The EUMME Project:Towards a New Philological Workstation

Andrea Bozzi[1], Sylvie Calabretto[2], Maria Sofia Corradini[3], Bruno Tellez[4]

[1] ILC-CNR
Via G. Moruzzi, 1
I-56124 PISA
andrea.bozzi@ilc.cnr.it

[2] LIRIS CNRS UMR 5205
INSA de Lyon
Bâtiment Blaise Pascal
7, avenue Jean Capelle
F-69621 Villeurbanne Cedex
sylvie.calabretto@insa-lyon.fr

[3] Dip. Di Lingue e Letterature Romanze
Università degli Studi di Pisa
Via S. Maria
I-56000 PISA
m.s.corradini@rom.unipi.it

[4] LIRIS CNRS UMR 5205
Université Lyon 1
Bâtiment Nautibus
43, boulevard du 11 novembre 1918
F-69622 Villeurbanne Cedex
bruno.tellez@liris.cnrs.fr

Abstract

The EUMME project[1] (Euro-Mediterranean Union within the framework of Medieval Medicine) is intended to strengthen the international collaboration among Centres involved in transmitting medico-pharmaceutical culture in order to witness the fundamental role it has played for the development of an European scientific community. In this respect, the contributions provided by different linguistic and cultural environments (Arabic, Hebrew) will be given special attention. Those ambiences were in fact strongly integrated in what we could call a universal dimension where there was full awareness of the supranational dimension of science, which is able to overcome geographical borders and ideological and religious barriers. These aims can only be achieved by means of modern technology which, similar to one of the objects of its analysis, overcomes borders and is not subject to ideological assessment, fostering the exchange of data among the Centres involved in the project.

More specifically, the aims are referred to three large sectors and can be summarized as follows:

- to increase the knowledge of the data relative to the subject treated (philological-textual sector);
- to increase linguistic knowledge (linguistic-lexicographic sector);
- to develop technological tools specific for the study and dissemination of the information produced (sector of technological tools).

[1] Actual partners are Institut für Romanische Philologie, Freie Universität Berlin (Germany), Istituto di Linguistica Computazionale, CNR, Pisa (Italy), LIRIS - Laboratoire d'InfoRmatique en Images et Systèmes d'information (Lyon-France), Facultad De Filología, Università di Salamanca (España), Department of European Languages at The University of Wales (UK) and the IRHT - Institut de Recherche et d'Histoire des Textes (CNRS-France).

1 Introduction

Isidor of Seville[2] assigned to medicine the statute of discipline comprehensive of all liberal arts, owing to its opening to different spheres of thought such as philosophy, logic, astrology, botany, theology, alchemy. He gave it an unitary value that makes medicine a true synthesis of the Medieval age. The period ranging between the 5th and 14th century is characterized by an extraordinary phenomenon that only in the modern age, with the development of new means of communication, has definitely been established: the interconnection among different disciplines and the interrelation among specialists belonging to various countries. In other words, in that period, we are for the first time in the actual presence of a number of scientists in different territories distributed across all the Mediterranean countries: Italy, France, Spain (on the North/North Western coast), north Africa (on the south/south Western coast), and those located in the east/south east, that is Egypt, Syria and Iraq.

At the root of this phenomenon there are certainly historical, political and socio-economic reasons, but the lack of original medieval documentation in some cases constituted by little-known and unedited sources has made it difficult to perform in-depth studies on the subject. The phenomenon is not only interesting in itself, but it also forms an invaluable basis for the knowledge of the medical science and, more in particular, of the human anatomy which has played a fundamental role also for humanistic and Renaissance art (Leonardo da Vinci, Michelangelo Buonarroti). Another important element to be considered is the linguistic aspect of this theme: through medieval medicine Greek, Arabic and Hebrew terms were introduced in the West by the activity of doctors and philosophers acting as mediators and translators with the merit of contributing, in collaboration with scholars of law - a subject which at that time was strictly related to medicine - to the foundation of one of the most ancient Universities in Europe.

The subject of Medieval science, also including medicine, has been faced by a large number of scholars who have continued to privilege sectors such as mathematics, astrology and botany. Only those who have devoted themselves to the study of botany have, although not systematically, highlighted the therapeutic functions of some herbs. A similar situation applies to the studies on mineralogy: attention was dedicated to the medieval concepts of the properties - more magical than chemical - of stones, as is read in numerous medieval lapidaries. The same could be said with regard to treatises on animal zoology ("bestiaries") in which the ancient authors provided imaginative descriptions of the relations (characterial, physiognomic, etc.) between animals and men.

An important step forward in the knowledge of the intercultural relations regarding the subject of medicine in the Middle Ages, with evident corollaries relative to botany (for pharmacopoeia) and zoology (for anatomy) could be made if the Centres involved in these studies were to set up a Consortium for scientific collaboration in the framework of a project of common interest.

From a technological point of view, the Institute of Computational Linguistics of the National Research Council in Pisa, Italy, and the LIRIS at the INSA of Lyon, France, have been studying a workstation for the management

[2] Cfr. *Etymologiae* (ed. Lindsay 1911), book IV, chapter 13.

of data concerning texts and images of ancient documents. Over the past years these two Institutes have collaborated towards the development of such a workstation which is specialized in handling large digital archives for linguistic and philological research (Figure 1).

The study developed within the UE Libraries Programme and completed in April 1997 was called "Better Access to Manuscripts and Browsing of Images (BAMBI)". The Consortium was formed by A.C.T.A. (coordinator, Florence), the Central National Library (BNR, Rome), National Research Council (ILC-CNR, Pisa), Pisa Research Consortium (CPR, Pise), LISI-INSA (Lyon) and the Max Planck Institut für Rechtsgeschichte (Frankfurt a. M.).

Some years later, the Italian and French institutions cooperated within the STEMA project (P.A.I Galilée 1999-2000, between LIRIS Lyon and ILC-CNR Pisa), and the Web version of the Philological Workstation BAMBI was designed and developed as a prototype.

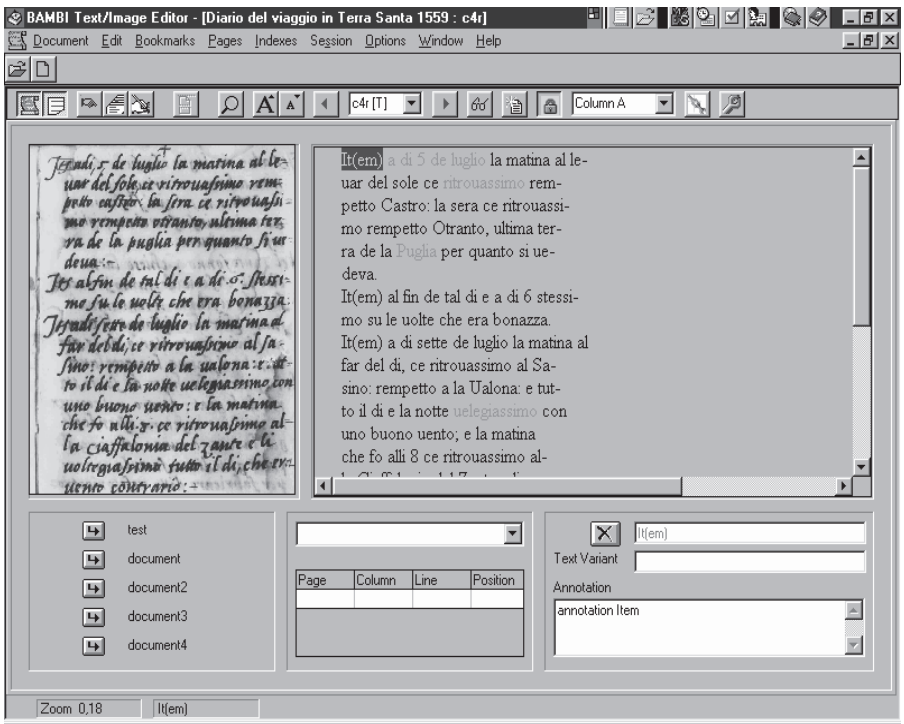

Figure 1: The BAMBI workstation displaying a medieval manuscript

2 Creation of EUMME digital archive

The aim is to implement an archive containing a large sample of texts relative to medieval medicine and pharmacology in Latin, Greek, Occitan, Catalan, Castilian, Italian and ancient French. The archive includes primary sources and a selection of secondary sources.

By the term ‘primary sources’ we intend handwritten documents which have already been reviewed and which contain works of medicine, pharmacopoeia, botany (and possibly alchemy). The linguistic field is represented by the main Romance languages, but the elements displayed by these languages in technical works

of the same period written in Hebrew will also be presented. This aspect will be dealt with by the group of research of the Freie Universität of Berlin directed by Prof. Guido Mensching, who has worked on a project on this subject (*Edition and Analysis of Medieval Hebrew Medical Synonym Lists with Romance, Latin and Arabic Elements*). Ancient printed editions will be included in the textual archive.

The 'secondary sources' are made up of the following elements:

- bibliographic information for each primary source;
- comments, studies, essays (reproduced in digital format), in the case of ancient material, extremely useful from a historical and documentary point of view;
- reference to Latin, Greek, Arabic and Hebrew sources.

3 Production of EUMME knowledge and lexicons

The aims of this second objective are to:

- encourage operations of logical, philological and linguistic links among those parts of the primary sources which are strictly connected with one or more secondary sources;
- make it possible to carry out digital reproduction of the secondary sources not protected by copyright which are difficult to recover. After they have been made available on the Web infrastructure, they can be easily accessed by specialists in the field;
- foster the on-line publication of essays and preparatory comments to more complex works, which can also play a fundamental role for didactics, involving Web-based technology in the University ambience.

The most important and final product of the project will be a selection of meaningful terms in the medico-pharmaceutical lexicon. It will be structured according to the international standards of computerized lexicography on the Web as they have been encoded by TEI (Burnard & Sperberg-McQueen, 2000) and which are currently used in a large number of lexicographical projects (with regard to medieval lexicography see, for example, the Anglo Norman Dictionary on-line edited by David Trotter).

The choice of XML-TEI codes will make the lexicographic data of medicine compatible with other projects which could be integrated one with the other.

4 Development of EUMME technological tools

The main technological component is represented by a workstation for the handling of texts and digital images reproducing the original sources on which the project will be based.

Such workstation makes it possible to:

- look-up an image archive with digital representation of the source documents on a high resolution monitor,
- transcribe, annotate and index the text presented in the images,
- view the transcribed version and the Index Locorum in a window adjacent to the display of the source document,

- automatically match each word of the transcription, of the Index Locorum and of the annotations with the portion of the source document image in which the word is found,
- allow the compilation of the lexicographical items from the textual and image archives,
- export information on manuscripts in the form of XML-formatted text.

The new philological tool will include collaborative aspects. Multi-sites and multi-users diffusion will have to be associated with an increased effort towards the security of exchanges and the protection of the intellectual property of each resource (image, transcription …). The integration of the existing philological workstation (Calabretto & Bozzi, 1998), (Bozzi & Corradini, 2004) in a distributed context needs to propose the appropriate architecture ensuring the integrity of information and the security of data in this new context. Moreover, the integration of digitized documents will become really complete if new tools allow to assist users in the content extraction from images of ancient manuscripts.

4.1 Workstation Architecture

The foundation of a distributed architecture is built on the ability to access distant documents. It opposes two approaches: a centralized vision and a node to node approach (like peer-to-peer). The centralized solution relies on the principle of information centralization and concurrent access by clients. On the other hand, the peer-to-peer (P2P) approach is based on the principle of direct linking among different users for file exchanges. If the centralized approach has undeniable advantages for information management (available in only one place and so, efficient for its collaborative features), problems of reliability and security have to be strongly considered. In addition to information scattering, the P2P architecture solution presents a more effective and precise control on the information, by offering dedicated architecture for file exchange. The definition of the system architecture will be proposed with respect to the user requirements generated by the collaborative aspects of the project. Indeed, users should be able to work together on the same document at the same time, or in successive order but with full awareness of work carried out by other partners. The follow-up of the document (workflow) also allows to enrich the system with a study of procedures that specialists implement to process their documents.

4.2 Security

As explained before, the first level of information security can be made through a network with connections on demand. It will protect the system from external intrusions. Moreover, information exchange from one node to another is provided so that information interception is difficult. A second level will consist in protecting the resource from direct access. It is possible to encrypt the resource with advanced encoding protocols (for example RSA encryption techniques). Open solutions like Ants or Mute provide solutions to combine P2P and encryption solutions. Finally, the last level of protection concerns the resource contents which will be protected. With regard to images, the watermarking techniques (Katsenbeisser & Petitcolas, 2000) can be chosen because they allow to insert marks inside the image itself and then to verify copyright for use and dissemination.

4.3 Image processing

In order to optimize and improve the works of specialists, it is necessary to obtain more than one image from the digitized documents. The nature of digitized documents (printed manuscripts) and also the possibility to dispose of transcriptions makes it possible to look ahead and answer user requirements. For example, searching the occurrences of a word inside the non-transcribed document requires almost complete automatic extraction of the image content. It will be then necessary to provide solutions for improving the documents (cleaning of the verso), extracting the information (segmentation) or finding similar patterns (techniques of correlation). Moreover, higher level techniques relative to perception will allow to obtain information on the structure of the document: globally by identifying paragraphs, dropped initials, illustrations or more locally by identifying the lines or words which constitute it.

Finally, to allow image dissemination, two conditions will be necessary: image securing (discussed before with watermarking and encryption) but also image compression (for optimising the diffusion of the huge images over the network). This last technique can also contribute to improve security. Indeed, the diffusion of compressed images (low resolution) naturally limits the unauthorized employment of these images.

5 Conclusion

The EUMME project appears able to give a challenge to the solution to the problem of transmitting any information (text and/or related images) to external users (either specialist or generic), so to be available in different historical and philological environments. In effect, the use of standards for the metadata description (Doerr, 2003), (Powell & Johnston, 2003), the exchange and the management of information allow users to access a huge amount of homogenous digital archives in a collaborative way.

References

Bozzi A. Corradini M.S. (2004). The Diphilos workstation : a computational system for digital philology. *Linguistica Computazionale,* 16-17, 47-77.

Burnard L. Sperberg-McQueen C.M. (2004). Guidelines for Electronic Text Encoding and Interchange XML-compatible edition, from http://www.tei-c.org/P4X/, April the 26th 2005

Calabretto S. Bozzi A. (1998). The Philological Workstation BAMBI (Better Access to Manuscripts and Browsing of Images). *International Journal of Digital Information (JoDI)*, 1-3, 1-17.

Doerr M. (2003). The CIDOC CRM – An Ontological Approach to Semantic Interoperability of Metadata. *Artificial Intelligence Magazine (AI)*, 24-3.

Katsenbeisser S. Petitcolas F. (2000). Information Hiding, techniques for steganography and digital watermarking. Artech House Inc, Norwood

Powell A. Johnston P. (2003). Guidelines for implementing Dublin Core in XML, from http://www.dublincore.org/documents/dc-xml-guidelines/, April the 26th 2005