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

Paragraph-oriented Structure for Narratives in Medical Documentation

**Christian Lovis, Robert H Baud, Claude Revillard,
Laura Pult, François Borst, Antoine Geissbuhler**

*University Hospital of Geneva,
Division of Medical Informatics, Geneva, Switzerland*

Abstract

The authors present a 6 years experiment using a document-centered electronic patient record, based on a central document repository. The document management system is paragraph oriented and all documents are built automatically before editing using predefined ordered sets of paragraphs. Paragraphs can be preloaded with templates, text or images. Once edited, signed and printed, documents are again decomposed in paragraphs and permanently stored. This system, though the compositional aspect of paragraphs is limited and their semantic content wide, offers numerous advantages. The typology is easy to build and to maintain, it has been implemented widely in our hospitals without need for any natural language processing techniques and is used daily within commercially available text editors. The actual state of the system is discussed, emphasizing the structure of the documents, the various attributes and properties that have been needed in order to meet user's needs.

Keywords:

Documentation, classification, knowledge representation, Electronic Patient Record

Introduction

Narrative text reports represent a significant and important source of clinical data. Tightly controlled and structured data entry can be a major burden for health care providers with high costs in time [1-3]. One of the major challenges in designing the electronic patient record (EPR) is to meet the needs for detailed documentation whilst keeping the burden for direct care providers in an acceptable range. Whereas narratives are still the most natural and used way to express medical information, they still suffer from being out of reach from robust natural language analysis [4]. Beyond the ongoing scientific discussion of knowledge representation and semantic analysis of narrative text reports, there is a strong need for structuring and characterizing narratives. The minimal levels of structure that can be implemented are the typology of documents and, within each document, the structure of paragraphs. To be able to describe an explicit structure of documents has many advantages. It allows the production of consolidated views of the documents for care providers, like concatenation of all history paragraphs chronologically. It is also the first step towards deeper analysis of the text, as lexical and rule-based knowledge needed to analyze various categories of medical texts can be very different, as shown recently by Barrows [5]. Finally, a good knowledge of document structure is of great help for automatic anonymization [6].

Our experiment is based on the Diogene 2 architecture [7, 8] that allows us to have a centralized repository for various kinds of clinical narratives, including complex discharge letters. More than 700 users in our hospital use this system in 40 medical services. All inpatient clinics are using the system for a coverage exceeding 80% of official reports. This includes radiology reports, pathology, surgical procedures and discharge letters among others, but not progress notes, which are handled separately. Most medical outpatient clinics are in the process of using the system for patient summaries and discharge reports. At the time of writing this paper, over 2 millions documents were available online. Among them, approximately 860.000 reports and discharge letters. Almost 6.000.000 paragraphs have been generated during this process, which represents approximately 7 paragraphs per document. Documents are stored in two formats. On one side, they are stored as sets of paragraphs linked in a relational SQL database and on the other side they are stored as read-only viewable documents in a document management system. This system is the cornerstone of hospital-wide dissemination of medical documents for healthcare providers within the EPR and is based on a three-layers XML-compliant architecture. The database can be queried either using a document-centered approach or a paragraph-centered approach. For example, it is possible to have all reports of a patient or, within all documents of a given patient, all history, physical examinations, diagnosis paragraphs, etc.

Since its first implementation in our hospital in 1994, this document management system has had a constant increase in both coverage of various medical services and document-type handled in each service. Figure 1 shows the evolution of the amount of radiology reports edited in then system since 1993, with a peak last year at 67.509 reports.

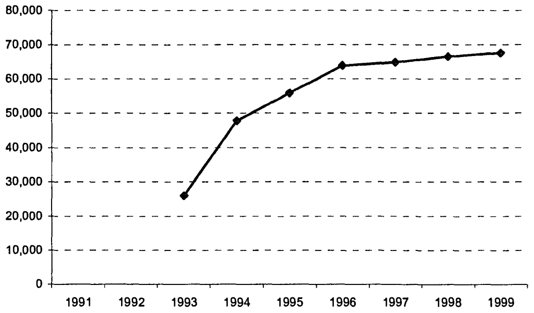


Figure 1 - Evolution of radiology reports between 1993 and 1999

Our approach differs from several previous publications as it introduces a strong *a priori* paragraph-based structure without any semantic or conceptual indexation [9]. In addition, most work relating to document structure has focused on word and concepts semantic, eventually using XML or SGML [10-15]. Its major originality is the development of a unified structured representation of medical narratives documents with a granularity that has been formulated up to the level of paragraph. We describe the structure to paragraphs using a relational database and are able to generate the final document structure using several formalisms, such as rich text format (RTF) in order to meet user's requirements for layout and quality printout. The system permits also to generate structured ASCII outputs for natural language analysis. When compared to the HL7 Clinical Document Architecture (CDA) semantics (previously known as the Patient Record Architecture, PRA) [16, 17], our architecture is very light as it only designates documents and paragraphs. They more or less represent the same type of content as the *containers* defined in the HL7 CDA.

Structure of documents

Medical documentation is characterized by a strong structure. This structure can be found at several levels, from medical domain sublanguage up to the EPR organization. Surprisingly, whereas there are numerous papers on the EPR organization, the medical NLP techniques and medical semantic representation, very few papers can be found about the overall structure of medical narratives themselves or the structure and typology of paragraphs used to build these documents [18]. It appears however that medical documents have a strong, though not yet explicit, paragraph oriented structure. An interesting use of the natural structure of medical documentation in the United States is the HCFA Evaluation/Management services system, at least for the seven main descriptors. This system is still a strong incentive for structured data acquisition in the EPR [19-21]

The paragraph-oriented architecture we present in this paper is used to create new documents and to store them. After editing, all documents are stored in the two repositories, database and document-management system. It must be emphasized that when users are editing these documents, the paragraph-oriented structure is only visible through layout within a rich text format compliant text editor.

There is no direct repository for empty documents, as they do not exist *per se* in the system. Any empty document must be created dynamically with an ordered set of existing paragraphs.

A central repository is used to enumerate all possible paragraphs. This repository contains a first table that describes the type, the title and several properties of the existing paragraphs. The body, or content, is empty. A second table describes all documents possible by enumeration of the paragraphs needed to build them. This description contains several fields, like title of the document, associated medical service(s), purpose, etc. So, each document can be dynamically created with a set of empty paragraphs. The steps needed to create a new document are summarized in Figure 2. Finally, templates can be applied to any document, with textual or structured preloaded information, in order to normalize and facilitate the edition. Typically, thoracic standard radiology reports will be preloaded with the description of a normal radiography. An example of ordered set of paragraphs is shown in Table 1. This is the set used to create an empty surgery transfer report.

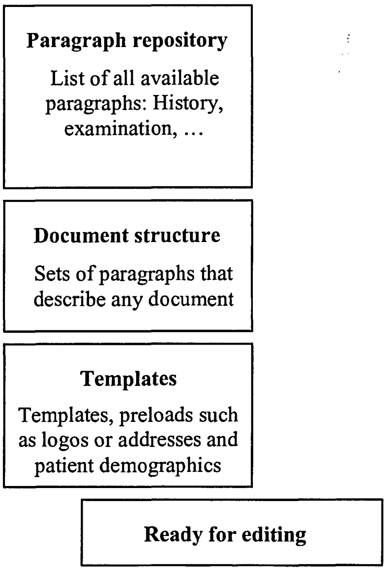
Table 1 - Example of set of paragraphs for a surgery transfer report

Header
Logo
Addresses
    Diagnostics
    Procedures
    Summary
    Points to follow
    Reason for transfer
    Treatment
Signatures, copies
Footer

Once the document is formed with the specific succession of paragraphs, a second step is involved that will apply the specific layout, logos or possible preloads in each paragraphs. There are three kinds of possible content that can be preloaded:

1. The general layout of the document, including font and paragraph formatting,
2. Static preloads, like the logo of the hospital, or the templates for several kinds of reports, such as ultrasounds.
3. Dynamic preloads, such as patient ID or demographics, date and duration of the stay, addresses of physicians.

At this point, the document is ready for editing and is downloaded on the target computer where the user can do the final edition. When the users edit the document, paragraph tags cannot be deleted, however the textual content of each paragraph, including eventually its title, can be freely modified. If a paragraph is completely deleted, it will not more appear visually or when the document is printed but will be kept in the database as an empty paragraph.



Paragraph repository

List of all available paragraphs: History, examination, ...

Document structure

Sets of paragraphs that describe any document

Templates

Templates, preloads such as logos or addresses and patient demographics

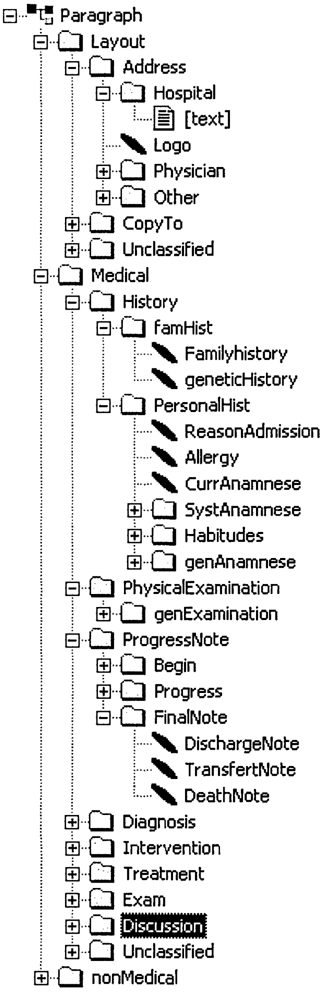
Ready for editing

Figure 2 - Steps for document production

Once edited, the document can be saved and will be published after valid electronic signature. The document is saved centrally in two different databases. On one side, it will be decomposed according to its paragraph structure, and each paragraph will be stored in a text field of a relational database. On the other side, the whole document with its complete layout will be stored in a document management database, as a viewable file, and made available within the EPR. The storage of full viewable documents in a separate document-oriented database has been made mainly to ensure a fast response time for display (< 1 sec) within the EPR.

Typology

In the present system, as used daily, the list of available paragraphs is still a flat list, however a typology is being built. There has been no clear need for a typology up to now and there is still debate as to what kind of typology could be useful. It appears that a medically pertinent organization has the most support. Such a typology consists of regrouping similar paragraphs under common a header. For example, *Social history, family history, history of current disease*, etc. would be regrouped under *history*.



```
Paragraph
  Layout
    Address
      Hospital
        [text]
      Logo
      Physician
      Other
    CopyTo
    Unclassified
  Medical
    History
      famHist
        Familyhistory
        geneticHistory
      PersonalHist
        ReasonAdmission
        Allergy
        CurrAnamnese
        SystAnamnese
        Habitudes
        genAnamnese
    PhysicalExamination
      genExamination
    ProgressNote
      Begin
      Progress
      FinalNote
        DischargeNote
        TransfertNote
        DeathNote
    Diagnosis
    Intervention
    Treatment
    Exam
    Discussion
    Unclassified
  nonMedical
```

Figure 3 - Example of typology for paragraphs

The level of granularity is determined by the smallest coverage of a single paragraph. For instance, if *Allergy* is considered as a possible paragraph, this will define the granularity of the representation. In the typology on which we are actually working, such considerations led to a four-level deep representation for medical history, as presented in Figure 3. At each level, there is a clear need for grouping paragraphs that have a similar content. A typical situation were there are similar paragraph in our system is illustrated when paragraphs are both available in French and in English. Our system has actually 477 different paragraphs. A first grouping with one level that takes into account paragraphs with similar content decreases this amount to 110. Each paragraph can, as it is the case for CDA containers, have specific properties like confidentiality status, encoding, link to other paragraphs. However, this is not yet used in our system where properties are only linked to documents. The only properties used are the possible templates for preloading and a flag indicating if the paragraph can be left empty.

Besides this medical grouping of paragraphs, others needs made a multiaxial typology useful for documents. An axis is needed for the organizational aspects of the workflow; another axis represents a medical sound structure and a third axis used for document types and categories. In the present system, documents are organized only in a bi-axial typology. The first axis is devoted to the medical organization of our hospital and is used to control the workflow of documents, compute the rights and accesses for health care providers in charge of edition, typing and signing documents. It is merely an institutional description. The second axis is devoted to the purpose of the EPR and dictates the way documents can be organized in a visual manner in the EPR; it is more a categorical structure and accounts the medical description. In this second axis, documents can be described as notes, reports, results or letters for example. Queries can be used to identify all radiology reports or discharge letters from any division of medical or surgical services. Such a descriptive typology is extremely useful in order to be able to categorize documents in a pertinent manner for users.

Properties of a document

There is a set of properties shared by all documents. These properties can be divided into four axes, according to their functionalities:

1. *Identification.* Typical mandatory elements are used to identify documents, like author, date and time, patient ID, medical service of emission, etc. In addition, we have information about versioning, because signed documents cannot be modified in our system. Documents modified after signature will be considered as new documents in a history of documents. Finally, each document receives a unique sequential identifier.
2. *Category.* The category identifies a document within the lifecycle of the information to which the document relates. The category is an ordinal within a list that includes *provisory, preliminary, intermediate, definitive, supplementary* and *additional.* Not all documents types have an entry in each category. Most documents types have only one (usually *definitive*) or two categories. Some documents types, like pathology reports, have all categories.
3. *Status.* The status defines whether a document is a draft or not. A signed document is not more a draft and cannot be modified in the future. As long as it remains a draft, the document will not be published outside the editing group, which includes all authors and typists. The status of a document is not dependent from the category; all documents have a status, whatever their category.
4. *Type.* The type is an entry in our hierarchy of documents.

Users needs and requirements

The major problem that we faced was the need of users for complex page layout and formatting capabilities. These needs include font and paragraph formatting, running headers and footers, tables, images and footnotes, among others. An important requirement was to use a commercially available text editor, as it was difficult and too expensive to ensure the teaching of all possible users. We had to use the office suite already installed on PC's. In addition, preloaded templates were required in order to decrease the burden of narratives acquisition for many reports. However, as said in introduction, the respect of these few points led to a high user's acceptance of the system.

The next steps to work on include finalizing the typology of paragraphs, eventually on the models that might be available as standards in the future, like HL7 or CEN. We are also in the process of defining paragraph-level attributes, such as confidentiality status.

Conclusion

We present a versatile way to structure and organize medical narrative documentation based on a document typology and formal paragraph segmentation. Though this system does not allow a deep modeling of medical knowledge and does not include any conceptual representation or structured data entry, it allows taking into account all medical narratives produced in our hospital and clinics and is an important step towards deeper natural language processing. It is a critical project within the EPR as it leads to normalization of narrative documentation and formalization of the many documents produced by clinical services within a unique architecture. In addition, it has been possible to implement the system using a commercially available text editor with all layouts and formatting facilities expected and needed by users, which was a critical point for user's acceptance.

Acknowledgments

The authors wish to thank the Clinical Support Group, Irène Ponsolle and Jalil Ben Hamza, who devoted untold hours to configure the Unidoc system in support of our users.

References

[1] Bates DW, Boyle DL, Teich JM. Impact of computerized physician order entry on physician time. *Proc Annu Symp Comput Appl Med Care* 1994:996.

[2] Tierney WM, Miller ME, Overhage JM, McDonald CJ. Physician inpatient order writing on microcomputer workstations. Effects on resource utilization. *Jama* 1993;269(3):379-83.

[3] Payne TH. The transition to automated practitioner order entry in a teaching hospital: the VA Puget

Sound experience [In Process Citation]. *Proc AMIA Symp* 1999(1-2):589-93.

[4] 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.

[5] Barrows Jr RC, Busuioc M, Friedman C. Limited parsing of notational text visit notes: Ad-hoc vs. NLP approaches [In Process Citation]. *Proc AMIA Symp* 2000(20 Suppl):51-5.

[6] Ruch P, Baud RH, Rassinoux A, Bouillon P, Robert G. Medical document anonymization with a semantic lexicon [In Process Citation]. *Proc AMIA Symp* 2000(20 Suppl):729-33.

[7] Scherrer JR, Revillard C, Borst F, Berthoud M, Lovis C. Medical office automation integrated into the distributed architecture of a hospital information system. *Methods Inf Med* 1994;33(2):174-9.

[8] Borst F, Appel R, Baud R, Ligier Y, Scherrer JR. Happy birthday DIOGENE: a hospital information system born 20 years ago. *Int J Med Inf* 1999;54(3):157-67.

[9] Charlet J, Bachimont B, Brunie V, el Kassar S, Zweigenbaum P, Boisvieux JF. Hospitexte: towards a document-based hypertextual electronic medical record. *Proc AMIA Symp* 1998:713-7.

[10] Lincoln TL, Essin DJ. A document processing architecture for electronic medical records. *Medinfo* 1995;8(Pt 1):227-30.

[11] Kimura M, Ohe K, Yoshihara H, Ando Y, Kawamata F, Tsuchiya F, Furukawa H, Horiguchi S, Sakusabe T, Tani S, Akiyama M. MERIT-9: a patient information exchange guideline using MML, HL7 and DICOM. *Int J Med Inf* 1998;51(1):59-68.

[12] Myers DL, Culp KS, Miller RS. Use of a Web-based process model to implement security and data protection as an integral component of clinical information management. *Proc AMIA Symp* 1999:897-900.

[13] Sokolowski R, Dudeck J. XML and its impact on content and structure in electronic health care documents. *Proc AMIA Symp* 1999:147-51.

[14] Kuikka E, Eerola A, Porrasmaa J, Miettinen A, Komulainen J. Design of the SGML-based electronic patient record system with the use of object-oriented analysis methods. *Stud Health Technol Inform* 1999;68:838-41.

[15] Kahn CE, Jr. Self-documenting structured reports using open information standards. *Medinfo* 1998;9(Pt 1):403-7.

[16] Dolin RH, Alschuler L, Boyer S, Beebe C. An update on HL7's XML-based document representation standards [In Process Citation]. *Proc AMIA Symp* 2000(20 Suppl):190-4.

[17] Dolin RH, Alschuler L, Behlen F, Biron PV, Boyer S, Essin D, Harding L, Lincoln T, Mattison JE, Rishel W, Sokolowski R, Spinosa J, Williams JP. HL7 document patient record architecture: an XML document architecture based on a shared information model. *Proc AMIA Symp* 1999:52-6.

[18] Laforest F, Flory A. The Electronic Medical Record : Using Documents For Information Capture. In: Hasman A, Blobel B, Dudeck J, Engelbrecht R, Gell G, Prokosh HU, editors. *Proc MIE 2000*; 2000: IOS Press; 2000. p. 617-621.

[19] AMA/HCFA. Documentation Guidelines for Evaluation and Management Services. 1997.

[20] McCormack J. Electronic records: key to HCFA compliance? *Health Data Manag* 1998;6(6):40-2, 44.

[21] Taragin MI, Lauer M, Savir M, Sivan E, Siesel D, Aufgang B. HCFA documentation guidelines and the need for discrete data: a golden opportunity for applied health informatics. *Proc AMIA Symp* 1998:653-6.

Address for correspondence:

Christian Lovis MD MPH
Division of Medical Informatics
Integrated Electronic Patient Record Group
University Hospital Of Geneva
21. Micheli-du-Crest
CH-1211 Geneva 4
Phone +41 22 372-6008
Fax +41 22 372-6198
email christian.lovis@dim.hcuge.ch