A conceptual analysis interface for automatic generation of medical surveys

Mourad Ouziri
LIRIS Laboratory
INSA of Lyon
7, avenue Jean Capelle
69621 Villeurbanne Cedex
France
mourad.ouziri@insa-lyon.fr

Christine Verdier
LIRIS Laboratory
Lyon 2 University
5 avenue P. Mendès France
69676 Bron Cedex
Hungary
christine.verdier@univ-lyon2.fr

André Flory
LIRIS Laboratory
INSA of Lyon
7, avenue Jean Capelle
69621 Villeurbanne Cedex
France
andre.flory@insa-lyon.fr

Abstract – **Epidemiological surveys are used to study population diseases. Manual or electronic questionnaries are built, the target population questionned and the results are obtained with statistical tools. Every new study supposes to create again new questionnaries and start again the statistical process. Today, medical records, which contain the medical data and epidemiological studies which need medical data are not linked. We propose in this article a new tool based on Topic Maps representation and description logic used to create both a own medical record and data classification. The data classifications represent a first step for epidemiological study. We complete this step with a document generation tool for epidemiological studies.**

Index Terms—**Conceptual data analysis, description logics, interactive analysis interface, epidemiological surveys.**

I. INTRODUCTION

In health area, information systems (IS) pose problems. The computerization of medical data is being really difficult because of data nature (complex, changing, rapidly evolutionary), organizational challenge (because of numerous places of care and health professionals), and strategic and political aims.

Health IS expanded without coordination, standards, defined transfer protocols, defined data structure and so on. What is the result? Excessive medical data dispatched in different medical IS which do not communicate.

Consequently two main problems are appearing:

(i) the difficulty to create a care process: to replace a data distribution based on places of care by one based on a patient anywhere he is treated;

(ii) the difficulty to aggregate data for a particular study (haemophilia prevalence in youth population), directly from the IS.

A. The care process

Many research projects are dealing with care process and electronic patient records. We can identify two main approaches.

Concerning the care process, the main idea is to create particular workflows to identify the patient's trajectory through the different places of care. This approach is particularly interesting for medical costs aim. In this category, we can cite community particularly interesting in care process [1,2,3] and scientific community interesting in management process [4,5].

Concerning the electronic medical record, several publications have been written since years [6,7,8]. The main ideas are to well-structure medical data in order to create relevant medical records, share medical data and communicate between heterogeneous systems. The numerous medical problems give numerous and multi-axis scientific projects.

B. Epidemiological studies

Epidemiology is a science which concerns the study of diseases, biological and social systems. It defines the diseases frequency and distribution through a statistical population and analyzes the factors of influence. All medical domains concern epidemiology: incidence-based diseases trends, drugs consumption evolution, quality of life, etc.

Epidemiology uses medical data but the medical record are not a really good media. Information needed in a study is hidden in the huge volume of data. So, it is really less difficult to create again the information. Indeed, in a questionnary, the information is immediately accessible. Many projects are dealing with epidemiological issues [9,10].

We build an interface based on medical data sources. This interface is helpful for the user to create his own medical records and for the epidemiologist to produce data classifications. The global representation of the system is shown Fig.1.

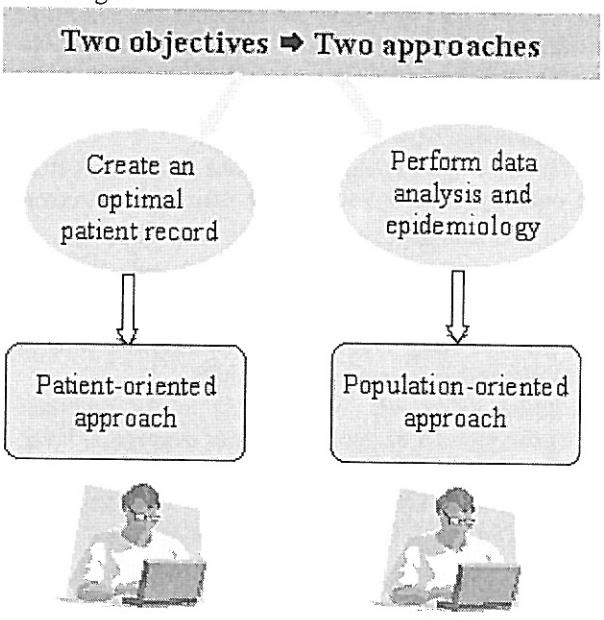

Fig. 1. Objectives and underlying approaches in health information systems

The patient-oriented interface is presented in [11].

We present in this paper the population-oriented approach.

II. THE INTERACTIVE INTERFACE

A. System architecture

The user is in the center of the system. Firstly, he uses the interactive interface to analyze data by introducing stratification criteria in a conceptual graph and builds his own data classification. These stratification criteria are named concepts in our system because they combine stratification criteria ("less than 65 years old"), attributes ("heart diseases") or entities ("doctor"). Secondly, he uses the results of this analysis to define the document structure of medical surveys. A medical survey is generated automatically by creating its corresponding relational database, which is filled-in automatically by the data collected in the document survey.

The system architecture is shown fig.2.

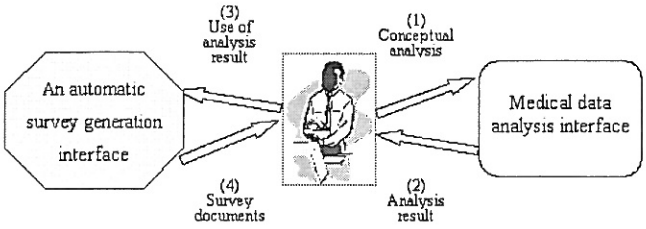

Fig. 2. System architecture for medical data analysis and survey generation.

B. Data representation

To represent medical data, we used description logics (DL) formalism [12]. It provides powerful reasoning at the terminological level like testing the satisfiability of a description, computing subsumption relationship, etc. The DL organizes a knowledge base (KB) into two parts: the intentional part (TBox) and the extensional part (ABox). The extensional part represents the concrete objects, called individuals. The intentional part is an abstraction of the extensional one. It contains the concepts descriptions of the individuals and the definitions of the roles (a role is a binary relationship between two concepts).

In a TBox, complex concepts and roles can be built from atomic ones by using a set of well defined constructors (universal and existential restrictions $\forall$ $\exists$, conjunction $\sqcap$, subsumption $\sqsubseteq$, negation $\neg$, inverse role $^{-1}$). The complex concept Person $\sqcap$ $\forall$ wife.Female $\sqcap$ $\exists$ child.Person denotes persons being able to have the attribute wife of type Female and having at least one child.

Description logics are well adapted to the problems involving the process of the data at their conceptual level. Indeed, it provides powerful reasoning facilities on KB conceptual part (TBox) rather than reasoning on individuals (ABox). The significant facilities in data management are:

- Consistency/Coherence of a description
- Subsumption between two descriptions
- Equivalence between two descriptions
- Disjunction between two descriptions

There are other useful reasoning such as *lcs* [13] (Least Common Subsumer between two or more descriptions), *msc* (Most Specific Concept/description of an individual), *Glb* (Greatest Lower Bound), etc.

The suitability of description logic-based data representation and classification is illustrated through diverse applications. It is used for video objects descriptions and indexing [14], for higher level description of video data (sequence, shot, etc.) and for describing and classifying multimedia objects [15].

III. A CONCEPTUAL DATA ANALYSIS AND AUTOMATIC GENARATION OF MEDICAL SURVEYS

A. Conceptual data analysis

Intelligent information systems represent a combination of intelligent computer-human interfaces and intelligent data integration, management, indexing and querying. The design of such systems is complex. It requires to jointly use databases and artificial intelligence techniques [16]. In accordance to this definition, neural networks [17], data mining [19] and machine learning [18] were used for data analysis and future predicting.

We propose to use automatic reasoning of description logics and statistics to built our interactive interface. The results of an analysis session allow to verify hypothesis about a population and to generate an optimal document of a survey for epidemiological studies.

The data collected from multiple datasources are classified and analyzed at the conceptual level. Data are described by concepts. The user does not interact directly with the data but with a conceptual graph, which describes the data.

The medical data analysis is given by the data distribution for the concepts and the links between them. These two elements are estimated with percentages. Some hidden knowledge and association rules can be inferred by the logic reasoning. The architecture of this part is given in fig. 3.

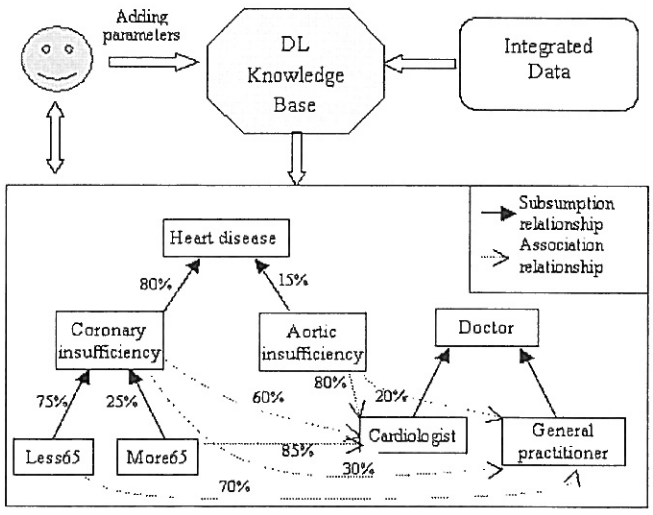

Fig. 3. The interactive interface for medical data analysis

The data used for the analysis purpose are integrated from multiple datasources. We used the Topic Maps [20] knowledge representation technique to represent and integrate the data. Like semantic network, Topic Maps represent data as concepts linked with associations. Thus, the data are previously translated into a description logic knowledge base. This former step is out of the scope of this paper.

At the beginning of an analysis session, the interface presents the concepts of the original knowledge base. These concepts are automatically classified in a hierarchy according to their descriptions. Two types of relationships are created between the concepts: the subsumption relationship and the association relationship. The subsumption relationships are automatically generated using the DL reasoning from the concepts descriptions. Let us consider the following knowledge base:

$$\text{patient} \doteq \text{person} \sqcap \exists\, \text{nss.number} \sqcap \forall\, \text{name.String} \sqcap \forall\, \text{disease.String} \sqcap \forall\, \text{examine}^{-1}\text{.doctor} \sqcap \geq 1\ \text{examine}^{-1} \ldots$$

$$\text{doctor} \doteq \text{person} \sqcap \forall\, \text{id.number} \sqcap \forall\, \text{name.String} \sqcap \forall\, \text{spéciality.String} \sqcap \forall\, \text{examine.patient} \sqcap \ldots$$

The concept patient represents all the persons examined at least once by a doctor and the concept doctor represents all the persons who necessarily examine patients. The two subsumption relationships are automatically deduced:

$$\text{patient} \sqsubseteq \text{person}$$

$$\text{doctor} \sqsubseteq \text{person}$$

We use also the realization reasoning of description logics for data analysis. This reasoning consists to affect an instance/individual to the most appropriate concept. Consider the following ABox:

person(p1)	person(p2)	patient(p3)
disease(p1,"Aortic")	examine(p2,p1)	
	examine(p2,p3)	

The two individuals p1 and p2 are declared in this ABox as persons. The realization reasoning use the assertion examine(p2,p3) to deduce that p2 is a doctor, that is doctor(p2). This deduced assertion is compatible with the ABox assertion person(p2) because doctor $\sqsubseteq$ person (all doctors are also persons). Then, using the assertion examine(p2,p1), the assertion examine^{-1}(p1,p2) is implied and then the realization reasoning deduce that p1 is a patient, patient(p1).

From the previous example, we show that new information can be deduced from the given data. The realization reasoning we use affects an individual to the most appropriate as the most specific concept it belongs to. In the example, the data says that p1 is a person and the reasoning adds more exact information, p1 is also a patient.

The association relationships are calculated from the Topic Map concepts associations and represented as roles (binary predicate that links concepts in DL) in the description logic knowledge base. For example, the *Patient* concept is associated with the *Doctor* concept in the Topic Map, so the role examine and its inverse role examine^{-1} are added to the DL descriptions of patient and doctor.

The conceptual analysis interface consists of a conceptual graph. The user interacts with the interface especially to add new analysis attributes or to ask to show the data distribution of an association. When the user specifies an analysis attribute, the interface creates a new concept (defined with its description), which represents the new class of individuals. For example, the user would get the data distribution about coronary insufficiency patients. The user specify, through a dialog box, the analysis attribute disease = "coronary insufficiency" on the concept *Heart-disease*. The interface adds automatically a new sub-concept of *Heart-disease* and asks for the name of the added concept. The user gives the name *Coronary-insufficiency*, for example, and then the interface creates the description (expressed in the DL $ALC(D)$ [21]):

$$\text{coronary-insufficiency} \sqsubseteq \text{patient} \sqcap \text{one of (disease, "Coronary insufficiency")}$$

and adds it to the KB. The DL reasoning classifies automatically the new concept in the concept hierarchy as sub-concept of *Heart-disease*. The individuals of the new-added concept are automatically calculated with the realization reasoning. This new concept inherits all the characteristics of its parent concept. Thus, this new concept is conceptually associated to the same concepts associated to its parent. As the concept *Heart-disease*, the concept *Coronary-insufficiency* is associated to *Doctor*.

The interface computes the percentage of the subsumption and association relationships from the new distribution of the individuals of the updated knowledge base. That is, the most important analysis information is given by the individuals/instances distribution over the conceptual graph. To add the new concept *Coronary-insufficiency* to the conceptual graph does not give information. At conceptual level, it inherits automatically all the associations of *Heart-disease*. But at the instances level, it can be different. For example, if the *Heart-disease* concept is associated to *Cardiologist* and *Generalist* concepts, the *Coronary-insufficiency* concept is also associated (at conceptual level) to these two concepts. But when we explore data, we can find that all the coronary insufficiency patients consult only generalist practitioners thus, at the instance level, the *Coronary-insufficiency* concept is not associated to the concept *Cardiologist*. This last information is the significant one.

As example, the knowledge expressed in the interface of the figure Fig. 3 are:

--60% of the coronary insufficiency patients are examined by cardiologists and only 10% of them consults general practitioners.

--80% of the aortic insufficiency patients are examined by cardiologists and only 20% of them consults general practitioners.

The first knowledge is insignificant because the percentage does not converge to 0% or 100%. The user refines the analysis by adding an analysis attribute on the old of the coronary insufficiency patients, for example. From the results presented in the figure Fig. 3., the user gets the significant information:

--Most of (85%) the coronary insufficiency patients being more than 65 years old visit cardiologists and those having more than 65 years old visit general practitioners.

As connection to datamining, the percentages calculated

from the concept hierarchy represent confidence of association rules. Indeed, the confidence of an association rule A → B is the conditional probability of B given A:

$$\text{confidence } (A \rightarrow B) = P(B / A) \qquad (1)$$

When we say that 80% of the aortic insufficiency patients are examined by cardiologists, this means that the conditional probability of *Cardiologist* given a *Aortic-insufficiency* is 0,8. That is,

$$P(\text{Cardiologist} / \text{Aortic-insufficiency}) = 0,8$$

Using the relation (1), we get the association rule:

$$\text{Aortic-insufficiency} \rightarrow \text{Cardiologist}$$

and,

$$\text{confidence}(\text{Aortic-insufficiency} \rightarrow \text{Cardiologist}) = 0,8$$

We show that we can discover some association rules and compute their confidence values. These results (association rules) are obtained from the actual data. The system allows to create a survey for two reasons:

--The user needs to confirm the results obtained (which can be considered as hypothesis) with other data samples,

--The data used in the analysis process does not allow extracting needed or desired information. Thus, a survey is generated to gather lacked data. This lacked data will compose the survey documents, which are created using the an automatic survey generation interface.

B. Automatic generation of medical surveys

A medical survey is composed of documents, which will be used by investigators to collect data. As seen in the previous section, the structure (in terms of attribute to be filled) of a survey document is known from the results of the previous data analysis step. The architecture of the system is given in the following figure.

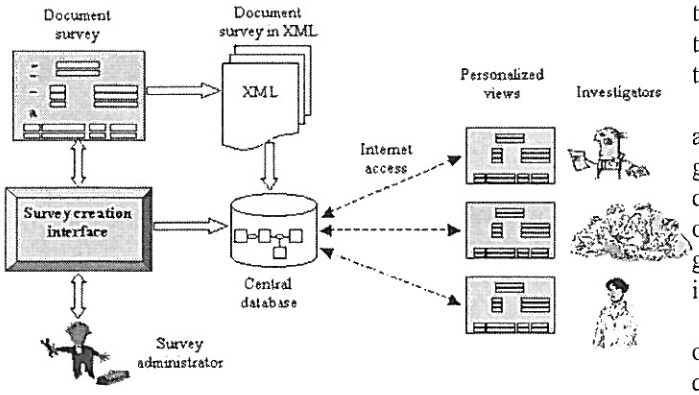

Fig. 4. Architecture for an automatic survey creation

A survey is created as follows:

- The investigator creates the questionnary of the survey using a convivial interface. The screen-copy of this interface is given in Fig.5.
- During an investigation, the investigators collect data and fill-in the document, which leads to many instances of the document survey. Some attributes of some document instances may not be filled. Thus, the instances will not have the same structure. Therefore, we used XML to format these irregular instances of the document survey.
- The interface generates automatically a relational database to centralize the storage of these document instances. For more information about our XML to database mapping approach is presented in [22].
- The interface proposes for every investigator a specific document, which gives a personalized view of the data according to his access rights. This interface proposes some tools to simple aggregate data such as average and variance.

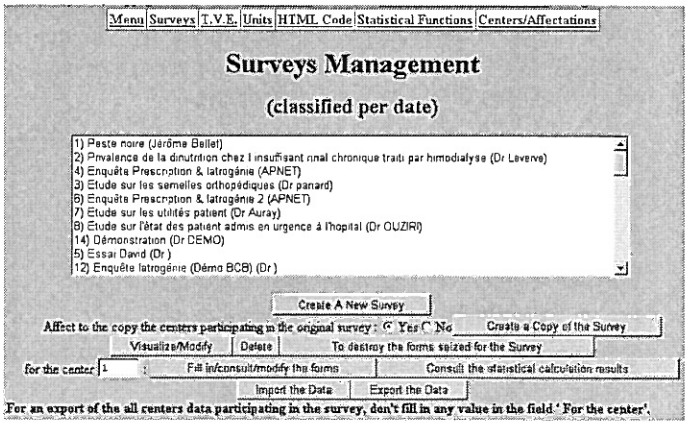

Fig. 5. A convivial interface from document survey creation

IV. CONCLUSION

We have presented in this paper an interactive interface for medical data analysis and automatic survey creation. The data analysis is done at the conceptual level. Then, the data are visualized as a concept hierarchy in which some statistical values are attached to concepts and links. The interface displays the data distribution and the correlation defined by the user analysis. The analysis results are expressed by association rules. The percentages represent the confidence of the association rules. The system is able to produce a new information that was probably hidden in the data sources.

To confirm these association rules or discover other association rules, we have developed an automatic generation of medical surveys system. The results of the data analysis are used to determine precisely the attributes of the survey. A relational database is automatically generated to store the data collected from the different investigations.

Actually, we are implementing an interrogation interface of the survey database. The idea of this interface is to query the database by using the same document survey used to collect the data. A document view is automatically generated from the database in which the user specifies query constraints. A query result is presented as instances of the same document survey.

V. REFERENCES

[1] N. Bricon-Souf, J.M. renard, R. Beuscart, "Dynamic workflow model for complex activity in intensive care unit," *MedInfo 1998, Seoul, Korea,* pages 227-231.

[2] E. Ammenwerth,,et al, "Analysis and modeling of the treatment process characterizing the cooperation within mult-professional treatment teams," *MIE 2000,IOS Press, A. Hasman et al (eds.),* pages 57-61.

[3] P. Dadam, M. Reichert, "Towards a new dimension in clinical information processing," *MIE 2000, IOS Press, A. Hasman et al (eds.),* pages 295-301.

[4] R. Rolland, et al, "L'ingénierie des processus de développement de systèmes: un cadre de référence," *ISI 1996,* vol. 4, n°6, pages 705-744.

[5] E. Tanin, et al, "Facilitating network data exploration with query previews: a study of user performance and preference," *Behaviour and Information Technology, 2000,* vol. 19, n°6, pages 393-403

[6] Richard S. Dick, Elaine B. Steen, and Don E. Detmer, Editors, *The Computer-Based Patient Record: An Essential Technology for Health Care*, The national Academic Press, 256 pages

[7] H. Kindler, D. Densow, H. Mall, T. M. Fliedner, "Visual Access Tool to the Computer-Based Patient Record," *HCI (1) 1997*: 757-760

[8] E. Bayegan, Ø. Nytrø, A. Grimsmo, "Ontologies for Knowledge Representation in a Computer-Based Patient Record," *ICTAI 2002, p. 114-121*

[9] J.G. Phadke and A.W. Downie, "Epidemiology of multiple sclerosis in the north-east (Grampian region) of Scotland--an update," *Journal of Epidemiology and Community Health, 1997,* Vol 41, 5-13

[10] K. Miura, H. Nakagawa, P. Greenland "Height-cardiovascular disease relationship: where to go from here? (invited commentary)," *Am J Epidemiol 2002;155(8):*pages 688-689

[11] M. Ouziri M, C. Verdier, "Utilisation des TopicMaps pour l'interrogation et la visualisation du dossier médical distribué," In *Document Virtuel Personnalisable* DVP2002, Brest- France, july 2002

[12] A. Borgida, "Description Logics in Data Management," IEEE Transactions on Knowledge and Data Engineering, October 1995

[13] F. Baader, R. Küsters, "Computing the least common subsumer and the most specific concept in the presence of cyclic ALN-concept descriptions," *Proceedings of the 22nd Annual German Conference on Artificial Intelligence, KI-98*

[14] J. Carrive, F. Pacher, R. Ronfard, "Using Description Logics for Indexing Audiovisual Documents," *Proceedings of The International Workshop on Description Logics (DL'98), Povo-Trento, Italy, pages 116-120*

[15] J. Fan,X. Zhu, M-S Hacid A.K. Elmagarmid, "Model-Based Video Classification Toward Multi-level Representation, Indexing and Accessing," *Multimedia Tools and Applications, 17(1), pages 97-120*

[16] P. Bresciani, "Some research trends in KR&DB," *KRDB 1996*

[17] R.W. Brause, "Medical Analysis and Diagnosis by Neural Networks," *In ISMDA 2001, LNCS 2199,* pages 1-13,

[18] D. Riano, S. Prado, "The analysis of hospital episodes," *In ISMDA 2001, LNCS 2199,* pages 231-237

[19] S.E. Brossette, *et al.* "Association rules and data mining in hospital infection control and public health surveillance," *In Year book of Medical Informatics 2000,* pages 134-142.

[20] ISO/IEC 13250, "Topic Maps," *Dec ISO/IEC FCD, 1999*

[21] F. Baader, P. Hanschke, "A Scheme for Integrating Concrete Domains into Concept Languages," *IJCAI 1991:* pages 452-457

[22] M. Ouziri, C. Verdier, "XML storage in relational databases: An approach combining description logics and statistics," *IRMA International Conference - Track on Text Databases & Document management, May 18-21, 2003, Philadelphia, PA*