A UMLS-based Method for Integrating Information Databases into an Intranet

Françoise Volot, M.D., M.S., Michel Joubert, Ph.D., Marius Fieschi, M.D., Ph.D, and Dominique Fieschi, Ph.D.
Service de l'Information Médicale, Hopital de la Timone - Adultes, Marseille, France
CERTIM, Faculté de Médecine, Université de la Méditerranée, Marseille, France

The Internet and the World Wide Web provide today end-users with capabilities to access universally to information in various and heterogeneous databases. The biomedical domain benefits from this new technology, specially for information retrieval by searching and browsing various sites. Nevertheless, end-users may be desoriented by specific ways to access information on different servers. In the framework of an Intranet design and development, we present a method for integrating information databases based on knowledge sources of the UMLS. The method provides designers of a Web site with facilities to implement an easy and homogeneous access to information. The pages are built dynamically and displayed according to a style sheet and their content stored in a database during the design phase. The database also describes the links between pages. Moreover, this organization provides administrators with powerful capabilities to manage Web sites.

INTRODUCTION

Massive volumes of data, unprocessed details, and heterogeneous sources of information are present in large information systems, such as Hospital Information Systems. They have serious consequences: a lot of data is manually processed, some information is unused, and costs are high. Two solutions for improving these systems are in broad terms: 1) the re-engineering of the software components with recent computer technics and architectures: client/server, object-orientation, communication standards, and so on, or, 2) including the use of modern technology, the total revision of these systems following a knowledge engineering approach. The first approach guarantees the software interoperability of the services inside a system. It leads to the notion of open systems which satisfies the constraint of heterogeneity of hardware and software. The second approach is an answer to the semantic heterogeneity of data processed by the applications. It proposes a semantic information integration which is necessary for the applications to process in a corporate work. It emphasizes the notion of conceptual analysis[1].

An ontology is an inventory of the things that are presumed to exist in a given domain together with a formal description of the properties of those things and relations that hold among them[2]. An ontology thus provides a representational vocabulary for a given domain with a set of rules that constrain the meaning of the terms in that vocabulary sufficiently to enable consistent interpretation of data framed in that vocabulary. Even if it is uneasy to build, validate and reuse[3], an ontology shared by different applications is the cornerstone of a semantic integration in information systems[4]. The UMLS includes in its knowledge sources (Semantic Network, Metathesaurus, Specialist Lexicon) a part of the medical knowledge, and above all, almost of the biomedical vocabularies[5,6]. As it seems possible in some cases to translate terms between controled vocabularies by mapping them on concepts in the Metathesaurus[7], we can consider today the UMLS knowledge sources as an operational and suitable ontology.

The various servers in a computerized information system implement data according to the way they represent medical concepts. The main problem for end-users to access the data in the servers databases is to match their own viewpoint on medical concepts with the representations made in the databases. The aim of the project ARIANE is to build user interfaces with heterogeneous information databases, and to provide end-users with easy and natural means to query them. In previous works we have shown how the UMLS knowledge sources are exploited with this aim in view. First, we modeled the UMLS components we exploit[8]. The prototype developed provides end-users with the capability to browse the UMLS knowledge sources and to build a query step by step by means of selections of concepts and semantic relationships[9]. The present work allows designers of a corporate Intranet a

1091—8280/97/$5.00

method to develop a conceptual interface to heterogeneous information databases, including queries to the databases. The produced pages and links implement only the part of the UMLS needed by the designers to provide forthcoming end-users with a Web-style access to information.

INTEGRATION OF DATABASES INTO AN INTRANET: A SCENARIO

With the intent to illustrate our approach we will use a scenario before we present in the next section the principle of the method for integrating information databases into an Intranet. The aim is to build a set of pages giving access to information in relation with digestive system diseases. Information and data are located in a bibliographical database, in a database of surgical reports, and in a database of clinical guidelines. At one point, the designer builds access paths to information in relation with pancreatic diseases. Then he/she focuses on pancreatic neoplasms and on other associated concepts in the frame of their diagnoses and therapeutics. For the scenario to be simple we restrict here his/her concern to these three topics which relate to the following types of concepts in the Semantic Network of the UMLS respectively: disease or syndrome, diagnostic procedure, and therapeutic or preventive procedure.

Paths to information are built on the basis of the UMLS following four successive steps:

- Path to the concept *pancreatic neoplasms*. Starting from the type of concepts *disease or syndrome* in the Semantic Network, the designer goes to the concept *digestive system diseases* in the Metathesaurus, then to the concept *pancreatic diseases*, and finally to the concept *pancreatic neoplasms*. Another path via the concept *digestive system neoplasms* was also possible. The designer validates the path he/she followed.
- Semantic relationships between concepts. In the Semantic Network the semantic relationships *diagnoses* and *treats* hold between the types of concepts *diagnostic procedure* and *therapeutic or preventive procedure* and the type *disease or syndrome* respectively. When used these relationships translate the designer's concern in pancreatic neoplasms from the viewpoints of both diagnosis and therapeutics, and they are ins tanciated as hypertext links between pages.
- Instanciation of the semantic relationships. The designer decides to link *ultrasonography* and *pancreatic diseases* by means of the relationship

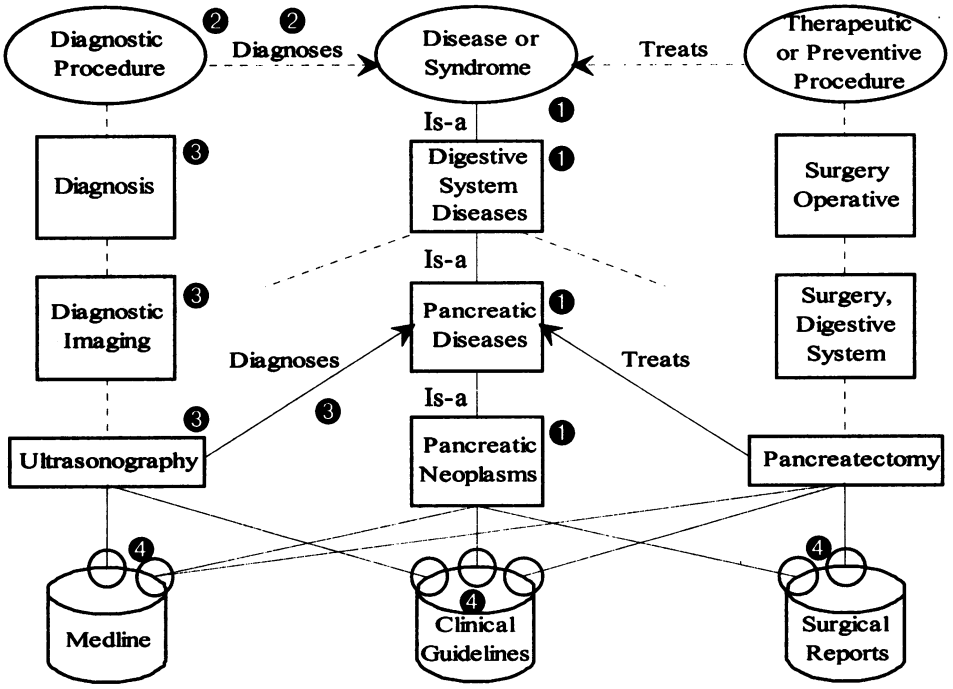

Figure 1. The priciple of integration of information databases into an Intarnet by means of UMLS knowledge sources.

diagnoses. This precision level is chosen because an ultrasonography is not used to diagnose every digestive system disease, such as an esophagial disease. Even if each of the pancreatic diseases is not automatically treated by a pancreatectomy, it may be relevant of this surgical act at one time. Thus, the designer associates *pancreatectomy* and *pancreatic diseases* by means of the relationship *treats*.

- Access to information databases depends on the kind of databases is queried. An access to a database may be a query expressed in the proprietary language of a bibliographical server. It may be a SQL query to a patients database. It may be a link to a page on a Web server.

These steps are schematized by the diagram of Figure 1. They are numbered from 1 to 4. The stored elements are represented in plain boxes, while those used but not stored are represented in doted boxes.

Once the designer has stored the results of the successive steps, he/she has implemented an interface with information databases. It is a set of pages connected by typed hypertext links. Other pages may be created following the same principle by the same designer and/or other designers until having paths to information in relation with all the digestive system diseases. End-users may follow later these paths to obtain information after some mouse clics in a well-known Web-style environment, independently of the UMLS knowledge sources the designers used to build the interface, and in a conceptual way they may feel natural. An end-user accesses firstly a home page (here it is the page of digestive system diseases). Then, following the links between pages, he/she navigates until he/she reaches the searched page (the page of pancreatic neoplasms). From here, he/she can either follow a link until the page of a concept semantically related to the concept of the current page (for instance : the page of ultrasonography or the page of pancreatectomy) and thus explore thoroughly, or activate queries to information databases.

A METHOD FOR INTEGRATING INFORMATION DATABASES INTO AN INTRANET

During the design phase, pages defined by designers, links between the pages et accesses to information databases are stored in a database. These data will be later exploited for building dynamically the pages displayed to end-users.

A style sheet is associated to each type of concepts. For instance, there is a style sheet for the type *disease or syndrome*, a different one for *diagnostic procedure*, and another one for *therapeutic or preventive procedure*. Each style sheet contains fields which will be fulfilled by the concept name and values of other attributes before a page is displayed. The other attributes are the concept definition when available, key-words extracted from the Metathesaurus, and so on. Fields are also foreseen for links and access to databases. Each time a page is defined, it is done according to the style sheet associated to the selected type of concepts to which the concept is connected. The definition of a page consists in the storage of its content in the following tables of the database.

- A table stores all the information related to the defined pages. It contains for each page an identifier of the page, its name and other attributes values. It contains also the unique identifier of the related concept and the identifier of the related type of concepts. Fields in the appropriate style sheet will be later fulfilled with these values.
- Hypertext links descriptions are stored in another table. It contains the identifier of each link, the identifier of the semantic relationship from which it is an instance, and the identifiers of the linked pages. An appropriate field in a displayed page will show the links which start from it.
- A third table contains the defined accesses to information databases. They are query masks identified by the pages identifiers and information sources. When the information source is a bibliographical server, the key-words stored with the page description are used to fulfill the mask and query the server. When the source is a RDBMS, the query mask is fulfilled by key-words to build a complete SQL query. In case of access to a Web site, its URL is stored.

The diagram of Figure 2 schematizes the dynamic building of a page according to its style sheet and its content stored in the database.

When a Web site becomes large, the number of pages increases dramatically, and their management becomes uneasy. The above method offers solutions to this problem since no static pages are created and stored. Pages are built dynamically when needed by means of their related style sheets and stored contents. The number of style sheets is reduced to the number of types of concepts in the UMLS. Even if the number of defined pages becomes large, they may be managed easily with regard to the capabilities a relational database management system

allows. The stored key-words enhance this management. They make possible to query the database for knowing how many pages are in relation with a specialty, a theme, and so on. They also make possible to control the redundancy and thus to contribute to maintain the consistency of a large set of pages. Since anchors of links are stored in a table of the database, the whole consistency may be ensured by queries intended to verify the real existence of the related pages.

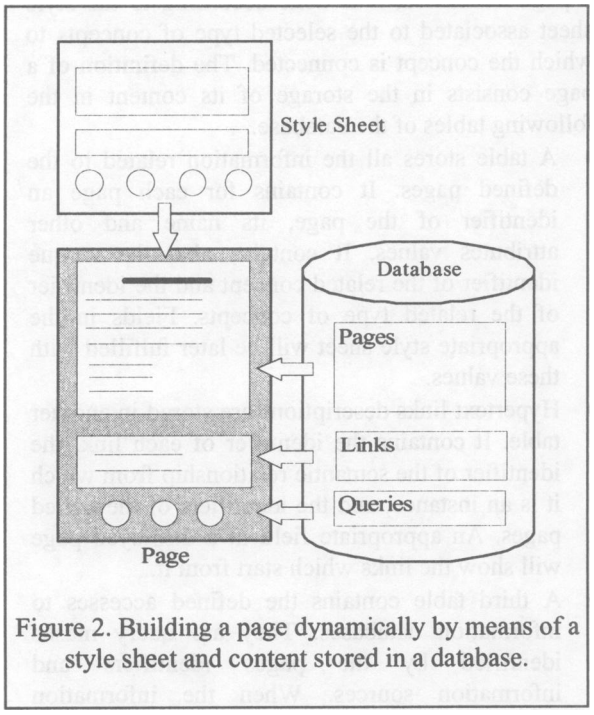

Figure 2. Building a page dynamically by means of a style sheet and content stored in a database.

DISCUSSION

As felt by its conceptors themselves, the UMLS may be a basis for building « better » user interfaces[10]. Few years ago the World Wide Web technology has emerged and has been successfully exploited by many medical informatics researchers and practitioners. Benefits taken from this technology come from the universal user interface provided by Internet browsers. For instance, the U.S. National Library of Medicine provides a direct access to the UMLS knowkedge sources over the Intrenet[11]. Other works are intended to provide end-users with easy to use and universally available access to information databases[12,13], supporting the integration of various types of biomedical information[14]. The method we presented uses the UMLS knowledge sources as a shared ontology for building a homogeneous way to access heterogeneous information databases in a corporate Intranet. Only the needed elements (concepts and relationships) are instanciated as pages and links respectively. But, following the same method, several designers may add new pages and links to the already existing ones without any risk to introduce inconsistency since the work done by previous designers can be: 1) either reused as it is, by linking new pages to the already existing ones, or 2) enhanced with new links into the existing pages. Both the first and second processes are operated according to the UMLS knowledge sources. The end-users benefit from: 1) the unique Web-style environment produced to facilitate information retrieval in the heterogeneous information databases, and 2) the natural organization of the pages due to the conceptual way to link them according to the UMLS knowledge. We are convinced that this kind of navigation until querying information databases reduces the effort to learn the query languages of the different servers, and decreases the desorientation the end-users may have when faced to other data organizations.

Our aims are closed to those of co-workers in the « Intelligent Information Integration » project (I*3). Following the I*3 concerns[15], our next works will be in relation with the processes we must implement for querying information databases efficiently. Among the needed query services are the following processes. Presently a query is sent to only one database at a time. Generally, a complex query, such as those to which we turned our attention in previous works, must be decomposed. The sub-queries resulting of the decomposition may be sent to different databases simultaneously. A query related to pancreatic neoplasms and their diagnoses may found results, for instance, in a bibliographical database and in a database of clinical guidelines. A query decomposition process is thus needed. For the query to be processed, it is also necessary to develop a process intended to select the sources of information. Another essential process is the translation of a sub-query from the internal representation into the query language of the selected information source. Previously we expressed queries in the *conceptual graphs* formalism due to its closeness to natural language representation and processing, and its ability to represent the meaning of a query by means of concepts and relationships[16]. Since this formalism is generally more expressive than the query languages of the information servers, it is easy to translate a query into one of these latter languages, as we have shown previously[9].

Acknowledgements

This work has been partially founded by the French Ministry of Education and Research. The authors thank the U.S. National Library of Medicine which gracefully provided them with the UMLS knowledge sources. This work has been achieved with the 1996 release.

References

1 Sowa JF. Conceptual Analysis for Knowledge-base Design. Meth Inform Med 1995; 34: 165-71.

2 Gruber TR. A Translation Approach to Portable Ontologies. Knowledge Acquisition 1993; 5: 199-220.

3 Musen M. Dimensions of Knowledge Sharing and Reuse. Comp Biomed Res 1992; 25: 435-67.

4 McCray AT, Scherrer JR, Safran C, Chute CG. Concepts, Knowledge and Language in Health-Care Information Systems. Meth Inform Med 1995; 34: 1-4.

5 McCray AT, Nelson SJ. The Representation of Meaning in the UMLS. Meth Inform Med 1995; 34: 193-201.

6 Tuttle MS, Suarez-Munist ON, Olson NE and al. Merging Terminologies. Proc. MEDINFO 95. Greenes R, Peterson H, Protti D, eds. North-Holland 1995: 162-66.

7 Qing Zeng MS, Cimino JJ. Mapping Medical Vocabularies to the Unified Medical Language System. JAMIA Symposium Supp. Cimino JJ, ed. 1996: 105-9.

8 Joubert M, Miton F, Fieschi M, Robert JJ. A Conceptual Graphs Modelling of UMLS Components. Proc. MEDINFO 95. Greenes R, Peterson H, Protti D, eds. North-Holland 1995: 90-94.

9 Joubert M, Robert JJ, Miton F, Fieschi M. The Project ARIANE: Conceptual Queries to Information Databases. JAMIA Symposium Supp. Cimino JJ, ed. 1996: 378-82.

10 Lindberg DAB, Humphreys BL. The UMLS Knowledge Sources: Tools for Building Better User Interfaces. Proc. 14th SCAMC. Miller RA, ed. IEEE Computer Society Press 1990: 121-25.

11 McCray AT, Razi AM, Bangalore AK et al. The UMLS Knowledge Source Server: a Versatile Internet-based Research Tool. JAMIA Symposium Supp. Cimino JJ, ed. 1996: 164-68.

12 Chute CG, Crowson DL, Buntrock JD. Medical Information Retrieval and WWW Browsers at Mayo. JAMIA Symposium Supp. Gardner RD, ed. 1995: 903-7.

13 Hersh WR, Brown KE, Donohoe LC et al. Cliniweb: Managing Clinical Information on the World Wide Web. JAMIA 96; 3: 273-80.

14 Detmer WM, Shortliffe EH. A Model of Clinical Query Management that Supports Integration of Biomedical Information over the World Wide Web. JAMIA Symposium Supp. Gardner RD, ed. 1995: 898-902.

15 Intelligent Information Integration (ARPA/ISO) project: http://dc.isx.com/I3/

16 Sowa JF. Toward the Expressive Power of Natural Language. Principles of Semantic Networks: Exploration in the Representation of Knowledge. Sowa JF, ed. Morgan Kaufmann 1991: 157-89.