modules. Providing as an input a query involving two concepts associated by a semantic relationship, a XSLT process produces a query to the related information source. A XSLT file is dedicated to PubMed, other files have been written to access other information sources such as the HON databases [25] and the C- 

Fig. 3 The XML representation of the conceptual graph expressing the query "ranitidine treats stomach ulcer" for PubMed.

```
<Request>
  <Source Name="PubMed" />
  - <Concept CUI="C0034665" HICD="D3.383.312.750"
     Designation="RANITIDINE" ID="1"/>
  - <Relation TUL="T554" Designation="TREATS"
     ID="1">
    <Concept CUI="C0038358" HICD="C6.405.748.860"
     Designation="STOMACH ULCER" ID="2"/>
  </Relation>
</Request>
```

Fig. 4 The representation of the request in the query language of PubMed.

```
(((Ranitidine/ADMINISTRATION and DOSAGE[MeSH Terms]) OR 
Ranitidine/THERAPEUTIC USE[MeSH Terms]) OR 
Ranitidine/PHARMACOKINETICS[MeSH Terms]) OR 
Ranitidine/PHARMACOLOGY[MeSH Terms]) AND 
stomach ulcer/DRUG THERAPY[MeSH Terms]
```

Fig. 5 The XSL transformation mechanism for two concepts and a semantic relationship.
Fig. 6 Series of transformations operated by components (treatment of PubMed results).

the mapping operated in the case of PubMed requesting. The XML representation of each element returned by PubMed is presented in the left part of the screen dump. The data schema on the left side, gets wrapped in a standard format, the schema that is shown on the right side. This example shows how a set of authors (each of them is represented in the PubMed schema by a last name, a forename, and initials) is transformed into a list. The same kind of transformation is done in the case of keywords. While titles and abstracts are not modified and directly transferred to the related user interface element.

3.2. Implementation of a tumor samples bank

The University hospitals in Marseille (AP-HM), France, are equipped with old-style computer systems, which link hospitals into a unique and huge network. For the most part they are: a patient management system (identity, stay number, entry and exit dates, identification of visited units, etc.); a system for managing clinical and surgical data; a laboratory system, including pathology; a blood products management system. These systems are independent one from each other, but they are connected to the patient management system. Moreover, some medical units maintain complementary databases for care and research. This is the case of a specific database the aim of which is to manage the biological resources (tumor samples) that a biology unit stores in its freezers.

French Ministry of Health and Ministry of Research proposed to standardize the existing or "under construction" tumor samples banks all over France. The aim of the harmonization of tumor samples banks is to contribute to cancer research, as it was decided in the UK by the National Cancer Research Institute (NCRI) initiative [26], for instance. The head of the AP-HM and clinicians of this institution, who are concerned with diagnosis and treatment of cancer, decided to build a tumor samples bank with the intent to provide access to different databases with a unique tool. Involved systems are the patient management system, the related clinical and surgical databases, the pathology database, and the tumor samples database management system. A prototype has been implemented according to the above principles. The various systems and databases continue to be exploited independently, the prototype is a tool intended to query them separately, to retrieve and display relevant data. The current release of the prototype includes data issued from organs oncology: neuro-oncology, gastroenterology, etc.

Fig. 7 shows the principles introduced in Fig. 1 instantiated in the case of the tumor samples bank of AP-HM: standard nomenclatures are ICD-10 and SNOMED; databases are clinical and surgical databases, pathology data, and tumor samples management, to which is joined the patients management system; the user interface is a standard
Fig. 7 Architecture of the tumor samples bank system of AP-HM.

Internet browser. The transformations operated in order to translate queries and connections with the proprietary systems have been realized for the occasion. Appropriate connectors have been developed in order to operate the expected mediation between the Web services and the various data resources. Results returned by the various and heterogeneous databases are transformed into XML files with the intent to be displayed homogeneously, as presented by Fig. 8, where the tree-view on the left of the screen is constituted dynamically according to the units visited by a selected patient (whose identification items has been rubbed), and the tab component on the central part of the screen displays information for a given unit (pathology here) at a selected date (the left tab here).

4. Discussion

Following the principles proposed in the ARIANE and CoMedias projects, it will now be easily feasible to mix factual data and medical information retrieval. That is to say, for instance, when consulting data about tumor samples, to obtain by a click on a ICD or SNOMED code information related to the pathologies, acts, or drugs the codes express. This information may be inside or outside the institution. It can be clinical guidelines, internal protocols or prescription and dosage of drugs as usually used in the institution. In such a case, queries dynamically attached to displayed data would be "glioma surgery" or "chemotherapy of glioma" depending on the kind of concerned data, surgical or clinical. It can also be information coming from outside, such as clinical trials or the latest bibliographical references published by PubMed on a topic. It is obvious that such a navigation needs usage of standard nomenclatures for data and information retrieval, and of an ontology for combining their designations semantically.

Since it does not need to modify an existing information system at all, our approach allows to quickly have at one’s disposal an expected service and to have time to let a hospital information system evolve coherently. This approach is also
Designing and implementing health data and information providers

interesting in various integration projects since it proposes a complementary capability to data warehousing which often imposes long and costly treatments and duplication of data [27]. Without intending to substitute for data warehousing, the objectives of which are different, it proposes a logical integration in order to query directly the source databases, and does not impose the physical construction of data warehouses. Transformations operated by Web services allow a representation of data different from their original representation in the databases, as data warehouses generally do. Integration and interoperability are an affair of software architecture [28]. From this point of view, the solution we propose is a middleware built on a set of Web services which cooperate with existing systems. Following this method, our objectives encourage the use of an ontology and standard nomenclatures because the Web services are conceived to exchange data and documents. Sharing coding rules is of the utmost importance at this level. Moreover, this should allow to move towards the promising direction of a semantic medical Web. The integration and/or interoperability of components are we concerned with needs: (1) a common ontology unifying concepts and vocabularies, and (2) descriptions of formats and protocols for sending and receiving messages. The UMLS knowledge sources constitute today a commonly recognized and operational ontology in the health domain.

The XML/XSL technology is the binding agent that makes the interoperability feasible and easy to do as possible. This architecture tested with PubMed, and other documentary resources, has been easily applied to other information sources on the basis of reusable functions we developed. Adopting the XML technologies for exchanging information between objects presents a two-fold advantage: (1) XML is a standard, and XSL transformations allow to modify XML files by the means of knowledge bases when exploited in accordance with appropriate functions, and (2) since Web sites, as well as data management systems, return today XML or HTML files, a generic XSLT filter allows us to wrap results returned by these systems.

Acknowledgments

The authors thank the National Library of Medicine who provided them with the successive releases of the UMLS products. They thank the French Ministry of Education and Research, which partially founded the project ARIANE. They also thank all the members of the LERTIM who cooperated to the project ARIANE. The authors thank the clinicians who brought their competencies in this building phase of the tumor samples bank of the AP-HM, as well as the “Direction de l’Informatique et des Réseaux” of AP-HM for its support and its active participation in this project. They thank the French Ministry of Health for its support.

References