Document Semantic Model :
an experiment with patient medical records.

Jean-Marie Pinon, Sylvie Calabretto, Line Poullet
Laboratoire d'Ingénierie des Systèmes d'Information - INSA de Lyon.
20, avenue Albert Einstein
F-69 621 Villeurbanne Cedex, France
e-mail: pinon@ifhpserv.insa_lyon.fr

ABSTRACT

Patient medical records contain a large amount of information distributed in different kinds of documents: diagnosis, prescription, symptom observations or radiology analysis, etc. Document heterogeneity makes specific information retrieval difficult for medical staff. This paper shows how a semantic model of documents assists in handling information stored in these documents. It allows the definition of a generic semantic structure of a medical record: this structure expresses the implicit content of each document element by specifying what kind of information is required. Moreover, it permits a display of relevant information for a specific reader.

1. Introduction

According to international standards, such as Open Document Architecture [ISO 8613] or Standard Generalised Mark-up Language (SGML) [ISO 8879], and according to the usual document processing software products, a document is considered to have conceptually two structures (Figure 1) :

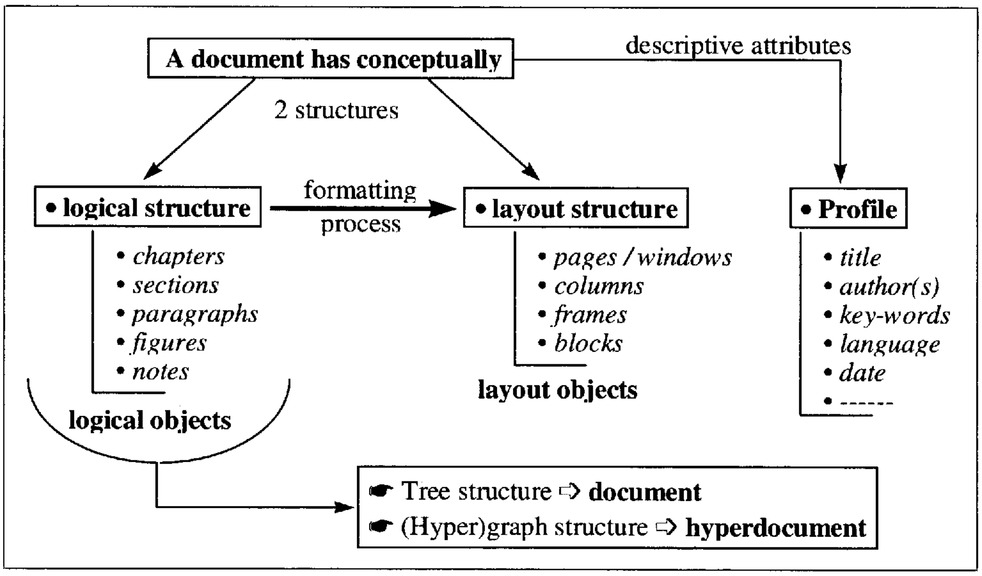


Figure 1 . Usual document architecture

- the *logical structure* representing the overall organisation of information. It is composed of logical objects such as chapters, sections, paragraphs, figures, notes, and so on;

- the *layout structure* representing the presentation of this document on paper sheets or on a screen. It is composed of layout objects such as pages, columns, frames, blocks, and so on. It is automatically generated from the logical structure and content portions thanks to the formatting process.

If the logical structure is a tree, we have a *document*, but if it is a graph or a hypergraph, we have a *hyperdocument.* Moreover, descriptive attributes are often attached to a document such as title, author names, language, date, and keywords. This attribute set is usually named *document profile.*

Our proposal consists of defining a third structure, *the semantic structure*, which is linked to the logical structure and which expresses the meaning of each logical element in a formal way. This semantic structure can be efficiently defined using SGML syntax with a few technical extensions. This semantic structuring provides a document description which is more precise and more powerful than a simple description by keywords. So, thanks to semantic structuring, it is possible to implement a new generation of more powerful tools such as documentary information retrieval tools or documentary information processing tools.

The first part of this paper presents an example to illustrate our semantic model, the medical patient record, which is a relevant application for handling semantic structured documents. The second part gives an overview of the model. The third part shows how semantic structuring of documents can be efficiently defined using SGML syntax. Using this document structuring standard, two levels of description may be defined: generic semantic structure (versus. Document Type Definition : *DTD*) and specific semantic structure (versus. *SGML instance*) in order to define an 'abstract interface' of information stored in documents.

2. Presentation of our example : a Patient Medical Record

In order to illustrate our semantic model, we have chosen an example of a patient medical record (Figure 2). The patient is Mr Smith who is hospitalized in the Fleming Hospital because he has a major headache (Mr Smith and the Fleming Hospital are purely imaginary, we apologize to the readers named 'Smith' and to the hospitals named 'Fleming' for the use of their name).

The *layout structure* of this document is made up of *composite layout objects* (the five pages, the 'age-address-phone ' frame in the top of the page 2, etc) and *elementary layout objects* (the 'patient-name' block in the bottom of the page 1, etc).

The *logical structure* is composed of the following *composite logical objects (CLO)* :

- the 'patient description' CLO is composed of several *elementary logical objects (ELO)* such as 'patient-name' ELO , 'age' ELO, 'address' ELO, 'phone number' ELO, and so on) ;
- the 'antecedent-set' CLO is optional ;
- the 'observation-set' CLO stores the observations made in the emergency service when Mr Smith arrived ;
- the 'test-set' CLO consists of several 'test' CLOs (for example, the third test is the Mr. Smith's electroencephalogram) ;
- the diagnosis CLO whose content is 'meningitis';
- the 'prescription-set' CLO is made up of several 'prescription' ELOs (for example, the second prescription is 'penicillin, six tablets per day';

- and so on.

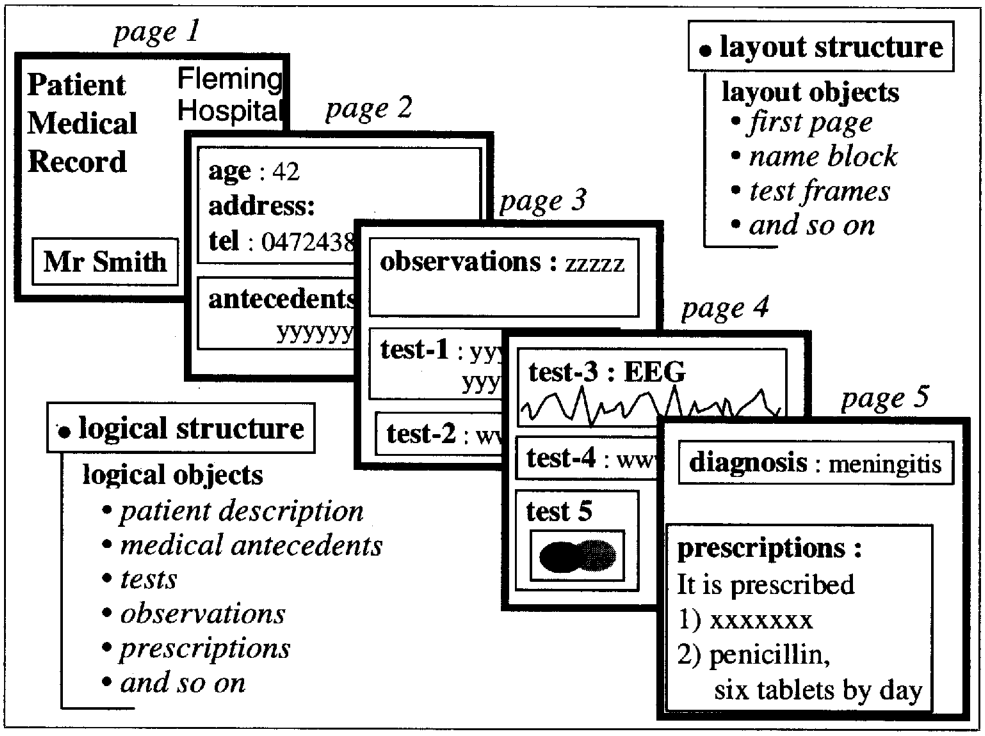


Figure 2. Example of document : Patient Medical Report

3. Model description

3.*1*. *Semantic model*

The semantic model relies on 'meaning representation' of information units (i.e. the logical units). This meaning representation is distributed in the overall architecture model : the model binds together the generic semantic structure, the generic logical structure of a document class (i.e. a SGML DTD) and a domain model (see Figure 3 and Figure 4).

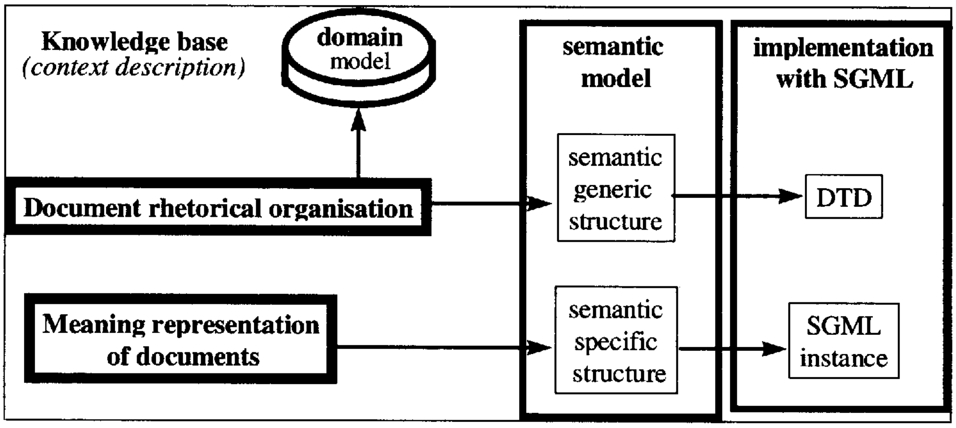

Knowledge base
(context description)
domain model
Document rhetorical organisation
Meaning representation of documents
semantic model
semantic generic structure
semantic specific structure
implementation with SGML
DTD
SGML instance

Figure 3. Implementation of our semantic model

The semantic model contains two levels of description :

- document rhetorical organisation (organisation of the discourse in the document class such as assertions, descriptions, examples, etc*).*
- meaning representation of information units (the Conceptual Graph formalism is used to represent semantics of document elements) ;

In our example, the domain model contains a medical ontology of concepts and relations between concepts, used for guiding the meaning representation of information units. The generic semantic structure defines the generic organisation of document content for a specific class of patient medical records. Each element of this structure defines the a priori semantic content of the associated logical element.

3.2. *Semantic Structure*

The semantic structure (Figure 4) includes a hierarchy of semantic objects. A semantic object may be composite or elementary. A composite semantic object is a hierarchical structure of semantic objects. An elementary semantic object contains a formal description of the corresponding logical object. This description is based on natural language representation formalisms such as description logic [Brachman 85] or the Sowa conceptual graph [Sowa 84].

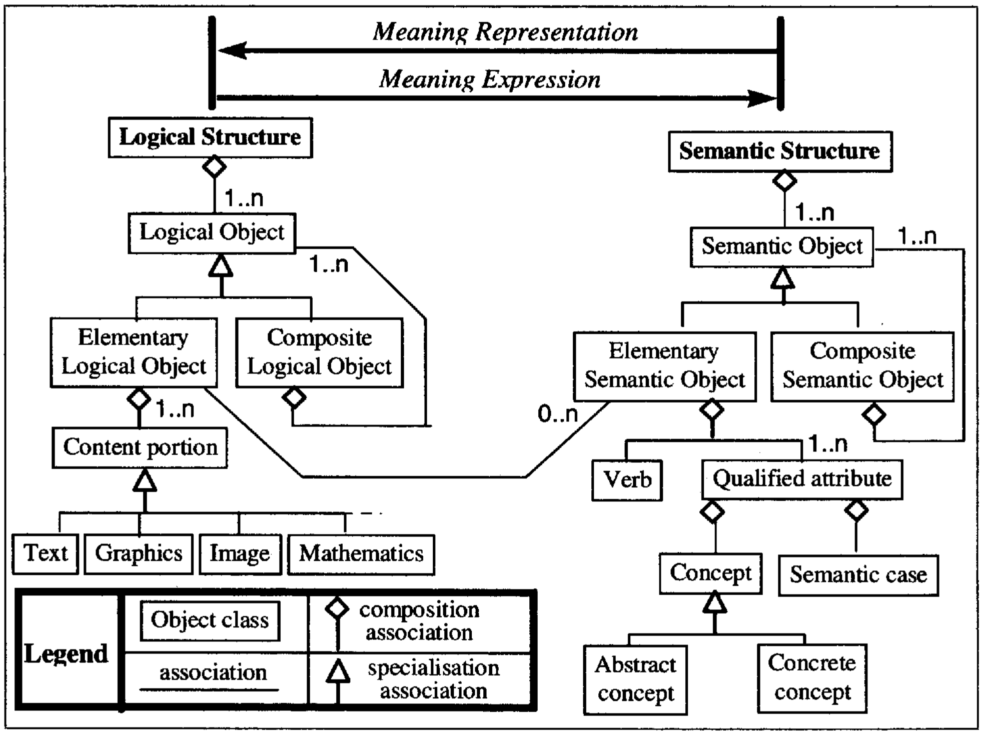


Figure 4. Logical structure and semantic structure of a document

An elementary semantic object is composed of a verb and a list of qualified attributes. *'To be prescribe', 'to suffer'* are examples of verb. A qualified attribute is composed of a concept and a semantic case. The notion of semantic case is similar to the notion of semantic role. *'Matter', 'state', 'recipient'* are examples of semantic case. A concept may be abstract or concrete. *'Drug', 'dosage', 'patient'* are examples of concept. In *'to take a drug'*, 'drug' is an abstract concept. In *'Mr Smith takes penicillin'*, penicillin is a concrete concept. The pairs *'drug-matter'* or *'penicillin-matter'* or *'patient-recipient'* or *'Mr Smith-recipient'* are examples of qualified attribute.

The main structure of a document is its logical structure. Elementary Logical Objects are connected to content portions. A content portion may be a text, an image, a graphic, a mathematics content, and so on. A semantic structure is thus bound to the logical structure by two types of process (these kinds of process may be used for elaborating conceptual documents [Nanard 1988]) :

- Meaning expression enables logical units to be represented by semantic units. Currently, this is a manual process.
- Meaning representation refers to the fact that semantic units represent logical units.

3.3. *Generic and Specific Semantic Structures*

Each specific document, such as *'the medical report on Mr Smith'*, may have a Layout Specific Structure, a Logical Specific Structure and a Semantic Specific Structure. These specific structures must be true to the corresponding Generic Structures. A document class, such as *'the medical report class'*, is usually defined by three generic structures (logical, layout, semantic).

Figure 5 shows a part of the specific semantic structure of the medical report on Mr Smith (Figure 2). The elementary logical specific object *'penicillin six tablets per day'* is indexed by this elementary semantic specific object . Its verb is 'to be prescribed'. This verb is completed by three qualified attributes : 1- the matter is penicillin, 2- the state is six tablets per day and 3- the recipient is Mr Smith.

When instantiating the generic semantic structure (for indexing or creating the document), the specific semantic elements are instantiated: Mr. Smith is the patient, he had meningitis, ...).

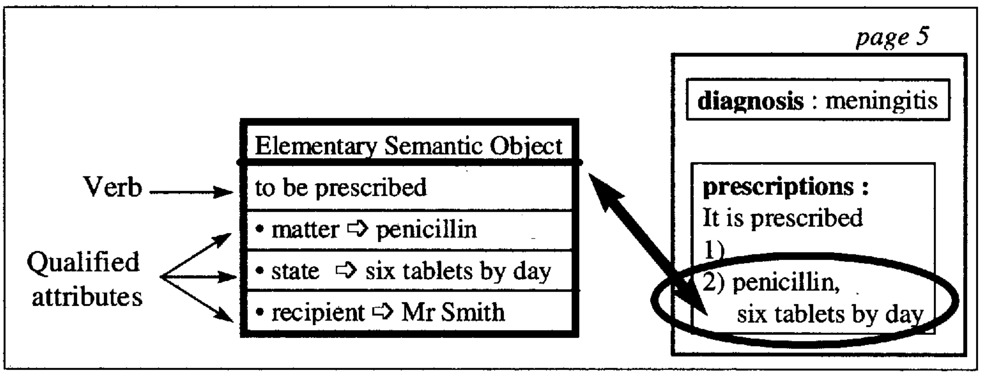


Figure 5 . Example of a specific elementary semantic object

4. Semantic structuring with SGML

SGML offers two levels of document representation:

1. A logical document model, called "Document Type Definition " (DTD), describes the logical organisation of a class of documents, such as the patient medical record class.
2. A SGML instance, true to a DTD, describes the logical organisation of an existing document, such as the Mr Smith medical record.

In the same way, we can define a generic semantic structure as a semantic DTD in order to model the specific semantic structure of documents belonging to the corresponding class. Moreover, we have defined a *conceptual semantic structure* which is a kind of *meta-DTD.* It provides the semantic syntax used for defining a semantic DTD.

We saw (section 3.1 and figure 3) that our model has also two levels of description :

- the document rhetorical organisation which is carried out by a semantic generic structure and a knowledge base. This semantic generic structure is defined as a SGML DTD ;
- the meaning representation of a specific document which is carried out by the semantic specific structure and is implemented by a SGML instance.

Let us note that a few SGML extensions are necessary. In particular, the SGML environment does not provide an inheritance mechanism. So we simulate it in this way (Figure 6):

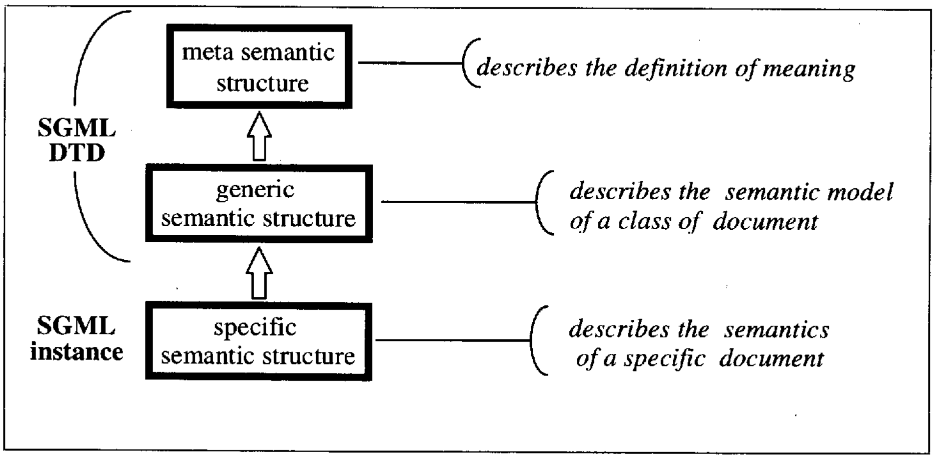

SGML
DTD
SGML
instance
meta semantic structure
generic semantic structure
specific semantic structure
describes the definition of meaning
describes the semantic model of a class of document
describes the semantics of a specific document

Figure 6. Implementation with SGML

1. we have defined a meta-semantic structure the aim of which is to implement the definition of meaning. This meta-semantic structure is a SGML DTD (Figure 7) ;
2. for each document class, it is necessary to define a semantic model, in other words a generic semantic structure which is also a DTD true to the meta semantic structure (Figure 8) ;
3. Finally, we describe the semantics (meaning the representation) of a specific document as an instance of the previous DTD (figure 9).

4.*1.* *Meta-semantic structure*

This abstract semantic structure implements the definition of meaning noted in section3. This abstract structure (Figure 7) shows how the meaning definition can be implemented in a SGML environment. This grammar defines a semantic structure (SemStructure) composed of semantic elements (SemElement), i.e. at least (+ occurrence indicator) one semantic element. Semantic elements are an Expression (elementary semantic elements), i.e. semantic representation of logical element(s) or are composed of other semantic elements (composite semantic elements).

```
<!DOCTYPE  SemStructure   [
<!ENTITY   Case                              ' SemanticCase '>

<!ELEMENT  SemStructure           - -   (SemElement)+ >
<!ELEMENT  SemElement             - 0   (Expression | SemElement)+>

<!ELEMENT  Expression             - 0   (Verb, QualifiedAttribute+)>
     <!ATTLIST  Expression        id        ID       #IMPLIED>
     <!ATTLIST  Expression        logEnt    IDREF    #IMPLIED>
     <!ATTLIST  Expression        Type CDATA  #IMPLIED      >

<!ELEMENT  Verb                   - 0   #PCDATA>
<!ELEMENT  QualifiedAttribute     - 0   (Concept  )>
     <!ATTLIST  QualifiedAttribute      case ENTITY  #FIXED>

<!ELEMENT  Concept                - 0   (ConcreteConcept | Expression)>
<!ELEMENT  ConcreteConcept        - 0   #PCDATA>                    ]>
```

Legend :	Connectors	Occurrence Numbers
	, : sequence *(ordered)*	+ : from 1 to many
	& : aggregate *(not ordered)*	* : from 0 to many
	| : or	? : 0 or 1 *(optional)*

Figure 7. The meta-semantic structure.

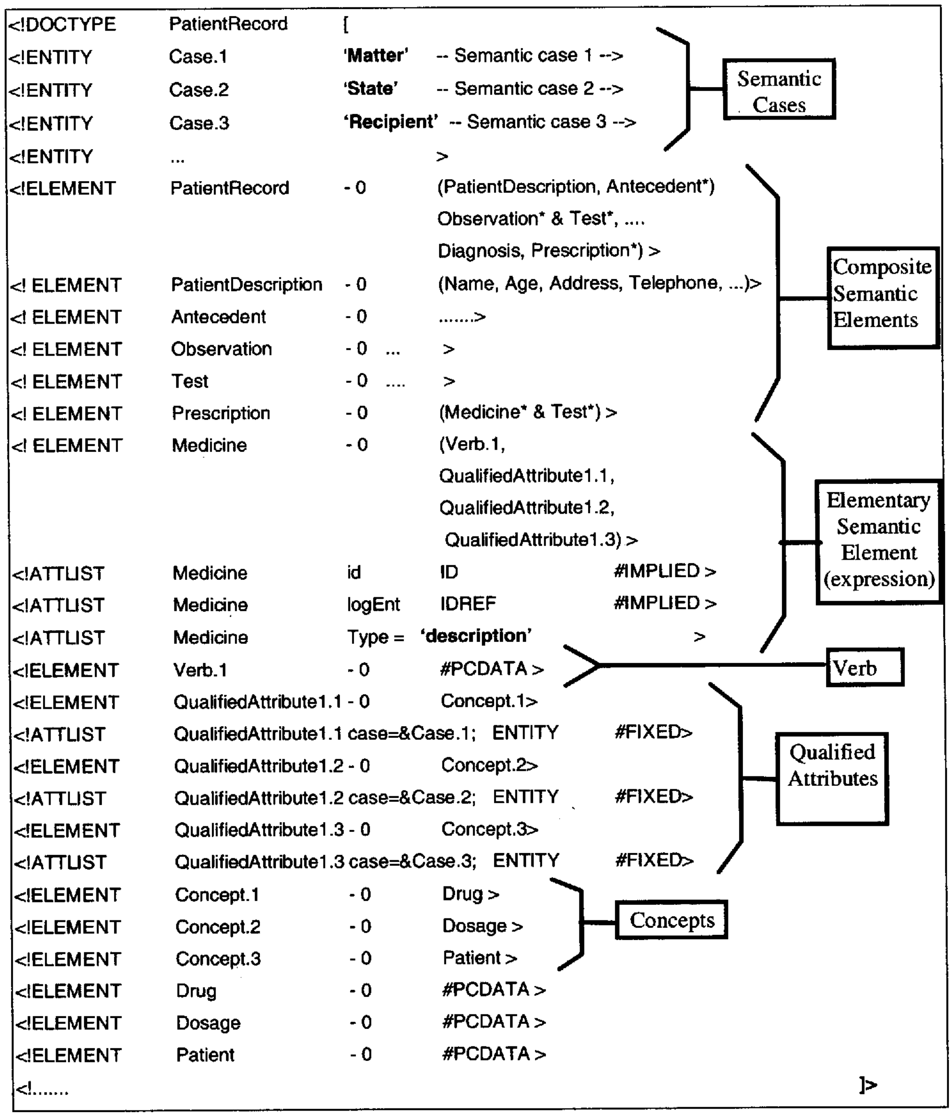
An expression is made up of a verb and a list of qualified attributes. Note that the SGML element 'Expression' has three attributes :

- - the *id* attribute enabling this element to be referred to, for example in a logical element ;
- - the *LogEnt* attribute enabling this element to refer to a logical element (the converse link);
- - the *Type* attribute enabling this element to have a rhetorical type .

The links between a logical element and the semantic expression representing it, are described by the ID/IDREF mechanism, allowing reference to a SGML element.

4.2. *Generic Semantic Structure*

By using this grammar, the generic semantic structure of the 'Patient Record' document class, can be described more precisely as follows (Figure 8). In this DTD, only the semantic element *Medicine* has been entirely described.



```
<!DOCTYPE    PatientRecord        [
<!ENTITY     Case.1               'Matter'    -- Semantic case 1 -->
<!ENTITY     Case.2               'State'     -- Semantic case 2 -->
<!ENTITY     Case.3               'Recipient' -- Semantic case 3 -->
<!ENTITY     ...                              >
<!ELEMENT    PatientRecord        - 0         (PatientDescription, Antecedent*)
                                              Observation* & Test*, ....
                                              Diagnosis, Prescription*) >
<! ELEMENT   PatientDescription   - 0         (Name, Age, Address, Telephone, ...)>
<! ELEMENT   Antecedent           - 0         .......>
<! ELEMENT   Observation          - 0 ...     >
<! ELEMENT   Test                 - 0 ....    >
<! ELEMENT   Prescription         - 0         (Medicine* & Test*) >
<! ELEMENT   Medicine             - 0         (Verb.1,
                                              QualifiedAttribute1.1,
                                              QualifiedAttribute1.2,
                                              QualifiedAttribute1.3) >
<!ATTLIST    Medicine             id          ID                  #IMPLIED >
<!ATTLIST    Medicine             logEnt      IDREF               #IMPLIED >
<!ATTLIST    Medicine             Type =  'description'                  >
<!ELEMENT    Verb.1               - 0         #PCDATA >
<!ELEMENT    QualifiedAttribute1.1 - 0        Concept.1>
<!ATTLIST    QualifiedAttribute1.1 case=&Case.1;  ENTITY           #FIXED>
<!ELEMENT    QualifiedAttribute1.2 - 0        Concept.2>
<!ATTLIST    QualifiedAttribute1.2 case=&Case.2;  ENTITY           #FIXED>
<!ELEMENT    QualifiedAttribute1.3 - 0        Concept.3>
<!ATTLIST    QualifiedAttribute1.3 case=&Case.3;  ENTITY           #FIXED>
<!ELEMENT    Concept.1            - 0         Drug >
<!ELEMENT    Concept.2            - 0         Dosage >
<!ELEMENT    Concept.3            - 0         Patient >
<!ELEMENT    Drug                 - 0         #PCDATA >
<!ELEMENT    Dosage               - 0         #PCDATA >
<!ELEMENT    Patient              - 0         #PCDATA >
<!.......                                                          ]>
```

Semantic Cases
Composite Semantic Elements
Elementary Semantic Element (expression)
Verb
Qualified Attributes
Concepts

Figure 8: A semantic structure for patient record - Description of the semantic expression "Medicine" (see figure 7).

In this DTD, we can see that the generic semantic structure of a patient record is composed of several composite semantic elements : patient description element, zero, one or several antecedent element(s), zero, one or several observation element(s) and/or test element(s), one diagnosis element and zero, one or several prescription elements.

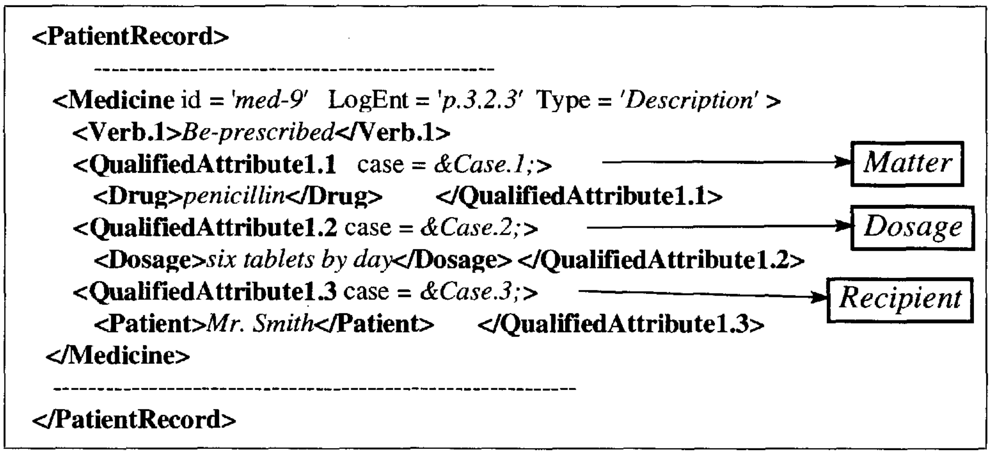
The medicine element is an elementary semantic element (we say also expression). It is made up of a "verb " element and three "Qualified Attribute " elements. In accordance with the meta DTD, Medicine has three attributes, 1- its identifier, 2- the identifier of its associated logical element and -3- its type whose *value* is 'description'. The three 'Qualified Attribute' elements are concepts ('drug', dosage' and 'patient') connected with semantic cases which are defined as attributes of these SGML elements. Their respective values (.'matter', 'state' and 'recipient') are defined thanks to SGML entities. So, matter qualifies drug, state qualifies dosage and recipient qualifies 'patient'.

4.3. *Specific semantic structure*

The specific semantic structure must be true to the generic semantic structure of its document class. Figure 9 details the elementary semantic element linked to the elementary logical object corresponding to the second prescription *'penicillin, six tablets per day'*.

Generic semantic elements permit the definition of an *a priori* meaning content by restricting concepts. This mechanism specifies what kind of concrete concepts and what kind of relations are concerned with the semantic elements. The *a priori* semantic content of a class of documents is then clarified. For example, in the generic semantic structure of the document class 'Patient Record' defined in section 4.2, the generic semantic element *Medicine* can be defined as relationship identified by the 'verb' between a 'drug' (which is a 'matter'), a dosage (which is a 'state', and a patient (who is a 'recipient ').

The specific semantic element *Medicine* points out that the logical element connected with it, expresses a relationship 'be-prescribed ' between a concrete drug (penicillin), a concrete dosage (six tablets per day) and a concrete patient (Mr Smith).



```
<PatientRecord>
  --------------------------------------------
  <Medicine id = 'med-9'  LogEnt = 'p.3.2.3'  Type = 'Description' >
    <Verb.1>Be-prescribed</Verb.1>
    <QualifiedAttribute1.1  case = &Case.1;>
      <Drug>penicillin</Drug>      </QualifiedAttribute1.1>
    <QualifiedAttribute1.2 case = &Case.2;>
      <Dosage>six tablets by day</Dosage> </QualifiedAttribute1.2>
    <QualifiedAttribute1.3 case = &Case.3;>
      <Patient>Mr. Smith</Patient>      </QualifiedAttribute1.3>
  </Medicine>
  --------------------------------------------------------
</PatientRecord>
```

Figure 9. Extract of a specific semantic structure PatientRecord. Description of the semantic expression Medicine

5. Conclusion

In this paper, we have proposed a formal model to assist users in the retrieval of information stored in documents. This model relies on a definition of semantic units, describing the meaning of information units. Documents have three bound structures: layout, logical and semantic structures. This model provides a way of defining a sort of implicit document, with an *a priori* clarified content.

Semantic elements can be used as a *semantic structured index* related to the documentary data:

- First, the main index level is the rhetorical type of semantic element. This index defines the kind of required discourse. Prescription of medicine as a description (medical prescription) appreciably differs from a prescription of medicine as an assertion (instruction for use).

- The second index is the verb which expresses the conceptual relations between the connected concepts: 'Be-prescribed' semantically differs from 'Be-forbidden' in the same type of semantic element.
- The third index is a the semantic case set.

This mechanism permits the indexing of parts of documents by following the 'meaning representation' links between semantic elements and logical elements. We believe that such a semantic structure provides tools for *information retrieval* in documents due to the formal representation of information. Statistical techniques [Lewis 96] have already been used for semantic retrieval mechanisms. Our approach allows for the extension of a definition of indexing and information retrieval, by taking into account formal semantic representations of text. It relies on the integration of two paradigms for representing the same information : structured documents on the one hand and a knowledge base on the other hand [Maret 96].

6. References

[Brachman 85] Brachman R. & Schmolze J. An Overview of the KL-ONE Knowledge Representation System. *Cognitive Science*, N° 9, 1985. p. 171-216.

[Dobson 95] Dobson, SA. & Burrill, V.A. Lightweight Databases. In : *Proceedings of the 3rd International World Wide Web Conference*, Darmstadt, April 1995. p. 282-288.

[Iso 86 13] International Standard. Information Processing - Text and Office Systems - Open Document Architecture (ODA) and Interchange Format (ODIF), 1988.

[Iso 88 79] International Standard. Information processing - Text and Office Systems - Standard Generalized Markup Language (SGML), 1986.

[Lewis 96] Lewis, D. & Spark Jones, K. Natural Language Processing for Information Retrieval. *Communications of the ACM*, 1996, Vol. 39, N° 1, p. 92-101.

[Maret 96] Maret, P., Poullet, L. & Pinon J-M Conceptual models for capitalizing information within an organisation in . *ISI (Ingénierie des Systèmes d'Information*), 1996, Vol 4, N° 4, pp 491 - 540

[Nanard 88] Nanard, J. et al. Conceptual documents : a mechanism for specifying active views in Hypertext. In: *Proceedings of ACM Conference on Document Processing Systems*, ACM Press, Santa Fe, 1988.

[Nanard 93] Nanard, J. & Nanard, M. Should Anchors be typed too ? An experiment with MacWeb. In : *Proceedings of the Fifth ACM Conference on Hypertext*, Seattle, November 1993. p. 51-62.

[Pinon 94] Pinon, J.-M., Richez, M.-A. & Flory, A. Support System for Cooperative Edition of Multimedia Documents. In: *Proceedings of OOIS'94*, London, 19-21 December 1994. London: Springer-Verlag, 1994. p. 416-421.

[Sowa 84] Sowa, J. *Conceptual Structures: information processing in mind and machine.* The System Programming Series. Addison Wesley publishing Company, 1984.

Dr Jean-Marie Pinon is a full professor in Computer Science at the Institut National des Sciences Appliquees of Lyon and Research Director at the Laboratory of Information Systems Engineering (LISI). Hé has supervised eight PhD dissertations and published two books and around 60 papers on various computing subjects, including document processing.

Dr Sylvie Calabretto is a lecturer in Computer Sciences at the Institut National des Sciences Appliquées of Lyon and Research Director at the Laboratory of Information Systems Engineering (LISI). She specialises in Semantic Indexing and Multimedia Databases.

Miss Line Poullet is a PhD student at the Laboratory of Information Systems Engineering (LISI). Her subject is the semantic representation of documents mixing both logical structure and semantic structure of content.