Hospitexte: towards a document-based Hypertextual Electronic Medical Record

J. Charlet*, Ph.D., B. Bachimont†‡, Ph.D., V. Brunie*‡, M.Sc., S. El Kassar*,M.Sc.,
P. Zweigenbaum*, Ph.D., J.-F. Boisvieux*, M.D., Ph.D.
* Service d'Informatique Médicale, AP–HP & Dép. de Biomathématiques, Univ. Paris 6, France
† Institut National de l'Audiovisuel, France
‡Univ. de Technologie de Compiègne, France

The patient record is a repository for knowledge about a patient. Work in Artificial Intelligence and knowledge representation has evidenced the intrinsic difficulty of formalizing knowledge for computer processing. It is therefore not a surprise that most attempts at computerizing the patient record have only had a limited degree of success or applicability. We claim that this is due to the fact that medicine is an empirical domain, and thus fundamentally resists formalization. Therefore, the only way medical knowledge can be fully expressed is through natural languages which is indeed what clinicians actually use. We proposed and designed an electronic medical record which adheres to this hypothesis and where structured documents play a prominent role.

INTRODUCTION

Documents are the core of medical reflection: in a Paper-based Medical Record (PMR), we found a lot (typically 150-200) of documents which give to the health-care professional board all the information necessary to follow up and treat the patient. These documents are of different natures (paper, image, etc) and come from different sources (laboratories, clinical departments, etc). Moreover, laboratory results come with textual reports which give important information about the exam. Assuming that we want to change the support of the PMR for an Electronic Medical record (EMR), few effective technologies may propose an appropriate solution.

In a traditional approach which tends to formalize knowledge, a view of the EMR is that of a database which holds items of coded, or standardized, information. In a different way, effective technologies like hypertext offer an opportunity to develop new systems based on the concept of "document processing". Our claim is that if we want to allow the clinician to get used to and work with an EMR we must adopt such a point of view in which information is described in natural language.

In the next section we argue for a document-based hypertextual electronic medical Record; we then present the HOSPITEXTE prototype and the research framework necessary to our approach. Finally we discuss some issues and future development of such an approach.

FORMALIZATION VERSUS DOCUMENT

The traditional solution of a computer approach of the EMR consists in formalizing knowledge. With this formalized knowledge it might be possible to run an artificial intelligence system which uses knowledge as data and provides conclusions to help a clinician in his diagnosis task. For an EMR, this approach leads to a traditional view where a database holds items of coded, or standardized, information. This approach has achieved some success. Nevertheless, it does require a substantial effort of standardization over medical information, and a complete formalization of medical information is theoretically not reachable. Moreover, experiences has shown that this point of view suffers from a lack of meaning. Two major reasons go against this approach[1]:

- Medicine is not a science in its daily practice. It is a practice in which context palys a major role. To force a clinician to express his knowledge in the form of records and fields is awkward. Moreover, each document may be considered as providing the context for the sentences it contains. For instance, if the sentence "the symptoms are attributed to an overdose of phenobarbital" appears in a clinical examination report, it is a hypothesis balancing several clinical observations. If it appears in the conclusion of a blood barbiturate measure result, it is the confirmation of a hypothesis. This is why medical information cannot be easily extracted from the documents which convey it.
- Medicine continues to evolve, and many of the medical categories, procedures and vocabularies change with time.

These requirements argue against data models which tend to impose a unique, universal, structured data encoding the patient. They do not argue again databases as systems which manage files securely and allow to record and gather data in order to perform (for example)

1091—8280/98/$5.00

epidemiological studies. Moreover, as we will see in a later section, the document-based approach also allows to build data structures devoted to particular applications – i.e. corresponding to particular interpretations of medical information. However, it does not compell to do so.

An Artificial Intelligence debate

This opposition between contextual, uninterpreted information as it may be found in the medical document and interpreted information as it is found in a rigid data model brings back the debate which holds in the "Knowledge Acquisition" community. In this community, when talking about Knowledge-Based Systems some discussions remain about "domain knowledge" and its near independence from the task assigned to the system. In this debate, it appeared that the concept of "ontology" – i.e. all the pertinent concepts of a domain and/or for an application – is at the heart[2]: is it possible to elucidate a domain ontology which may be described independently from the task? In contrast to the CYC approach[3] which is an attempt to build an universal task-independant ontology, we argue that we can tend towards the independence but that it is never fully reached[4].

Returning to the EMR and following the ontological debate, (i) we set down that the task of a clinician using an EMR is not unique. For example, Nygren & Henriksson[5] have elucidated four goals in the reading of an EMR by a clinician: *getting a fast overview and understanding of a case, triggering of a memory-picture, searching for facts (using the record as a dictionary), problem solving*. Moreover, (ii) we argue that it is impossible to reach a constrained data model which may be adapted to all medical tasks and therefore to all medical contexts. Only the textual, free-form document can give the uninterpreted – i.e. re-interpretable – context necessary for a clinician to capture the full range of clinical information[1, 6].

Nevertheless, some challenges remain in developing a document-based approach: even though the PMR may have a poor organization and is not always available, it is familiar to the users and they are satisfied with it[7]. In developing a hypertextual document-based EMR we must be careful to give the end users tools and facilities in order to access information without being lost and to work with this information.

Marking up the document

We wish to stress the fact that we need to choose a means to mark up the contents (structure, information, annotation, etc) of document. Managing structure and annotations in a text can be realized by inserting in it tags which mark this structure and identify the annotations. The publishing industry has pushed forth an initiative which led to the adoption of an ISO standard for this purpose: the Standard Generalized Mark-up Language (SGML)[8]. SGML has since then spread to a wide variety of industries and research disciplines. HTML (HyperText Mark-up Language), the language of the World Wide Web, is itself a fixed application of SGML.

SGML basically allows to define the nature and structure of the "tags" which can be inserted in a document. Tags delimit "elements" of a document, i.e., categorized segments, such as a section title, paragraph, diagnosis or person name. Tags may also associate "attributes" with elements, such as the level or number of a section title, the sex of a person or the standardized code for a diagnosis. The set of allowed tags in a class of documents and their authorized order of appearance and nesting are specified in a "Document Type Definition" (DTD), also written in SGML.

HOSPITEXTE: AN EXPERIMENT

Hospitexte Overview

HOSPITEXTE is a project which follows the DOME project (partly funded by the European Union, MLAP #63-221). The major result of DOME was the elucidation of the needs of the health care professionals for a multimedia, hypertextual document-based EMR with browsing functionalities. These needs were materialized in a running mock-up of a document-based EMR using WWW technology[9]. In the HOSPITEXTE project we follow two objectives:

Creation of a virtual patient record This objective addresses the problem of the reconstruction of a unique EMR for a patient from information that may reside at several sites and associated issues such as security and access control;

Design of a professional reader's workstation Here, we address the problems of the disorientation of the reader in front of a hypertextual interface and his capacity to work with it.

The solution to the first problem lies in the architecture of the system: (*i*) building a distributed hypertextual document-oriented EMR and (*ii*) using an "Intranet" approach with a WWW client-server architecture. The second problem will be discussed at the end of the paper.

Hospitexte in practice

A project like HOSPITEXTE needs to be connected to a Hospital Information System in order to be provided with patients indexes, biological laboratories results, im-

agery results, etc. This will shortly be the case in the hospital "La Pitié-Salpêtrière" (France).

The practical objective of the project is the realization of an industrial prototype in a clinical department. The first result is the realization of a research prototype (with all the functionalities) of a readable EMR. This prototype has been designed to provide:

A full document centerd approach This means that the system manipulates documents and presents documents to the user;

Structured documents Every document must be stored under a structured representation, *i.e.* it must belong to a type and have a structure conforming with this type. It must be easy to add a new document type and its structural description in a declarative way;

Dynamic synthesis tools Structuring documents must be dynamically generated, according to record contents. Dynamically-generated documents are submitted to the same structural constraints as other documents.

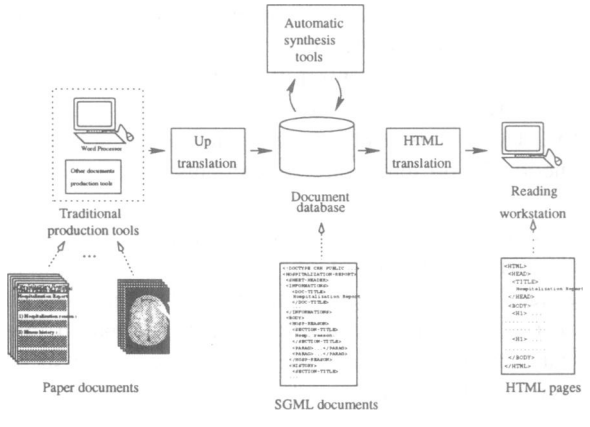

Figure 1: Hospitexte Architecture.

The architecture of the EMR we have developed is shown on figure 1. Documents are produced by the usual production tools used in the hospital, essentially word processors and automata. These traditional tools produce proprietary format documents which must be translated to their SGML representation. This step is known as *up-translation*, emphasizing the fact that physical layout mark-up must be raised up to structuring mark-up. SGML documents are then stored into a document database (currently implemented with the computer's file system). It is then complemented with new documents built with synthesis tools. Every document contained in the database is then translated into HTML to be displayed on a reading workstation. An excerpt of an anonymized EMR is browsable at `http://www.biomath.jussieu.fr/hospibin/patients.tcl`.

Each automatic synthesis tool uses only documents of a pre-defined set of types as inputs. At this time, automatic syntheses are pre-calculated every time a new document is added to the database, so that every synthesis tool using a given type as input is activated on the addition of a new document of this type.

In terms of structured documents, the major result of our work is the description of fifty document types which cover almost all the documents in the medical records of a pneumology department. These types are grouped in a general DTD for the EMR, MedDoc. The medical records of 10 patients are now recorded in the HOSPITEXTE experiment. This corresponds to 2,000 medical documents.

HOSPITEXTE : DIRECTIONS

The problems encountered in working with an hypertextual EMR may find solutions which deal principally with *hypertextual semantics* and *Medical Language Processing* (MLP).

Hypertextual Semantics

Hypertextual semantics is a generalization of textual semantics. According to the latter, linguistic units have their meaning determined by their position in a textual unit: for example, a linguistic unit given in an introduction does not have the same meaning as the identical unit given in a conclusion. According to the hypertextual semantics, document units have their meaning determined by their position in the hypertextual network. In our document centered approach, the position in the hypertextual network is defined by the SGML tags[8]. As a consequence, hypertextual semantics is for us the semantics of the SGML mark-up. SGML defines a syntax for hypertextual structure for which hypertextual semantics aims to provide a meaning.

While defining what a textual unit means according to its SGML mark-up – *i.e.* according to the position of the mark-up in the global hypertext DTD – the hypertextual semantics defines which structural transformations of SGML instances are meaningful. For example, it is meaningful to dynamically build a table of every risk factor marked up in the EMR, as a synthesized document. Such a document is meaningful because its construction relies upon the meaning of SGML tags. Thus hypertextual semantics is a semantics of the structure, that therefore we call *hypertextual structural semantics*. It must be understood that the structural semantics re-

flects in its rules the standard meaning or usual semantic value of textual units according to their hypertextual position. Finally, in elucidating structure of documents we are able to build synthesis documents which include knowledge which answer to "standard" questions/tasks.

Medical Language Processing

MLP programs can be applied to a document to make its information content more explicit. The basic idea is that MLP can identify that some text portions have a specific meaning. This results in a general categorization of the identified text portion (*e.g.*, it is a diagnosis), or even in a precise data representation, such as can be expressed in a coding system[10, 11, 12].

However, instead of replacing the text with the extracted "data", we propose to use it to annotate the document (see Sager and colleagues[13] for an exemple). This adds *content-based* annotations to the document (*structure-based* markup is dealt with in the up-translation process at a relatively low cost). Once "information" has been extracted, the original text is not "forgotten". It is still the reference that the reader of the patient record may want to consult when examining this patient's data, and that we keep in our EMR.

General natural language processing architectures where components work on an SGML encapsulation of natural language data have been defined by projects such as Multext (see, *e.g.*, `http://www.lpl.univ-aix.fr/projects/multext/`). The basic flow of control consists in selecting input segments, passing them to the MLP program, and then inserting the results back into the enriched document structure. A typical processing chain could be: select specific sections of a discharge summary (*e.g.*, admission, previous history, conclusion); run an encoding program, which delimits diagnostic expressions and produces a code (*e.g.*, ICD) for each of them; embed each diagnostic expression in a `diagnosis` element, and store the resulting ICD code into an `ICD` attribute of the `diagnosis` element.

One advantage of this approach is that several interpretations of the same text can be added monotonically to the same document. For instance, one MLP process could determine the ICD encodings of the relevant expressions in the text, whereas another one could identify and categorize drug dosage information[14]. Moreover, an analysis of only some parts of a text can be handled, and the actual piece of text which gives rise to a specific analysis can be precisely identified. For instance, a diagnosis will not have the same meaning if it is presented as *the* diagnosis of a patient (attached to the whole report) or as a diagnosis found in the "reason for admission" section or in the "previous history" section. Finally, search criteria operating on several levels of interpretation can be combined together, resulting in greater search power. For instance, one can search for a given diagnosis code in a specific section of a discharge summary — and search or view the corresponding original text.

Note that navigation in a MLP-enriched document can rely not only on explicit text and links, but also on the attached, underlying MLP-produced data. For instance, given the appropriate tools, one can navigate from one diagnosis to the next in a text or search for text segments whose attached analysis at a given level satisfies some criterion (e.g., find all occurrences in the text of conditions affecting the lower limbs, based on a SNOMED encoding "hidden" in the annotations). This opens up a whole range of dynamic navigation mechanisms[15].

DISCUSSION

The enriched document paradigm presented here relies on an SGML encoding of medical text and data. As mentioned earlier, SGML is making its way into the medical informatics community[6] (see also `http://www.mcis.duke.edu/standards/hl7/committees/sgml/`). We would like to stress two important specificities of our proposal. First, we use SGML to tag the logical content of documents, rather than their external presentation. The actual presentation and layout of documents must be determined by separate style sheets which pilot the final translation from SGML to HTML. This contrasts with methods which directly encode the presentation of documents with HTML*.

Second, we are concerned with partially structured information: a mixture of text and data, where natural language itself can undergo some variable degree of structuring. This feature of traditional (paper-based) medical records is considered essential by authors such as Nygren[5]. This is different from approaches where structured data is accessed and displayed through HTML front-ends[16]. Regarding the HL7 initiative, we agree with Alschuler *et al.*[6] and we can say that *SGML and SGML-based tools need not replace existing technology and need not be seen in opposition to other standards initiatives within medical informatics.* We think that a document-centered approach is the best way to capture all the breadth of medical information and that it can help us to build better and numerous epidemiological databases from EMR.

*It seems that XML will be the new standard of the *document engineering* on the Web (`http://www.w3.org/XML/`) but this does not change the spirit of our propositions. Moreover, the translation from SGML to HTML would be avoided: the structured document would respect XML DTDs and would be displayed according to XLS style sheets.

FUTURE DIRECTIONS

The design of a professional reader's workstation goes through annotational tools. With respect to *hypertextual structural semantics* (cf. supra) we can say that SGML tags reflect the standard meanings of textual units, but not the particular meanings projected by a given reader. However, these personal meanings are of primary importance for the legibility of a hypertext. A reader must be provided with some means to express on the hypertext what textual units mean according to his own hypertextual semantics. Annotation tools enable a user to tag the textual content and attribute them a hypertextual value. He can also use some tools to manipulate (organize into a hierarchy, aggregate, etc.) the tagged units and build his own reading of the browsed documents. These annotational tools may be theorized within an *annotational hypertextual semantics*. By allowing the user to annotate documents we are able to build synthesis documents including knowledge which answers personal questions/tasks.

Finally, the definition of DTDs for medical documents is an independently motivated goal for the document-centered medical record[17] which should be taken in charge by scientific societies of each discipline. Under these conditions, (i) choosing tags for documents is not the task of an isolated user, and (ii) it will be possible to exchange structured documents between the physicians who want to communicate their medical patient data.

Acknowledgments

We would like to thank Dr. T. Similowski (Pr. Derennes Dept.) for his medical contribution.

References

1. Lincoln TL, Essin DJ, Anderson R, and Ware WH. The introduction of a new document processing paradigm into health care computing – a CAIT white paper. Technical report, Los Angeles County + university of Southern California Medical Center, 1994. Available at `dumccss.mc.duke.edu/standards/SGML/proposals/CAIT-white-paper.txt`.
2. van Heijst G, Schreiber AT, and Wielinga BJ. Using explicit ontologies in KBS development. *International Journal of Human-Computer Studies* 1997;45(2/3):183–292.
3. Guha RV and Lenat DB. Cyc: A mid-term report. *The Artificial Intelligence Magazine* 1990.
4. Bouaud J, Bachimont B, Charlet J, and Zweigenbaum P. Methodological principles for structuring an "ontology". In: IJCAI'95 Workshop on "Basic Ontological Issues in Knowledge Sharing", August 1995.
5. Nygren E and Henriksson P. Reading the medical record. I. Analysis of physicians' ways of reading the medical record. *Computer Methods and Programs in Biomedicine* 1992;39:1–2.
6. Alschuler L, Dolin R, and Spinosa J. SGML in healthcare information system. In: Proceedings of GCA SGML Europe'97, 1997:195–204.
7. Tange HJ. The paper-based patient record: Is it really so bad? *Computer Methods and Programs in Biomedicine* 1995;48:127–31.
8. Goldfarb CF. *The SGML handbook*. Oxford University Press, 1990.
9. Bouaud J and Séroussi B. Navigating through a document-centered computer-based patient record: a mock-up based on WWW technology. *Journal of the American Medical Informatics Association* 1996;3(suppl):488–92.
10. Sager N, Lyman M, Nhàn NT, and Tick LJ. Medical language processing: Applications to patient data representation and automatic encoding. *Methods of Information in Medicine* 1995;34(1/2):140–6.
11. Gundersen ML, Haug PJ, Pryor TA, et al. Development and evaluation of a computerized admission diagnoses encoding system. *Computers and Biomedical Research* 1996;29(1/2):351–72.
12. Zweigenbaum P, Bachimont B, Bouaud J, Charlet J, and Boisvieux JF. A multi-lingual architecture for building a normalised conceptual representation from medical language. *Journal of the American Medical Informatics Association* 1995;2(suppl):357–61.
13. Sager N, Nhàn NT, Lyman M, and Tick LJ. Medical language processing with SGML display. *Journal of the American Medical Informatics Association* 1996;3(suppl):547–51.
14. Evans D, Brownlow N, Hersh W, and Campbell E. Automating concept identification in the electronic medical record: an experiment in extracting dosage information. *Journal of the American Medical Informatics Association* 1996;3(suppl):388–92.
15. Zweigenbaum P, Bouaud J, Bachimont B, et al. From text to knowledge: a unifying document-oriented view of analyzed medical language. *Methods of Information in Medicine* 1998;37(4).
16. Cimino JJ, Socratous SA, and Grewal R. The informatics superhighway: Prototyping on the World Wide Web. *Journal of the American Medical Informatics Association* 1995;2(suppl):111–5.
17. Séroussi B, Baud R, Moens M, et al. DOME final report. DOME (EU Project MLAP-63221) Deliverable D0.2, DIAM — SIM/AP-HP, 1996.