in ACM Symposium on Document Engineering, ACM Doc Eng 2003, Grenoble, France, pp. 55-57, 2003.

Digitizing Cultural Heritage Manuscripts: the Bovary Project

Stéphane Nicolas
Université de Rouen
Laboratoire PSI
UFR Sciences et Techniques
F-76821 Mont-Saint-Aignan Cedex
Stephane.Nicolas@univ-rouen.fr

Thierry Paquet
Université de Rouen
Laboratoire PSI
UFR Sciences et Techniques
F-76821 Mont-Saint-Aignan Cedex
Thierry.Paquet@univ-rouen.fr

Laurent Heutte
Université de Rouen
Laboratoire PSI
UFR Sciences et Techniques
F-76821 Mont-Saint-Aignan Cedex
Laurent.Heutte@univ-rouen.fr

ABSTRACT
In this paper we describe the Bovary Project, a manuscripts digitization project of the famous French writer Gustave FLAUBERT first great work. This project has just begun at the end of 2002 and should end in 2006 by providing an online access to an hypertextual edition of "Madame Bovary" drafts set. We develop the global context of this project, the main objectives, the first studies and the considered outlooks for the project's carried out.

Keywords
Digital libraries, Genetic edition, Hypermedia, Indexation, Document image analysis.

1. INTRODUCTION
Libraries and museums contain collections of a great interest which can not be shown to a large public because of their value and their state of conservation, therefore preventing the diffusion of knowledge. Today with the development of the numerical technologies, it is possible to show this cultural patrimony by substituting the original documents by numeric high quality reproductions allowing to share the access to the information while protecting the originals. These last years, numerous libraries have started digitization campaigns of their collections. Faced to the mass of the numerical data produced, the development of digital libraries allowing access to these data and the search for information becomes a major stake for the valuation of this cultural patrimony. However such task is difficult and expensive, so requires prior studies concerning the technical means to use.

Although they have a great interest for the study and the interpretation of literary works, modern manuscripts have been adressed by few digitization programs because of the complexity of such documents and the lack of adapted tools.

We present in this paper the Bovary Project, a digitization project of modern manuscripts concerning especially FLAUBERT's manuscripts and we discuss the underlying principles, difficulties and technical aspects related to such a project.

In section 2 we present the global context of this project and the aimed objectives. We overview in section 3 the few critical editions avalaible in electronic format and the projects dedicated to the digitization of manuscripts. Related problems and possible solutions are adressed in section 5. We conclude in section 6 with some outlooks.

2. GENERAL PRESENTATION OF THE PROJECT
Recently the municipal library of ROUEN has begun a program for digitizing its collections. For this purpose an efficient system of digitization allowing a high resolution display of the digitized documents has been purchased. One of the first aims of this program is the digitization of a manuscript folder compound of almost 5 000 original manuscripts issued from "Madame Bovary", a well known work of the french writer Gustave FLAUBERT. This set of manuscripts constitutes the genesis of the text, it means the successive drafts which highlight the writing and rewriting processes of the author. This digitization task is still in progress, and will be completely achieved in a couple of month.

The final aim of this program is to provide an hypertextual edition allowing an interactive and free web access to this material. Such an electronic edition will be of great interest for researchers, students, and anyone who wants to see FLAUBERT's manuscripts, especially because there is no critical edition of a full literary work of this author available on the web. This project called "Bovary Project" is a multidisciplinary project which implies people from different fields of interest: librarians, researchers in literary sciences and researchers in computer science.

The digitized corpus is composed of almost 5 000 manuscripts. This set of manuscripts constitutes the genesis of "Madame Bovary", it means the pre-text material of this work. Flaubert's drafts have a complex structure. They contain several blocks of text not arranged in a linear way, and numerous editorial marks (erasures, words insertion,...). So these manuscripts are very hard to decipher and interpret. To provide an electronic version of a such corpus is a challenging task which have to respect some requirements in order to meet users needs.

3. GENETIC EDITION REQUIREMENTS

In literary sciences the study of modern manuscripts is known as genetic analysis. This analysis concerns the graphical aspect of the manuscripts and the successive states of the textual content. In fact the nature of a manuscript is dual. A manuscript should be considered as a pure graphical representation or as a pure textual representation. A manuscript is a text with graphical interest [1].

As modern manuscripts reflect the writing process of the author, they may have a complicated structure and may be difficult to decipher. So in a genetical edition, transcriptions are generally joined to the facsimile of manuscripts. A transcription allows an easier reading of the manuscript. One can distinguish two transcription types: the linear one and the diplomatic. A linear transcription is a simple typed version of the text, which uses an adapted coding to transcribe, in a linear way, complex editorial operations of the author (deletation, insertion, substitution) sometimes located over one or several pages, even though the diplomatic one have to respect the physical aspect of the manuscript, it means the disposition of graphical elements in the page (text line, erasure, insertion,...).

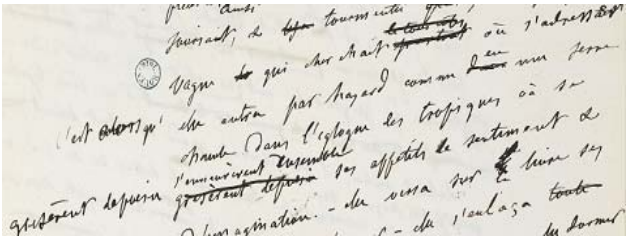

Figure 1: a manuscript fragment

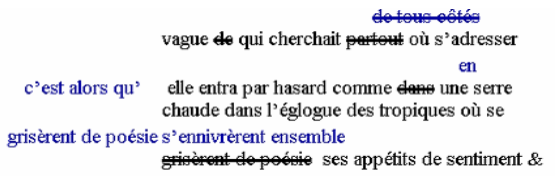

Figure 2: associate diplomatic transcription

The composition of a genetic folder consists in locating and dating, ordering, deciphering, and transcribing all pre-text material. A genetic publishing presents the results of such a work, it means an ordered set of manuscripts constituting the genesis of the text and their associated transcriptions, and allows to glance through this set of heterogeneous data.

An electronic version of a such critical edition must provide the same functionnalities than a traditionnal one, but furthermore have to allow to deal easily with the textual representation and the graphical representation of the manuscripts and create links between each other , using the capabilities of structured languages and hyperlinks.

4. RELATED WORK

In spite of the development of multimedia technologies and the possibilities provided by structured languages and hypertext, not many electronic versions of genetic publishing have been released up to now. This is probably due to the difficulty of such a task and the lack of professionnal tools dedicated to the manipulation of ancient or modern manuscripts.

In the following we overview the existing genetic editions available in electronic format (CD-ROM or website), and we discuss the capabilities and limitations of these editions. In a second part, we overview the latest projects related to the digitization of manuscripts.

4.1 Critical Publishing

Among the numerous digitized literary works available in text or image mode on Gallica[1], the server of the French National Library, two thematical editions related to the genesis of Emile Zola's work "Le rêve" and Marcel Proust's work "Le temps retrouvé" are proposed. These electronic publishing allow to visualize images of the author's handwritten notes (in black and white TIFF format or Adobe PDF files) and associated textual transcriptions in HTML. These releases contain a lot of explicative notes about the history of these works, but don't provide any capabilities to work on the manuscripts. They are more pedagogic editions than genetic ones.

The critical edition of Flaubert's work "Education sentimentale" proposed by Tony Williams[2] is very interesting even though only one chapter of the work is adressed. In fact it is the only genetic publishing of a work of Flaubert available in electronic format and it shows the complexity of Flaubert's writing process. In spite of the wish of the designers to develop a website dedicated to a large public, its use is quite difficult for non experts of literary work. The interface is not ergonomic and no tools for the study of manuscripts are provided.

The commercial genetic edition of André Gide's work "Les caves du Vatican" available on CD-ROM [3] is certainely the most achevied critical publishing available in electronic format. It allows to visualize Gide's manuscripts and associated transcriptions. Tools for image manipulation are provided and multiple acces to the text are possible using different thematic tables (characters, places, keywords, ...) and a search engine, allowing different reading of the work, according to user's needs . As we can notice, few critical editions are available in electronic format, and generally concern material with less genetic complexity than Flaubert's manuscripts or don't deal with the full text.

4.2 Manuscripts Digitization Projects

Considering the lack of tools adapted to the work on manuscripts and to the edition of electronic documents from handwritten sources, some projects have tried these last years to define the needs of the users of virtual libraries and to propose different environments devoted to the study of handwritten material. Most of them have led to the development of a workstation prototype. We briefly present the projects which are similar to the Bovary project and discuss the proposed technical solutions.

The problems related in PHILECTRE [3] are very similar to ours. The aim of this project was to explore the techniques needed by researchers in literary sciences, especially geneticians and medievists. In the context of this project, a workstation prototype dedicated to the edition of critical material was developed. This workstation integrate document analysis modules and interactive tools in order to provide an help for the transcription of

[1] http://expositions.bnf.fr/

[2] http://www.hull.ac.uk/hitm/

manuscripts and to perform an automatic coupling between this structured textual representation and the manuscripts' images. However the proposed algorithm for document analysis doesn't work well with complex documents like Flaubert's drafts. Furthermore this propotype has never been used for the edition of a full literary work.

Similar aspects have been adressed in BAMBI [4] in the case of ancient manuscripts. The aim of this project was to provide tools facilitating the transcription of medieval texts. However the problems encountered when dealing with such documents are very different to the ones related to the work on complex modern manuscripts, in particular in the case of document analysis.

The DEBORA project [5] tried to define the needs related to the use of virtual libraries and electronic books. The aim of this project was to develop tools for the digitization and the access to a selection of book of the 16th century. The results of this work have allowed to provide an environment dedicated to collaborative studies on such digitized collections. This workstation is not designed for authorial manuscripts but some important technical aspects have been adressed during this project, and among them the problem of images format and compression.

Once again we can notice that few projects have adressed the problem of electronic manipulation of manuscripts and they have proposed specific solutions adapted only to limited categories of manuscripts (generally medieval text).

5. RELATED TECHNICAL ASPECTS

5.1 Image Compression

One of the first problem we have to adress concerns the storage and the format of manuscripts images. In fact as manuscripts are studied for their graphical aspect, the facsimile has to be as conform as possible to the original, it means that the images must have a high quality. However a high resolution leads to files of great size (several mega-octets), and so implies long time of loading in the case of an online edition. So it is necessary to compress the images in order to reduce the files to a size less than 1 Mo. The choice of an adapted compression algorithm is a real problem. With the GIF and JPEG formats commonly used on the Internet, the degradation on the images is too important. New algorithms like DjVu seem to be more adapted to the compression of document images, and are more and more used in digital libraries. They provide good compression ratio while preserving good image resolution, using two different methods for the compression of the foreground (text) and the background. An improved algorithm using the same principle has been developed in DEBORA for printed documents.

5.2 Indexation and Information Retrieval

The advantage of an electronic edition is to provide multiple access to the text and to allow the user to find specific information. To provide such capabilities it is necessary to proceed to an indexation of the manuscripts images. But image indexation is a difficult task. A possible solution is to provide structured transcriptions of the textual content of manuscripts. So the choice of a structured language adapted to the encoding of critical edition is necessary.

5.3 Transcription Production

The transcription of the manuscripts requires a collaborative work with the researchers and the specialists of Flaubert. At the begining of this project we don't have any transcriptions of the manuscripts. Generally this task is performed manually using simple text editors. In order to facilitate this task and to produce a structured textual representation adapted to users' requirements, it is necessary to provide an editing environment integrating document analysis methods and interactive tools.

5.4 Document Analysis

As projects like PHILECTRE and BAMBI have shown, document analysis can bring an help for the transcription and indexation of manuscripts. Document analysis consists in extracting the structure of a document. In the case of manuscripts, it consists in extracting text lines and graphical elements like erasures. In spite of progress maded in the analysis of machine printed documents, the analysis of handwritten documents is still a challenging task. The few proposed methods [6] are not robust enough to deal with complex documents like modern manuscripts, so it is necessary to develop a new algorithm for the analysis of such documents.

6. CONCLUSION AND FUTURE WORK

As we have seen, document analysis can provide an help for the transcription of modern manuscripts, and allows a text-image coupling between these structured textual representation, which is require for computerized processing, and the image representation which describes the graphical aspect of the manuscript. However no document analysis method is now able to deal with such complex and weakly structured documents. Therefore our future work will consist in the development of a robust method taking into account specific features of such documents and using machine learning techniques.

7. ACKNOWLEDGMENTS

Our work is sponsored by the "Conseil Régional de Haute-Normandie".

8. REFERENCES

[1] J. André, J.D. Fekete, H. Richy. Mixed text/image processing of old documents. In congrès GuTenberg, pages 75-85, 1995.

[2] A. Goulet. Genetic publishing on CD-ROM of André Gide's book "Les caves du Vatican". Gallimard Eds.

[3] L. Likforman-Sulem, L. Robert, E. Lecolinet, J-L. Lebrave, B. Cerquiglini. Edition hypertextuelle et consultation de manuscrits : le projet Philectre. Revue Hypertextes et Hypermédias. Vol. 1, No. 2-3-4, pages 299-310. Hermès, Paris, Sept. 1997.

[4] A. Bozzi, A. Sapuppo. Computer-aided Preservation and Transcription of Ancient Manuscripts. Ercim News, Vol. 19, pp. 27-28, 1994.

[5] C. Belisle, C. Hembise. Etat de l'art sur les pratiques et sur les usagers des bibliothèques virtuelles. DEBORA project, release n°2.1, 1999.

[6] E. Bruzzone, M.C. Coffetti. An algorithm for extracting cursive text lines. ICDAR 1999: 749-752