Modelling a Medieval Manuscript Database with HyTime

Sylvie Calabretto and Jean-Marie Pinon
LISI-INSA de LYON, 20, avenue Albert Einstein
F-69621 Villeurbanne Cedex
E-mail: cala@if.insa-lyon.fr

ABSTRACT

Our project enhances the accessibility of ancient manuscripts and provides new ways of working with them. More precisely, we aim to produce a software tool allowing historians, and more particularly codicologists and philologists, to read manuscripts, write annotations, and navigate between the words of the transcription and the matching piece of the image in the digitised picture of the manuscript. In this paper, we present the design of such a Hypermedia Workstation.It is based on the standardized hypermedia language HyTime (Hypermedia/Time-based Structured Language). We describe how HyTime can be used as a modelling language to describe works on manuscripts. We present relevant parts of the HyTime model and prove that the model thus obtained can also serve as a basis for implementation.

Introduction

The transfer of texts and images to digital media presents an interesting set of possibilities for those who, in various capacities, are concerned with the conservation of written documents, such as members of library staff, as well as for anyone carrying out studies in the field of philology. While in the recent past it was impossible to establish a link between these two spheres owing to the lack of suitable technology, there are real prospects for the emergence of concrete products today, thanks to the reliability and capacity of optical memories on the one hand, and to the use of object-oriented programming and hypermedia standards on the other.

The European Libraries project BAMBI[1] (Better Access to Manuscripts and Browsing of Images) fits into this context and is aimed at two categories of user: the first is represented by the general users of a library who wish to examine manuscript sources; the second category of users envisaged by BAMBI is made up of professional students of texts. In other words, the latter group consists of philologists or, to be more specific, critical editors of classical or medieval works that are hand-written on material substrates of various types (paper, papyrus, stone). It includes, therefore, students of ancient texts such as papyrologists, epigraphists, palaeographers, and codicologists: all those, in short, who are interested in studying, annotating, or transcribing the text contained in digitized and accessible manuscript documents.

[1] The different European partners of the BAMBI project are *ACTA S.p.a* (Computer Society of Florence), *CNR* (Consiglio Nazionale della Ricerche - Istituto di Linguistice Computazionale di Pisa), *BNR* (Biblioteca Nazionale Centrale V.E.II di Roma), *MPI* (Max Planck Institüt für Rechtsgeschichte (München)), *CPR* (Consorzio Pisa Ricerche), and *LISI*.

Thus, the BAMBI project aims to provide a "philological workstation " for European libraries. Its kernel is therefore a software tool which allows historians to enter information (or critical apparatus, or notes) on manuscripts linked to images and transcriptions of individual pages of those manuscripts, with especially close linkage of image and transcription. To achieve this aim, we have chosen to use the International Standard HyTime (Hypermedia/Time-based Structured Language) which defines the structure of multimedia and hypermedia documents. In this paper, we elaborate on the usage of HyTime for ancient manuscript database modelling. We derive requirements for modelling and show how HyTime can be used for such purposes: we show the relevant part of the HyTime code. In order to prove that the model can also serve as a basis for implementation, we have developed a prototype.

Features of a philological workstation

The BAMBI project entails various major technological phases and components [Cal 1996]. Key technical elements in BAMBI are the following:

- Grey level high speed microfilm scanning,
- Image compression and storage,
- Pattern recognition and image processing,
- Image browsing, indexing and retrieval techniques,
- Hypermedia processing.

Our work in the project is to model and prototype the Hypermedia database. Important results may, in particular, be expected in the following areas: transcription and annotation of manuscripts, production of fully indexed and commented critical editions of ancient texts, linkage of image and transcription, linkage between manuscript pages and manuscript retrieval.

Transcription: The transcription of the manuscript is a process which leads to noting the full transcript of a given language by means of the system of signs provided by a conversion language. In the era under consideration, the early and high Middle Ages, abbreviations can be divided into the following types : syllabic abbreviation (omission and elision of letters), abbrevation by suspension (an example is provided by the names of jurists : ac. = Accurcius, bul. = Bulgarus, etc), abbreviation by contraction (which have endings written on the line), and the use of special signs.

Annotation: Further information can be added for the historian who works on manuscripts:

- *annotations*: personal commentaries on a text. An annotation can be either indisputable (number of copies, reference to a bibliography, etc), or subject to polemic (criticism of the content, etc),
- *critical apparatus*: annotation of a particular type appearing at the bottom of a page in transcriptions, or criticisms of text to indicate differences from other studies published in the same area.

Printing the results of research such as annotations and critical apparatus is very useful. Annotations and critical apparatus are linked to words, or groups of words, in the transcriptions.

Linking:

- Consultation of the BAMBI documentary database is linked to the concept of hypertext browsing. The workstation includes the possibility of creating links between manuscripts or manuscript portions. They can be linked in terms of topic, period, etc. Thus, the user creates an operating environment around a document.
- An original aspect of the BAMBI project is that it allows *links between parts of images and the corresponding parts of texts*: original manuscripts frequently contain abbrevations and this connection can be very useful. The decomposition of the transcription into semantic sets is relatively easy to carry out, because it is sufficient to define a markup language of the text or to use a normalized language (i.e. SGML).

Indexing: The function of searching for a sequence of characters in the transcriptions of the manuscript under examination is activated by selecting the word to be found in an *index verborum*. The use of more than one script within the same manuscript (Greek and Latin for example) requires the creation of an *index verborum* for each alphabet. These indexes take the form of a list containing all the words appearing in the transcription. Each entry in the index is followed by the number of times it occurs in the manuscript, as well as on the page under examination. The aforementioned *index verborum* is coupled with a function (*index locorum*) which can display the positions in which each word occurs in the manuscript. The reference to a given word takes the form of a list containing the page number, column number, line number, and word number.

Retrieval: The philological workstation offers the user a number of search tools that can speed up the selection of documents. The user is provided with three methods of retrieval: selection from a classified list, multi-criteria search, and the use of keywords.

Modelling with HyTime

Introducing HyTime

HyTime (Hypermedia/Time-based Structured Language) [Afnor 1994]; DeRose 1994; Newcomb 1991] is an international standard which defines the structure of multimedia and hypermedia documents. HyTime is an extension of SGML (Standard Generalized Markup Language) [Afnor 1993; VanHerwijnen 1990] mainly addressing the problems of:

- *locating data,* of any type, by using a standard notation that is independent of the processing system and the data themselves,
- describing *links within and between documents*,
- *structuring contents*,
- describing *relations between temporal and spatial events* occurring in documents.

HyTime is characterized by its meta definition capabilities. It consists of a Document Type Definition (DTD) which defines a set of SGML element types, each having a semantic meaning. The elements are called *architectural forms*. Architectural forms (FA) can be compared to abstract classes in object-oriented programming. Their specification is expressed by a narrative text combined with a formal definition and associated programs. The set of

specifications of architectural forms authorized by HyTime is gathered under the name of *meta-DTD* because it represents the DTD model of the HyTime application. An architectural form corresponds therefore to the definition of a meta-element. It represents a basic information structure. An element inherited from this meta-element can then be used in any DTD. It can be specialized (content model, attributes) following rules defined by the HyTime standard. These architectural forms are then linked by relations such as generalization/specialization analogous to those encountered in object models.
Figure 1 shows an example of a hyperdocument, which includes a very simple link. This type of link corresponds to the architectural form *clink*. Figure 2-1 shows the architectural form *clink*, whilst figure 2-2 outlines the derivation of an element model 'lien' in the DTD of the hyperdocument. The element model " lien " is exemplified in the figure 2-3.

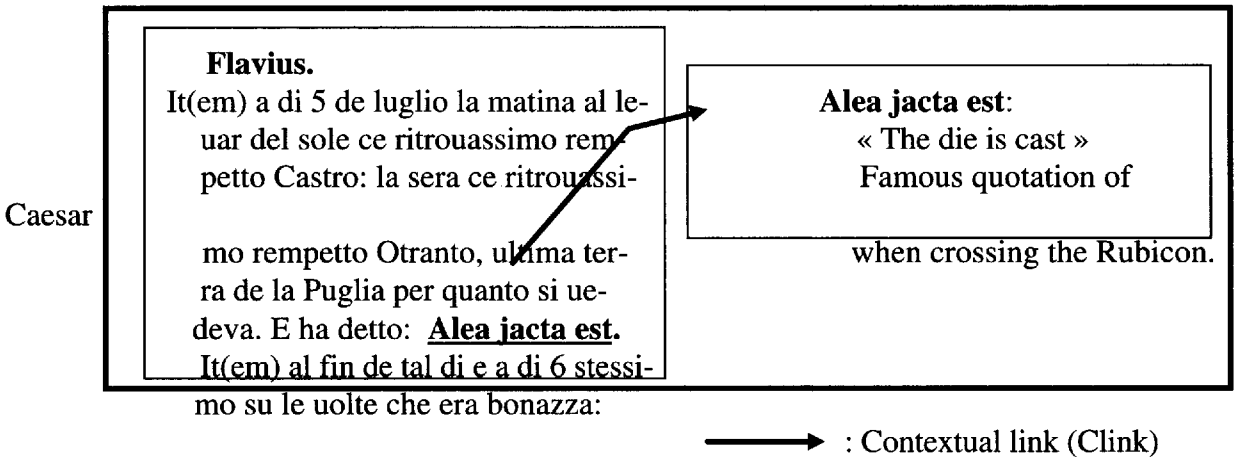

Figure 1: Hyperdocument example containing a contextual link between two nodes

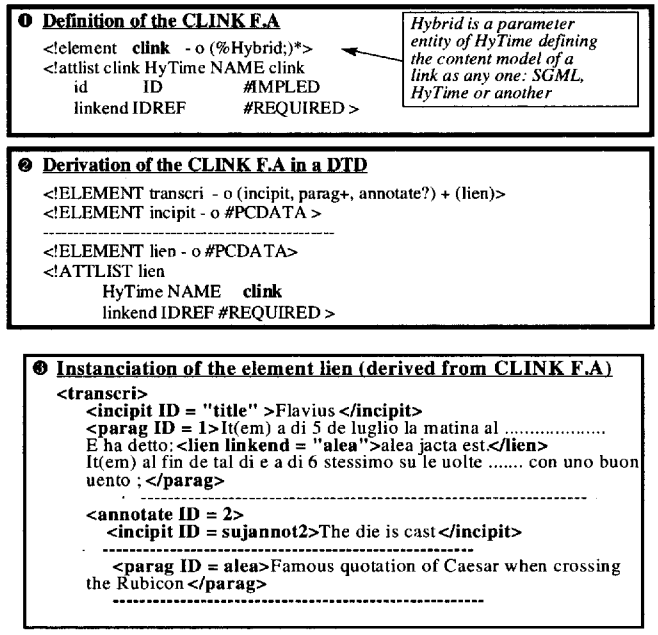

Figure 2: HyTime description of the hyperdocument presented in Figure 1

The main application area of HyTime is seen in electronic publications, as multimedia slide-shows, hypermedia dictionnaries, and the like [Burger 1995]. Thus, with regard to the modelling requirements (see section 2), we use HyTime for the kernel of the workstation for ancient manuscripts. The main reason is that SGML does not allow the specification of links between text and part of image (part of an object).

The BAMBI Document Type Definition

This section describes the HyTime conforming Document Type Definition for BAMBI (see Figure 3).

```
<!-- DTD for a class of document exported from BAMBI project -- >
<!ENTITY % doctype "MANUSCRI"    >
 <!-- Document STRUCTURE -->
<!--    ELEMENTS  MINCONTENT (EXCEPTIONS) -- >
<!ELEMENT %doctype; - -    (InfoManu, Pages*, Fin?) >
<!ELEMENT InfoManu   - -
(UserName,Title,Author,Library,Incipit,Material,Date,Size,
Languages,Handwriting,Bookmark*) >
<!ELEMENT UserName  - -   (#PCDATA)   >
<!ELEMENT Title  - -   (#PCDATA)   >
<!ELEMENT Author  - -   (#PCDATA)   >
<!ELEMENT Library  - -   (#PCDATA)   >
<!ELEMENT Incipit  - -   (#PCDATA)   >
<!ELEMENT Material  - -   (#PCDATA)   >
<!ELEMENT Date  - -   (#PCDATA)   >
<!ELEMENT Size  - -   (#PCDATA)   >
<!ELEMENT Languages  - -   (#PCDATA)   >
<!ELEMENT Handwriting  - -   (#PCDATA)   >
<!ELEMENT (Bookmark | Fin)- -(#PCDATA)   >
 <!-- Page STRUCTURE -->
<!ELEMENT Pages  - -   (Image,Transcri)+ >
<!ELEMENT Image  - -   (CoorMots*) +graphic >
<!ENTITY % CoordXY "(X1,Y1,X2,Y2)"   >
<!ELEMENT CoorMots  - -   (%CoordXY;) >
<!ELEMENT (X1,Y1,X2,Y2)  - -   (#PCDATA)   >
<!ENTITY % Annot "(Annot1|Annot2|Annot3|Annot4|Annot5|Annot6)"  >
<!ELEMENT Transcri  - -   (Curpage,(Column,Line,Mots+,(%Annot;)*))*>
<!ELEMENT Curpage - -   (#PCDATA)   >
<!ELEMENT Column - -   (#PCDATA)   >
<!ATTLIST ColumnNumColCDATA #REQUIRED   >
<!ELEMENT Line  - -   (#PCDATA)   >
<!ATTLIST Line  NumLine CDATA #REQUIRED   >
<!ELEMENT Mots  - -   (#PCDATA|Mots*)  >
<!ATTLIST Mots  Police CDATA #IMPLIED   >
<!ELEMENT (Annot1|Annot2|Annot3|Annot4|Annot5|Annot6)
  - -   (#PCDATA)   >
```

Figure 3. The BAMBI DTD

A manuscript is composed of a description (*InfoManu*, in the DTD) and of pages (*pages*, in the DTD). The element *pages* is composed of an image (ELEMENT *Image*) and a transcription (ELEMENT *Transcri*). Each word in the image is described by its co-ordinates - that is X1, Y1, X2, Y2 (ELEMENT *CoorMots*). X1 represents the X up-left corner of the rectangle around the word in the image, Y1 represents the Y up-left corner, X2 the X down-right corner and Y2 the Y down-right corner.
Each word in the transcription is located by the number of the line and the number of the column where it appears. Moreover, if an annotation exists for the word (or for a group of words), it is specified with the ENTITY *Annot*.

Instantiation of BAMBI DTD

Manuscript description

The model used for the description of the manuscript is the computerized register for the cataloguing of manuscripts recently developed by the Istituto Centrale per il Catalogo Unico (ICCU) of the Italian libraries, that is: identification of the manuscript, location, call mark, support, date, type of script, heading. The following HyTime code corresponds to the description (Element *InfoManu*) of the manuscript "Diario del viaggio in Terra Santa 1559".

```
<INFOMANU>
<USERNAME>Mario</USERNAME>
<TITLE>Diario del viaggio in Terra Santa 1559</TITLE>
<AUTHOR>Luca da Gubbio</AUTHOR>
<LIBRARY>1</LIBRARY>
<INCIPIT>Unknown</INCIPIT>
<MATERIAL>Cartaceo</MATERIAL>
<DATE>Sec. XVI 2° Meta</DATE>
<SIZE>CC 98</SIZE>
<HANDWRITING>8</HANDWRITING>
<BOOKMARK> Diario del viaggio in Terra Santa 1559 : c4r</BOOKMARK>
<BOOKMARK> Diario del viaggio in Terra Santa 1559 : c5r</BOOKMARK>
</INFOMANU>
```

Figure 4. Element InfoMenu

Link between part of image and part of text

As we can see from the portions of HyTime code below, we can define links between the part of an image and the corresponding part of a text in the transcription using the architectural form *hotspot*. The links between a word in the transcription and the corresponding part of the image is defined using the architectural form *link*. The pieces of HyTime code shown in (figure 5) give an example of such mechanisms.
For example, the word «I(tem)» in the transcription is linked with the corresponding part of the image using the architectural form *link*, and the part of the image is linked with the word (in the text) using the architectural form *hotspot*. RX, RY, RW and RH specify the co-ordinates of the part of the image corresponding to the word I(tem). The parentheses signify that " I " appears in the manuscript image and is an abbreviation of Item.

```
<IMAGE>
<HYLOC>
<HOTSPOT ID=H1_1_1 GRAPHIC = Image5  REF=T1_1_1  RX= «205,02»  RY=«75,64»
RW=«128,52» RH=«69,54»
.....
</HYLOC>
</IMAGE>
<TRANSCRI>
<CURPAGE>c4r</CURPAGE>
<COLUMN NumCol=1>
<LINE Numline=1>
        <LINK ID=T1_1_1 LINKEND=H1_1_1>I(tem)</LINK>
        ..............
</LINE>
</COLUMN>
</TRANSCRI>
```

Figure 5. Link between part of image and part of text

The BAMBI prototype

In the previous sections, we have described the usage of HyTime for manuscript database modelling. We derived requirements for modelling and showed how HyTime can be used for such purposes. In order to prove that the model can also serve as a basis for implementation, we have developed a prototype.

The philological workstation architecture (Figure 6) is composed of the following elements: the HyTime application, the HyTime engine, the SGML parser, and a relational database management system.

- The *HyTime application* is the visible part of the production chain. The HyTime application is a program which manages hyperdocuments. It determines how the document must be presented. It permits the translation of the HyTime concepts in presentation format [Buford 1994]. While the HyTime engines and the parsers are designed for all applications, the HyTime application must be developed individually. In the HyTime treatment chain, application checks all interactions with the user. In reality, when a SGML or HyTime document is treated, the application uses the HyTime engine [Newcomb et al. 1991].
- A *HyTime engine* [Buford et al. 1994; Koegel 1993] is a program (or portion of a program or a combination of programs) that recognizes HyTime constructs in documents and performs application-independent processing of them [Afnor 1994].
- A *SGML parser* is a syntactical analyzer. It checks if the document is consistent with respect to its DTD.

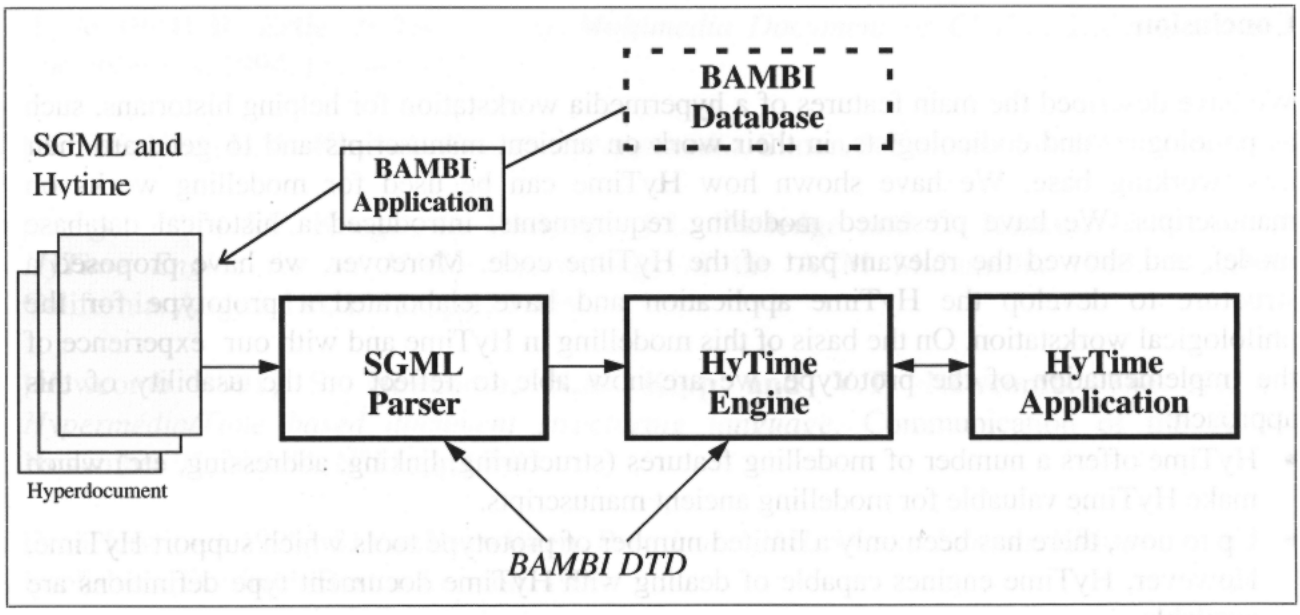

Figure 6. The BAMBI architecture

The HyTime engines supply a minimal user interface. In the BAMBI project, we have used the Synex Viewport user interface (http://www.synex.se) (Figure 7).

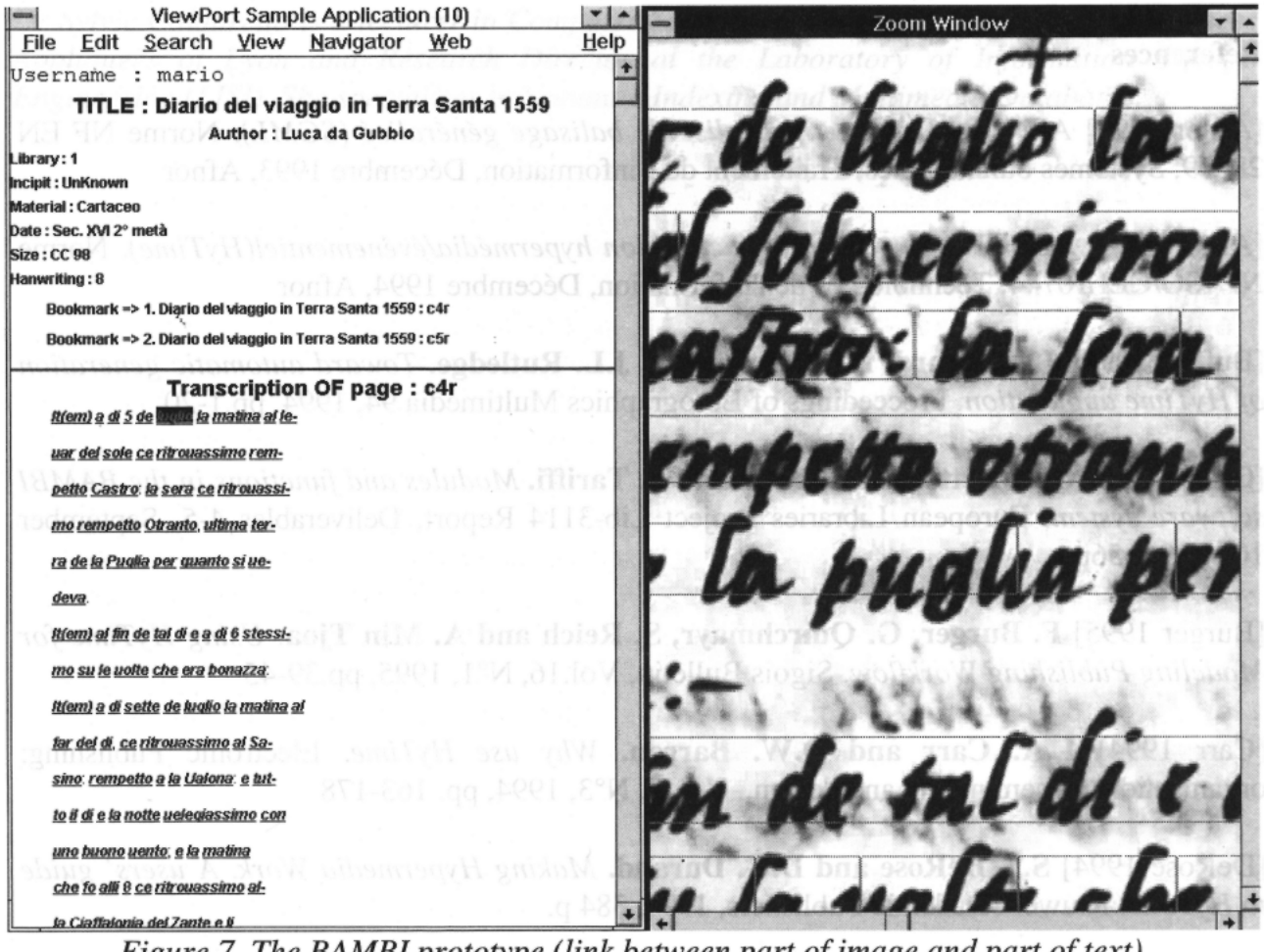

Figure 7. The BAMBI prototype (link between part of image and part of text)

Conclusion

We have described the main features of a hypermedia workstation for helping historians, such as philologists and codicologists, in their work on ancient manuscripts and to generate their own working base. We have shown how HyTime can be used for modelling works on manuscripts. We have presented modelling requirements, introduced a historical database model, and showed the relevant part of the HyTime code. Moreover, we have proposed a structure to develop the HyTime application and have elaborated a prototype for the philological workstation. On the basis of this modelling in HyTime and with our experience of the implementation of the prototype, we are now able to reflect on the usability of this approach:

- HyTime offers a number of modelling features (structuring, linking, addressing, etc) which make HyTime valuable for modelling ancient manuscripts.
- Up to now, there has been only a limited number of prototype tools which support HyTime. However, HyTime engines capable of dealing with HyTime document type definitions are available.
- HyTime is an international standard. Though there are a number of advantages and drawbacks of international standards, we consider that standardization and exchangeability for philological workstations is an advantage. In the future, we envisage the development of an information server for the on-line distribution of digital images through the Internet and using the visual tools of the World Wide Web. The link between SGML and HTML (HTML is a SGML DTD) will facilitate the transfer of our tool on the Web.

References

[Afnor 1993] AFNOR, *Langage normalisé de balisage généralisé (SGML).* Norme NF EN 28879, Systèmes bureautiques, Traitement de l'information, Décembre 1993, Afnor

[Afnor 1994] AFNOR, *Langage de structuration hypermédia/événementiel(HyTime).* Norme NF ISO/CEI 10744, Technologies de l'information, Décembre 1994, Afnor

[Buford 1994] **J.F. Buford, L. Rutledge and J.L. Rutledge.** *Toward automatic generation of HyTime application.* Proceedings of Eurographics Multimédia 94, 1994, pp.1-20

[Cal 1996] **S. Calabretto, A. Sappupo and F. Tariffi.** *Modules and functions in the BAMBI software system.* European Libraries Project Lib-3114 Report, Deliverables 4-5, September 10th 1996, 56p.

[Burger 1995] **F. Burger, G. Quirchmayr, S. Reich and A. Min Tjoa.** *Using HyTime for Modeling Publishing Workflow.* Sigois Bulletin, Vol.16, N°1, 1995, pp.39-45

[Carr 1994] **L.A. Carr and D.W. Barron.** *Why use HyTime.* Electronic Publishing: organisation, dissemination and design,. Vol. 7, N°3, 1994, pp. 163-178

[DeRose 1994] **S.J. DeRose and D.G. Durand.** *Making Hypermedia Work. A users' guide to HyTime.* Kluwer Academic Publishers, 1994, 384 p.

[Erfle 1994] **R. Erfle**. *HyTime as the Multimedia Document of Choice*. IEEE Computer Society Press, 1994, pp. 445-454

[Kimber 1995] **E.W. Kimber.** *Practical Hypermedia*. Simon & Schuster Ed, 1995.

[Koegel 1993] **J.F. Koegel, L. Rutledge, J.L. Rutledge and C. Keskin.** *HyOctane: A HyTime Engine for an MMIS*. Proceedings of the ACM Multimedia 1993, Anaheim, California. August 1-6, 1993, pp.129-136

[Newcomb 1991] **S. Newcomb, N.A. Kipp and V.T. Newcomb.** *The HyTime Hypermédia/Time based document structuring language*. Communication of the ACM, November 91, Vol.34, N°11, pp.67-83

[VanHerwijnen 1990] **E. VanHerwijnen.** *Practical SGML Second Edition*. Kluwer Academic Publishers, Dordrech/Boston/London. 1990

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). She 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 Appliquees of Lyon and Research Director at the Laboratory of Information Systems Engineering (LISI). She specialises in Semantic Indexing and Multimedia Databases.