Bridging the gap between medical narratives and structured data in the computerized patient record

Christian Lovis, Alexander Lamb, Anne-Marie Rassinoux, Antoine Geissbuhler

Division of Medical Informatics, University Hospitals of Geneva

Corresponding author:
Prof. Dr. med. Christian Lovis
University Hospitals of Geneva
21, rue Micheli-du-Crest
CH-1211 Geneva 4
Switzerland

Phone: +41 (0)22 372 6201
E-mail: christian.lovis@hcuge.ch

Abstract

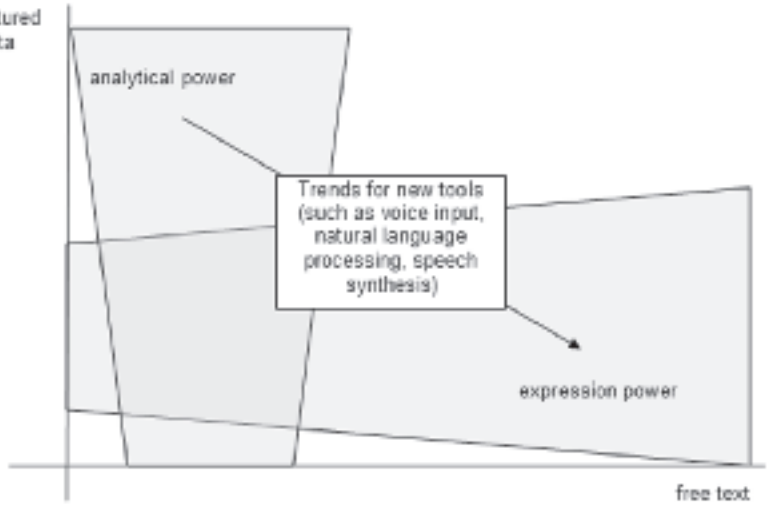
Care documentation is a central activity of care delivery and a mandatory step to the development of predictive and supportive care informatics in a collaborative paradigm. Beside its importance for classical data processing in healthcare such as reimbursement claims, scientific research or teaching, care documentation must also fit within the daily work of healthcare providers without intrusion and remain a precise and life biography of the patient. In this view, human-machine interfaces and philosophy behind data acquisition and restitution interfaces are of major importance. There have always been some antagonisms between narratives and structured data entry, both having advantages and disadvantages, supporters and detractors. In real practice, most documents used in clinical settings are made both of typed or structured data and narratives or free texts. In order to try to have a common source for all these information, we developed a unified representation, acquisition and storage system for medical information. To use this system in our computerized patient record, we use a middleware based on HTTP and XML that permits standardized exchanges between applications and data repositories. This paper is devoted to the description of some part of our system as well as its real implementation in a CPR working in the five Geneva University Hospitals.

Introduction

We do present a concept that permits the unification of semi-structured information such as what can be found in free texts and strictly typed data such as fields in questionnaires, allowing a common representation at several levels, from user interfaces to data storage. This concept is currently being implemented and already used at the university Hospitals of Geneva.

The ultimate goal of documentation is to provide accurate and timely clinical information for patient care and complete documentation for all stakeholders [1]. The diversity of stakeholders is a characteristic of a large health care organization. Each professional group deals with its personal and often partial view of the global system, with its knowledge, culture and terminologies. Moreover, the same professional can have multiple roles: a nurse can be planning the care of a patient, checking vital signs at the bedside or rounding with physicians. In each role, the information needs and the ergonomics of the system will be different: a computer at the nurses' station would be necessary for the care planning while a handheld device would work best at the bedside. However, computer-based systems have historically been built to help each professional group to deal with their information needs, thus respecting or reinforcing the functional boundaries of the groups. These systems have then been interfaced to others in order to form hospital-wide information systems, generally in a way that mimicked existing paper-based transactions. The development of healthcare delivery networks is increasing the diversity of different stakeholders, thus complicating the task of integration. The need for resource management and integrated clinical pathways requires a transversal understanding both of care structures and of data representation, management and acquisition. The main challenge is the ability to represent the knowledge in a way that is usable, maintainable and meaningful to diverse users. However, this implies numerous requirements that are sometimes in contradiction. One of these apparent contradictions concerns the discussion of structured data versus free texts [2, 3]. The underlying problem is far

from trivial and major consequences may have to be faced in the long term. On one extreme, one could consider to have a CPR by just scanning all paper-based documents. Such a system would answer immediate



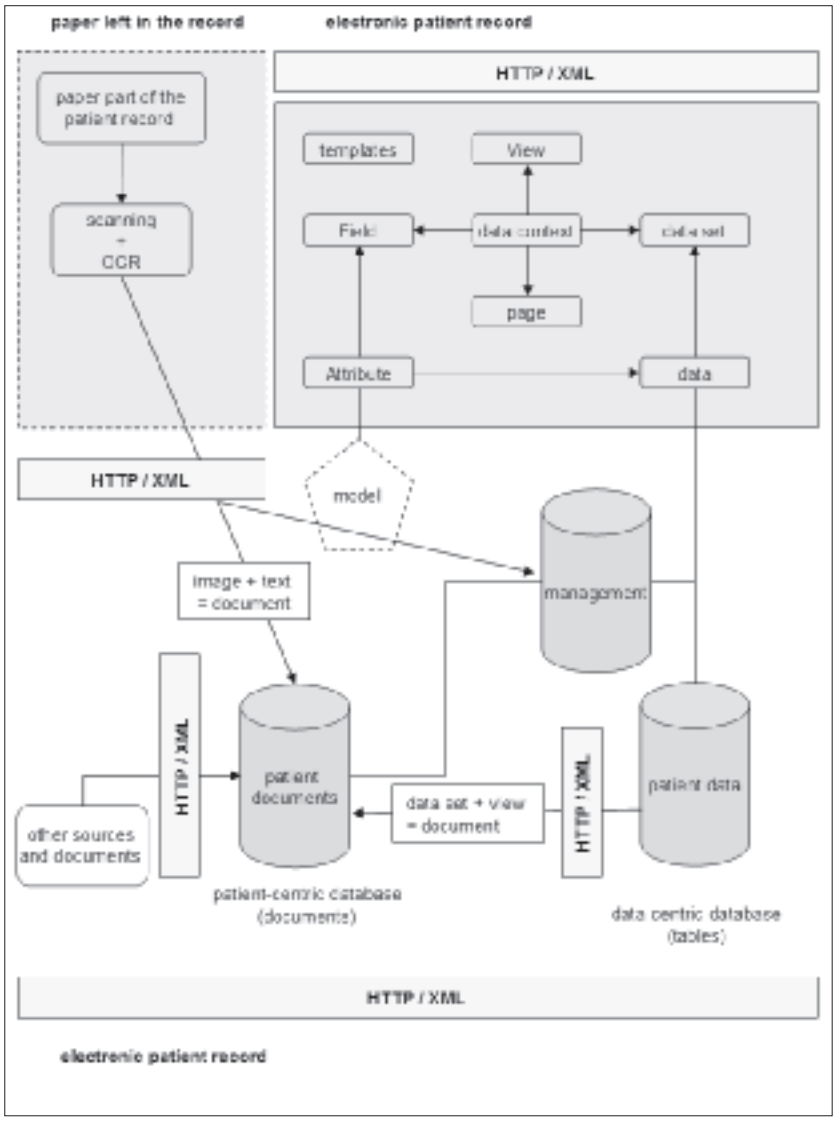
and important needs, such as ease of access, ubiquitous record, centralized management and right accesses, etc at reasonable cost. However, it would obviously not answer some important needs such as data processing and analysis, decision support or clinical pathways. At the other extreme, a patient record could be entirely model driven with a "Point&Click" user interface, such as developed within the Pen&Pad project [4]. In this system, a complete patient record can be mapped to a semantic model and the user interface for patient data acquisition is driven by the underlying model. Between these two extremes, many clinical information systems (CIS) do have to face with the coexistence of both documents and structured data management. There is a challenge in trying to bridge the gap between structured data and free text at several levels, from data acquisition to information processing and knowledge representation.

- *Structured versus narrative data acquisition (questionnaires versus documents)*
  There is a long tradition of controversy between the advantages and disadvantages of using structured data entry compared to free text. The two major arguments underlying the use of structured data entry are a) the analytical power of structured data and b) the ease and speed of data entry. Both points are of high importance in medicine, as they allow analysis of data, decision support, knowledge coupling, etc. On the other side, free text entry as a far better power of expression, allows layout formatting such as emphasis and is easier to use in numerous situations [5, 6].
- *Closed versus open user interfaces (Menus versus command-line)*
  historically, one have moved from commend-line driven systems, such as MS-DOS to menus and dialogs driven which are at the core of MS-Windows, Apple MacOS, etc. The explosion of computers proves better as anything that menus and dialogs driven-systems are easier to use. However, this does not prove to be true when very large number of items must be available, such as classification, order entry, etc. Several systems offer now again command-line entry, or their "modern" equivalent such as completers, natural language analyzers, parsers, etc. that allow users to type in a way more or less similar to free text and have the system interpret the entry.
- *Patient oriented versus data-type oriented storage (patient record versus data warehouse)*
  Most computerized patient record are built upon relational databases, which are most often structured regarding data type attributes rather than patient centric. So, for instance, there might be some tables for laboratory data, some other for administrative information, some other devoted to drug prescription, etc. Accesses to the whole record of any patient might require consolidating data coming from numerous tables or databases. On the other hand, some document oriented system can store all data of any single patient at a unique place, therefore facilitating and speeding accesses for patient-centric queries, but making cohort

studies almost not possible, or at the cost of high processing power.

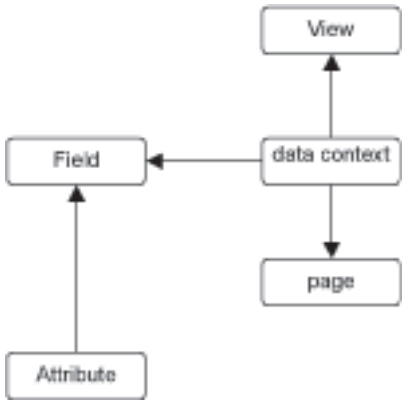
Merging questionnaires and documents

In a move to merge document-based medical narratives and questionnaires data acquisition, we consider that questionnaires and free texts documents have the same structure and representation; they are built upon a description of fields. Two different user interfaces allow the acquisition of data. Both do allow the acquisition of structured and typed data as well as blocks of free texts. One interface has a strong emphasis on layout control and is more devoted to such documents as discharge letters or reports. The other is more devoted to fast data acquisition and has more a structure of questionnaires. Both interfaces do use the same data representation and same repositories and all transactions are made using a common http/xml based middleware. Questionnaires use mostly basic attributes types, such as Boolean, dates, lists or numeric whereas documents are mostly formed using paragraphs. Documents have therefore been highly structured based on their paragraphs, such as patient history, discussion, conclusions. All documents and questionnaires do share a set of similar characteristics used for document management and workflow, privacy control and versioning.



Data model

The data model is based on the idea that all documents are built using a set of attributes (the data fields) grouped with a document class (the data context) (see figure 2). The model is separated from the references to attributes. It is based on the idea that all documents are built using a set of shared attributes with a given type and meaning. Therefore, the fields of a document are only references to attributes. The values are represented as lists of attribute-value pairs connected to an instance of a document for a given patient at a given date. Once filled,



the resulting values will be stored in a separate repository, which contains attributes-values pairs. The elements of figure 3 are explained thereafter:

- *Data context*
  Logical model that regroups fields. It is, in fact, the list of all fields needed to create a questionnaire, a document or a specialized record such as the record of anesthesiology consultation.

- *Field*
  It is the label of an attribute in a questionnaire or a document. Each field is linked to a specific attribute. For example, in a given questionnaire, it might be a field "Smoker?" linked to the attribute *tobacco_use*
- *Attribute*
  An attribute is one entry in the common dictionary of classes of facts. For each attribute classes, several properties can be specified, such as data type, description, links to external models, etc. Within the dictionary, each attribute has a unique internal identifier that cannot be removed. Each attribute can have dates of validity, so that attributes are never deleted but only inactivated. Each attribute belongs to one of the seven basic data types, which are *enumerated, date, decimal, image, integer, long text* and *short text.* The values that can take an attribute can optionally be limited by a *code value*, such as "Yes", "No". If needed, these possibles values can be aggregated in *Groups* such as ["Yes", "No"] for example. *Attributes* can be linked to external classification, and numerous are already linked, to ICD10 for example.
- *View*
  A *view* is a way a *data context* will be displayed. It does not define which *fields* (and linked *attributes*) are used, but the format, layout, authorization schemes and components used for display. An example of *view* is PDF for Adobe Acrobat Portable Document Format®. Another of the *Views* of a *data context* could be a Microsoft Rich Text Format *template* that can allow complex editing and layout.
- *Page*
  Page allows to group fields and will be used to produce automatically acquisition user interfaces. It is useful to avoid this kind of very long document that must be scrolled. A *page* can be assimilated to panels in a web-based questionnaires or to pages in a Word document.
- Once a *data context* has been instantiated, such as a questionnaire or a document for example, *attributes* will receive *data*, that is, the attributes' *values*. All *data* of a *data context* represent its *data set*. A link is maintained between *data* and *attribute* to allow *attribute-values* to be retrieved, as well as between *data set* and *data context*, so that the whole context in which data have been acquired is kept.

Semantic model

A semantic model acting as a "semantic" shield over the list of attributes has rapidly proven to be necessary with the increase of the size of the number of attributes and the apparition of synonyms or duplicates for example. One of the main problems encountered has been to have a way to create rapidly new attributes without loosing the added-value of a semantic representation. However, there is a large pressure for having new attributes fast whereas organizing these attributes in a semantic model can take long time and discussions. We therefore completely separate models, the data model and the semantic model. The dictionary of attributes can be considered as a flat list, or almost equivalent. However, the process of building a model on the top of this dictionary is ongoing. Attributes are linked to concepts with only four relation types:

- *isA.* It is a subsumption relation that can be used in various cases, such as *cl_femur isA cl_bone* or *cl_headache isA cl_pain*.
- *partOf.* is the usual partitioning link.
- *equiv.* It is an equivalence relation used to express synonymy or medical equivalence. It allows the creation of "grapes" attributes that are similar.
- *isNot* is a negation that can be used to ease the matching of similar concept, but that would have Fig-

ure 2. Overall architecture *(dashed = not yet implemented)* been expressed using negation in the Attributes.

The same relations are used to built a network of concepts, avoiding a deep model whenever possible [7]. The pragmatic goal is to reuse data in the CPR and to be able to help decision-support.

Storage

The storage is made in the two main repositories (in darker in figure 2) of the CPR, the patient-centric database, DoMed (documents médicaux) and the data-centric database, BDOC (Banque de Données Opérationnelle Clinique). All accesses to the data are made through the middleware MUSIC (Middleware Unifié du Système d'Information Clinique) in lighter in figure 2 using HTTP / XML messages. Both databases do share a common set of components for management, accesses, auditing, etc.

Beside questionnaires and documents, BDOC does hold all structured data on clinical activity about patients, such as order entry and laboratory. All these data have the common characteristics to have very defined acquisition pathways. In the contrary, the DoMed database is much more heterogeneous and can hold pretty much any kind of data being file oriented. So, the scanning system that will be used in the near future to scan all documents still on paper will send all its outputs, scanning and texts from optical character recognition (OCR) to DoMed. All *data sets* that have been validated in the BDOC database will also be saved as consolidated documents, in general using PDF format, in the DoMed database. So is it also for the laboratory results too. This allows being able to display the document of the CPR very fast to the users, without having to do complex queries in order to consolidate data. However, data are available in a data-centric database, therefore permitting analysis and decision-support.

Conclusion

The patient record is made of very heterogeneous documents originating from numerous sources. Most of these documents are made both of structured data and narratives. Some of them, such as admission notes from general practitioners, will not be available largely in electronic form for a long time and must be scanned. We have developed a way to have a unique repository for all these documents and data in implementing a dual storage architecture that is tightly integrated. This system is part of our n-tiers component-based architecture and can be used with XML formatted messages and the HTTP protocol. Up to now, more than 6'000'000 documents are stored in the DoMed database and more than 100'000 data sets in the BDOC database.

The separation of the semantic model and the pragmatic attributes dictionary layer has allowed a fast growth of the number of data contexts, although the building of the semantic model remains an important problem and challenge that will have to be solved.

References

1 Tang PC, LaRosa MP, Gorden SM. Use of computer-based records, completeness of documentation, and appropriateness of documented clinical decisions. J Am Med Inform Assoc 1999;6(3):245-51.

2 Sideli RV, Johnson SB, Clayton PD. Full-text document storage and retrieval in a clinical information system. Comput Methods Programs Biomed 1999;60(3):153-81.

3 Lovis C, Baud RH, Planche P. Power of expression in the electronic patient record: structured data or narrative text? [In Process Citation]. Int J Med Inf 2000;58-59:101-10.

4 Kirby J, Rector AL. The PEN&PAD data entry system: from prototype to practical system. Proc AMIA Annu Fall Symp 1996:709-13.

5 Stein HD, Nadkarni P, Erdos J, Miller PL. Exploring the degree of concordance of coded and textual data in answering clinical queries from a clinical data repository. J Am Med Inform Assoc 2000;7:42-54.

6 Tange HJ, Hasman A, de Vries Robbe PF, Schouten HC. Medical narratives in electronic

Acknowledgment

Part of the work has been funded by the Swiss National Science Foundation 632 -066041