MEDINFO 2001
V. Patel et al. (Eds)
Amsterdam: IOS Press

ARIANE: A Mediation Framework with Health Information Sources

Michel Joubert[a,b], Sylvain Aymard[a], Dominique Fieschi[a], Marius Fieschi[a,b]

[a] *LERTIM, Faculté de Médecine, Université de la Méditerranée, Marseille, France.*
[b] *Service de l'Information Médicale, Hôpital de la Timone – Adultes, Marseille, France.*
http://cybertim.timone.univ-mrs.fr/cybertim

Abstract

Objectives: to design and implement mediators dedicated to access heterogeneous information sources in a homogeneous way. Method: processes translate a query into the syntax of a target source built thanks to the UMLS knowledge sources and a catalog of information sources. Communication services connect users with information sources at the point they deliver results. Results: examples show the benefits healthcare professionals may found in searching information in this way. Discussion: improvements on the current developments may be done according to the current architecture of ARIANE. Specially, the user interface should be easier to use than the present one.

Keywords:

Information science; Information storage and retrieval; Unified Medical Language System; Internet.

Introduction

The dramatic increase in volume and number of health information sources on the Internet creates problems for healthcare professionals. Usual tools used to search information become inefficient today. Robots produce large listings in which many results are not appropriate (low *precision*) and expected ones are not always present (weak *recall*). Indexes, directories, and portals, that register good quality information sources, tend to reduce *noise*, but a lack of precision remains significant. Mediation with heterogeneous information sources is a recurrent problem the scientific literature addresses [1]. The ease to retrieve information in Internet sites dedicated to health depends on their internal design. Many of them exploit now ontologies or classifications both for indexing pages and links, and for retrieving them. For instance, systems and sites such as CISMeF [2], Cliniweb [3], and HON [4], among others, manually or automatically index pages by means of MeSH and provide users with the capability to browse this classification and retrieve documents. Others, such as MedWeaver [5] and ARIANE [6], exploit the UMLS knowledge sources [7] to allow powerful search capabilities.

The aim of the project ARIANE is to provide healthcare professionals with an efficient access to information sources helpful in their daily practice [8, 9]. Users need to access expected information quickly in a seamless way. Information sources may be either integrated inside the network of their institution, or in web sites on the Internet. In all cases, the aim of ARIANE is not only to connect users with servers, but above all, to query servers directly, and then to connect users with servers at the point where they deliver results. A prototype has been implemented according to a middleware architecture [10, 11]. It consists of:

- a user interface that provides users with the ability to express their queries in a conceptual way,
- a "broker" which aim is to select sources whose descriptions in a catalog match a query,
- mediators that translate queries into the syntax of selected servers and connect users and servers, depending on these latter access characteristics.

At its own level, each component exploits UMLS knowledge sources and an "Information Sources Catalog" (ISC). This latter stores descriptions of resources content and their access characteristics. This is illustrated by Figure 1.

This paper focuses on the capabilities of mediation with various information sources that ARIANE offers to healthcare professionals and on the benefits they may find of searching information in this way.

Method

The architecture of ARIANE must be viewed as a mediation platform with existing information sources. The mediation process exploits navigation and search capabilities of these sources as far as possible. For instance, a target source can be a web site. Some web sites propose to retrieve information in their directories by means of robots. In such cases, ARIANE activates these tools with a user's

query instead of simply redirect him/her to a home page. A database server may be encapsulated inside a web site. In such a case, ARIANE formulates the appropriate SQL query that translates the initial user's query.

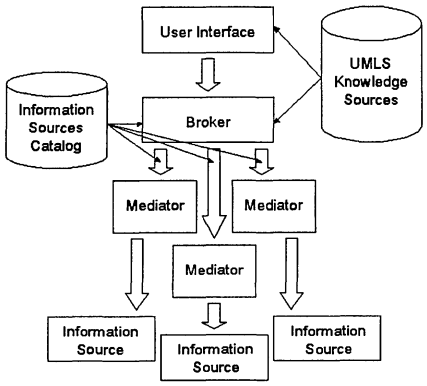

Figure 1. The three-tier architecture of ARIANE.

The way to access a source is a "mediator" built dynamically according to knowledge bases, translation processes, and communication services:

- The UMLS knowledge sources represent the concepts and semantic relationships. The ISC describes the mechanisms used to access a source (communication protocol, network address, ...), and the type of mediator to build.
- A first process translates the concepts and semantic relationships present in a query into the corresponding terms in the vocabulary of a selected source. A second process translates a query into the syntax of the source. A third process formats a query for the exploitation by a communication service.
- Communication services operate the connection to sources. Low level techniques may be used, such as: HTTP, OLE-DB, a file server, ... More sophisticated techniques can be implemented: RPC, SOAP, access to mainframes, ...

In this way, a user's query expressed by means of concepts and semantic relationships is translated into the query language of a source. If this language is rich, then the expected results will be of high quality. This is the case of PubMed, for instance whose query language is based on the MeSH hierarchy of terms. When this language is poor, the quality of listed results cannot be guaranteed. This is the case of web sites whose search tools only offers Boolean operators, and do not treat lexical variants of terms.

The ISC contains for each registered information source both a description of its content and characteristics to access it. The content description of an information source is a list a elementary conceptual graphs of the following type:

[Concept 1] → (relationship) → [Concept 2]

where concepts are either generic types of concept issued from the Semantic Network (SN) of the UMLS either concepts issued from the Metathesaurus (and then attached to at least a type of concept), and the semantic relationship is issued from the SN. For instance, the web site of the French "Association Nationale des Gastro-Entérologues des Hôpitaux Généraux" [12], is indexed by the following graphs:

[*] → (prevents) → [Digestive system diseases]
[*] → (diagnoses) → [Digestive system diseases]
[*] → (treats) → [Digestive system diseases]

that are a representation of: prevention, diagnosis and treatment of gastro-enterological diseases ("*" means any of the concepts involved by the related relationship according to its definition in the UMLS knowledge sources). A site focused on a specific disease, such as the "Association François Aupetit" [13] specially concerned by the Crohn disease, is partly indexed in the ISC by:

[*] → (diagnoses) → [Crohn disease]
[*] → (treats) → [Crohn disease]

When a user asks ARIANE to retrieve information about "the diagnosis of Crohn disease", represented by the graph

[*] → (diagnoses) → [Crohn disease]

the broker exploits the content of the ISC. It finds, for instance, that the two above sites may give an answer. The latter allows a direct answer, when the former is less precise even though its description includes a possible answer since the Crohn disease is a kind of digestive system disease in the UMLS ontology.

To formulate its listing of sources that directly or potentially answer to a user's request, the broker not only compares the query and the sources descriptions, but it computes, on the basis of the UMLS ontology, a semantic distance between the query elements and those involved by the sources definitions. This calculus is based on theoretical works conducted in the field of applications on conceptual graphs to query relational databases [14]. The result of this computation leads to the capability the broker has to propose the relevant sources according to a "conceptual ranking". It prioritizes the sources whose description matches the best a query. The principle of the conceptual ranking is as follows:

- It compares the elementary graphs issued from the query to the graphs that describe each source registered in the ISC, by the mean of nomenclatures present in UMLS.
- It evaluates a "distance" (number of nodes) between concepts in the related hierarchical nomenclatures every time the matching operates successfully.

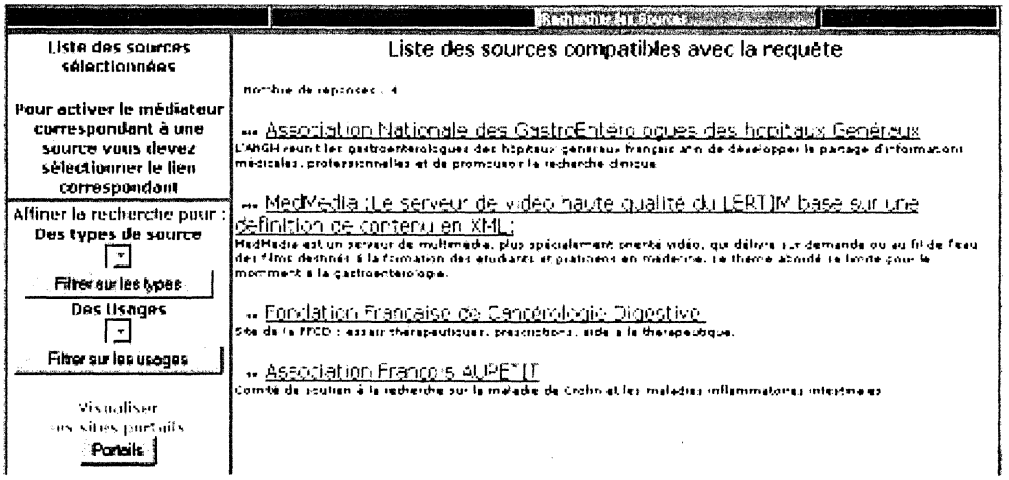

Figure 2. Results listing proposed by the broker of ARIANE to the query "diagnosis of Crohn disease".

- Finally, it takes into account the number of graphs that define a source and that match the query graphs.

Figure 2 illustrates the ranking of links to information sources produced by the broker as an answer to the query "diagnosis of Crohn disease".

Results

The two mediation layers, that are the broker and the mediators, have been prototyped. The operating system used is Windows NT. Software and technology used are[1]:

- Internet Information Server and ASP for the Internet communication and delivery,
- SQL Server for the data storage of UMLS knowledge sources and the ISC,
- COM for the development of components related to mediators,
- Site Server and MS-Search for indexing distant sites and information retrieval inside a local server,
- Windows Media for playing video records.

The graphical interface and the broker have been implemented according to the current, and now well known, technology. The most specific implementation issues concern the mediators that are dynamically built when connection with sources are required. Four modules are involved in the definition of a mediator:

- Vocabulary: the language(s) of the source content .
- Syntax: expression of the query in the language of the source. It is a COM object dedicated to a type of syntax: characters string with Boolean operators and use of parentheses, SQL query, ...
- Electronic format: prototype of the query according to the target source characteristics. It is a COM object : ADO in the case of SQL or MS-Search, SOAP (Simple Object Application Protocol), ...
- Service: the way to connect the source in a network. It is a component of the operating system: HTTP (HyperText Transfer Protocol), SQL Server, MS-Search, ...

The addition of a new source in the ISC requires to define each of these modules. XML (eXtended Markup Language) files are used to customize each of these modules. Figure 3 shows the object classes and associated XML parameters in the case of PubMed.

Discussion

The UMLS knowledge sources provide an operational ontology on which ARIANE internal mechanisms are based efficiently. The addition of the ISC database allows to store information sources descriptions as concepts and semantic relationships issued from the UMLS. These knowledge bases allow ARIANE to provide users to formulate queries in a conceptual way rather than in a purely syntactical way that most of the current robots offer today. Moreover, it facilitates the search process of end-users who may be unfamiliar with the query language of sources since it translates automatically their conceptual queries into the sources languages. For instance, a user unfamiliar with the PubMed query language

[1] Windows NT, Internet Information Server, SQL Server, Site Server and Windows Media are trade marks of Microsoft Corp.

MODULE	CLASS	XML PARAMETERS
Vocabulary	SimpleCustomisableVocabulary	<ARIANEVocabularyDefinition> <Concept> <ConceptCode></ConceptCode> <ConceptDesignationType>TRUE</ConceptDesignationType> <ConceptLanguage>Anglais</ConceptLanguage> <ConceptDesignationSource>MESH</ConceptDesignationSource> </Concept> <Relation> <RelationCode></RelationCode> <RelationSubHeadingsType>TRUE</RelationSubHeadingsType> <RelationDesignationType></RelationDesignationType> <RelationLanguage></RelationLanguage> </Relation> </ARIANEVocabularyDefinition>
Syntax	SimpleCustomisableSyntax	<ARIANESyntaxDefinition> ... <PROTOTYPE>PUBMED</ PROTOTYPE> ... </ARIANESyntaxDefinition>
Electronic format	SimpleString	<ARIANEFormElecDefinition> <SpaceSimplification></SpaceSimplification> <URLEncode>TRUE</URLEncode> <EncodeVerISO></EncodeVerISO> ... </ARIANEFormElecDefinition >
Service	HTTPService	<Service>HTTP</Service> <ADDRESS>http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?term= </ADDRESS> <PARAM1 TYPE=« cmd »>search</PARAM1> <PARAM2 TYPE=« db »>PUBMED</PARAM2> <PARAM3 TYPE=« dispmax »>50</PARAM3>

Figure 3. Object classes and XML parameters involved in the mediator dedicated to PubMed.

could formulate his/her query as "Crohn disease AND Diagnosis". Such a query produces 3,219 references when limited to 1990-99. ARIANE exploits the search capabilities in PubMed definitions and produces the query "Crohn disease/diagnosis", introducing the MeSH subheadings as defined in its PubMed mediator components. It produces 1,844 references during the same period of time, that is to say 42% fewer results.

The middleware architecture of ARIANE and the choices made for its implementation allow to index and access various types of information sources, not only located on the Internet. The design of the broker and of the mediators allow to access various kinds of documents: texts, images, videos, ... The video sequence pointed in Figure 3 (MedMedia server) may be played as illustrated by Figure 4. Moreover, the design of mediators as COM objects and XML files for accessing a kind of resource permits to reuse the components every time a same kind of source must be addressed, but exploiting appropriate parameters.

Conclusion

ARIANE must be considered as a portal of a "second generation" that contributes to a semantic mediation with health information sources, rather than the syntactical search that current robots allow. It must be now validated by numerous end-users. To experiment ARIANE in the daily practice of healthcare professionals, its interface must be extended. Presently, a user expresses a query by means of a sequence of dialog boxes in which she/he selects concepts and relationships. An easier way could be to analyze a sentence written by a user in her/his usual language and to recognize UMLS concepts and relationships. Another way for promoting the advantages end-users may find in the use of ARIANE could be to couple it with an existing health portal, as a "plus".

The analyze of sentences and recognition of UMLS elements should also facilitate the registration of new sources in the ISC. This could be the mean to analyze automatically the description of a source given textually by its authors and to derive its conceptual representation.

Acknowledgments

The authors thank the U.S. National Library of Medicine that provided them with the UMLS knowledge sources. This work has been partly funded by the French Ministry of Education and Research.

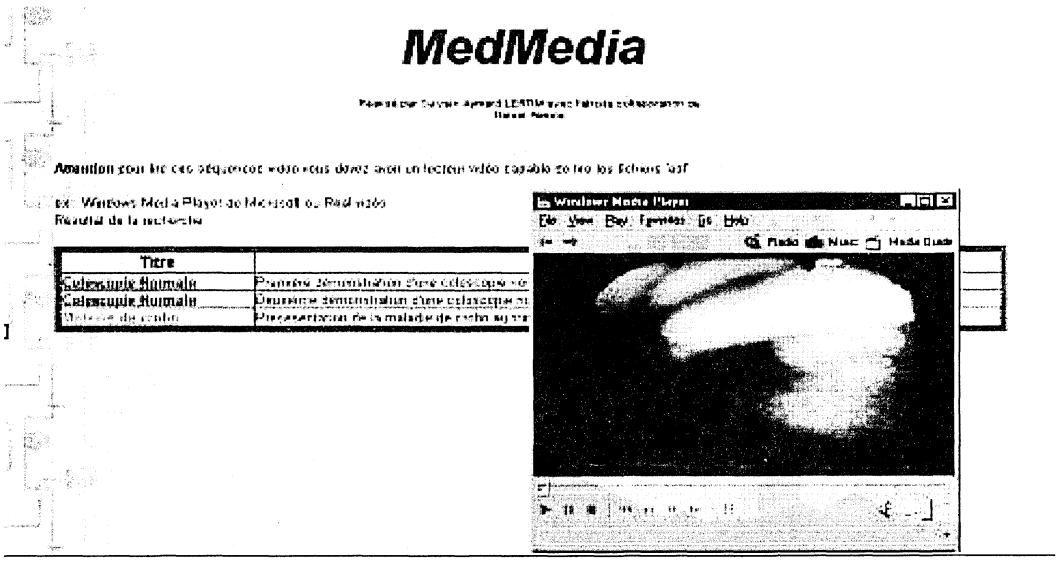

Figure 4. A video clip illustrating the Crohn disease diagnosis as pointed in Figure 3.

References

[1] Wiederhold G. Mediation in information systems. *ACM Computing Surveys* 27; 1995: 265-7.

[2] Darmoni SJ, Leroy JP, Baudic F et al. CISMeF: a structured health resource guide. *Meth Inform Med* 39; 2000: 30-5.

[3] Hersh WR, Brown KE, Donohoe LC et al. Cliniweb: managing clinical information on the world wide web. *J Am Med Inform Assoc* 3; 1996: 273-80.

[4] Baujard O, Baujard V, Aurel S et al. A multi-agent softbot to retrieve medical information on Internet. *Proc. MEDINFO'98* (Cesnik B, McCray AT, Scherrer JR, eds). IOS Press, 1998 : 150-4.

[5] Detmer WM, Barnett GO, Hersh WR. MedWeaver: integrating decision support, literature searching, and web exploration using the UMLS Metathesaurus. *Proc. AMIA Annu Fall Symp* (Masys D, ed). 1997: 490-4.

[6] Joubert M, Fieschi M, Robert JJ et al. UMLS-based conceptual queries to biomedical information databases - an overview of the project ARIANE. *J Am Med Inform Assoc* 1998 ; 5: 52-61. Also published in: *Yearbook 99 of Medical Informatics* (van Bemmel JH, McCray AT, eds). Schattauer, 1999: 500-9.

[7] McCray AT, Nelson SJ. The representation of meaning in the UMLS. *Meth Inform Med* 34; 1995: 193-201.

[8] Volot F, Joubert M, Fieschi, M, Fieschi D. A UMLS-based method for integrating information databases into an intranet. *Proc. AMIA Annu Fall Symp* (Masys D, ed); 1997: 495-9.

[9] Joubert M, Volot F, Fieschi D, Fieschi M. Conceptual integration of information databases into an intranet. *Proc. MEDINFO'98* (Cesnik B, McCray AT, Scherrer JR, eds). IOS Press, 1998: 161-5.

[10] Aymard S, Fieschi D, Volot F et al. Towards interoperability of information sources within a hospital intranet. *Proc. AMIA Annu Fall Symp* (Chute C, ed); 1998: 638-42.

[11] Aymard S, Joubert M, Fieschi D, Fieschi M. Mediation services with health information sources. *Proc. AMIA Annu Fall Symp* (Overhage JM, ed); 2000: 37-41.

[12] Association Nationale des Gastro-Entérologues des Hôpitaux Généraux. http://assoc.wanadoo.fr/angh/.

[13] Association François Aupetit. http://afa.asso.fr/.

[14] Carbonneill B, Haemmerlé O. Standardizing and interfacing relational databases using conceptual graphs. *Lecture Notes in Computer Science*, 35. Springer, 1994: 311-330.

Address for correspondence

Michel Joubert

LERTIM - Faculté de Médecine

27, boulevard Jean Moulin

13005 Marseille. France

E-mail: mjoubert@ap-hm.fr