Open Source Eminescu's Manuscripts: A Digitization Experiment

Gabriela Dumitrescu[1], F. G. Filip[1], Angela Ioniță[2], Cornel Lepădatu[1]

[1] Romanian Academy Library,
125 Calea Victoriei, Bucharest 1,
gabi_biblacad@yahoo.com; ffilip@biblacad.ro; cornel_lepadatu@biblacad.ro

[2] Research Institute for Artificial Intelligence, Romanian Academy,
13 Calea 13 Septembrie , Bucharest 5,
aionita@racai.ro

Abstract: The paper describes a practical digitization project which was carried out by the Manuscripts Department and the Information Technology (IT) Department of the Romanian Academy Library (BAR) under the coordination of the Romanian Academy and with the support the Romanian Ministry of Culture. Mihai Eminescu is the greatest Romanian poet and his manuscript collection consists of 48 notebooks, over 14,000 folios, without a strict chronological order and related topics. Main objectives of the project were a) ***creating*** digital collections of Mihai Eminescu's manuscripts in order to preserve the original and prepare a facsimile edition; b) allowing users to users through special collections available online for the research, publications on CD-ROM or DVD-ROM, and the WWW.

Keywords: archival resources; digitized manuscripts; facsimile edition; IT platform; preservation.

1. Introduction

Mihai Eminescu (1850-1878) is Romania's national poet (http://www.mihaieminescu.eu/). His literary manuscripts, correspondence and biographical notes are kept in the Romanian Academy Library (BAR; www.biblacad.ro): BAR holds several collections of a great interest (the collection of Mihai Eminescu's literary manuscripts is one of them) which can not be displayed on the shelves to a large public because of their value and physical preservation reasons. Nowadays, with the development of new technologies, it is possible to allow access to this cultural heritage by substituting the original documents with high quality digitized reproductions allowing sharing the access to the information while protecting the original.

In this paper the term "open source" will not be used as in the IT specialized language, but in a larger sense, in order to express the intention of "opening" the treasure of the cultural resources for all the readers, at three different levels (Figure 1). The project is open because it allows *building/creating the resources* represented by digital collections of Mihai Eminescu's manuscripts; it is also meant for *designing* a system of access easily to be used at different levels, from advanced scholars to university and high school students and the general „reader"; having a strategy for *delivering.*

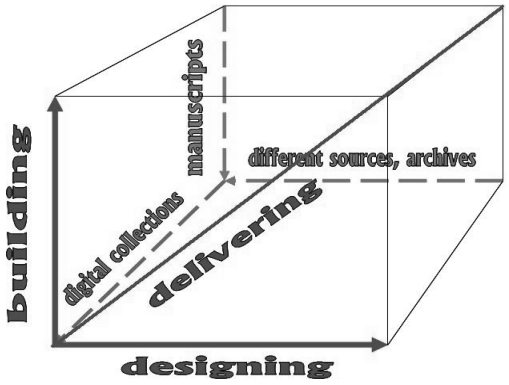

Figure 1. Three dimensions of e project *Mihai Eminescu – manuscripts*

The project entitled "Mihai Eminescu - Manuscripts" was carried out by The Romanian Academy Library and it was initiated and coordinated Acad. Eugen Simion, former president of the Academy, and supported by the Ministry of Culture. The remaining part of this article is organized as it follows according to several criteria. After some preliminary considerations the activities during the project are to be reviewed. Then several technical issues are presented such as preservation, formats, reproducing the text as an image and so on. The production of facsimile volumes is mentioned before the conclusions of the paper.

2. Preliminary Considerations

In the context of digitizing manuscripts of patrimony, the study of modern manuscripts

is known as *genetic analysis* (Cerquiglini, Lebrave, 1997). This analysis concerns the graphical aspect of the manuscripts and the successive states of the textual content. Actually the nature of a manuscript is dual. It could be considered either as a pure graphical representation or as a pure textual representation. A manuscript is a text of graphic interest as well. As modern manuscripts, Eminescu's work (see Fig. 2) reflects the writing process and style of the author. Consequently, they may show a complicated structure and may be difficult to decipher.

The data structure was a major decision in the design and implementation of a data repository. There are several options to consider (Knoll, 1997). such as the complexity of the pure SGML and the accessibility of documents encoded in the HTML. Therefore, the library option of cataloguing and classification has been the foremost tool in organizing the content, the library's own cataloguing and classification rules and international standards, such as ALEPH® system of ExLibris (2009) and Mark bibliographic description (Fritz, 2002) were utilised.

project utilised the state-of-art digitizing technologies; a Zeutschel scanner, Book cradles for careful document fixation, and digitization software.

Digital reformatting of rare library materials such literary manuscripts lead to the creation of large quantities of images, aiming to serve users' requests. There are various techniques to make users work easier. These techniques must consider the conditions under which the images will be used and their implementation requires a special effort from the institutions assuming the task of providing online image database.

The user will need to access our data via Internet and therefore we have to base our project structure on this access type.

Due to the technical and institutional challenges involved, the project was also intended as a test-bed for the problems and challenges involved in creating multimedia archives of global scale (Waring, 2001). The project pursued a number of strategies for creating more flexible modes of use. Each digital copy of a manuscript had to be presented in a certain structure which should correspond

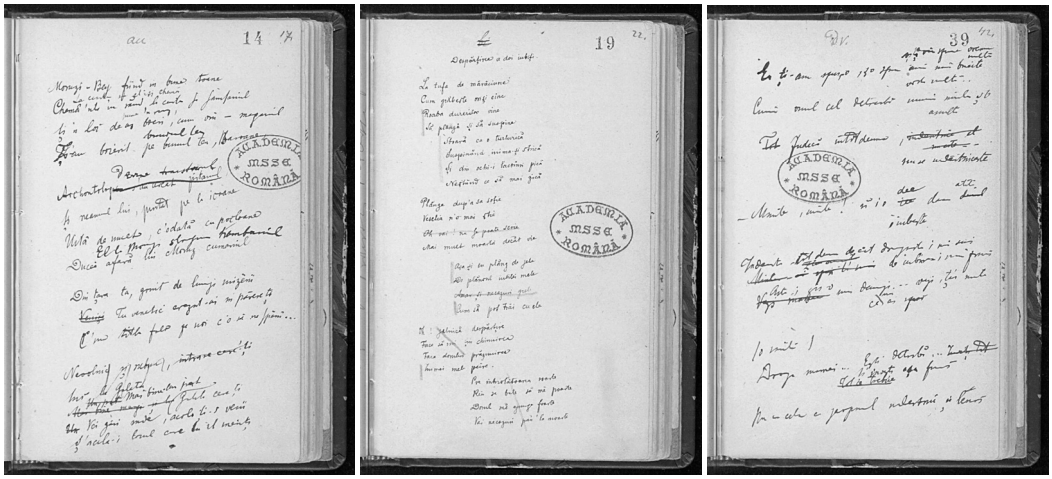

Figure 2. Mihai Eminescu's manuscripts

3. Activities

Since 2001 the team of the ***Eminescu's manuscript*** project has been constructing electronic environments for teaching and research based on digital copies of primary documents, including texts, high resolution page images of manuscripts. As a collaborative work between the Manuscript's Section and the IT Section of BAR with the funding from the Ministry of Culture, the

intuitively to the component parts of the original. On the base of the experience of Mayer and Knoll (1996) and after exploring various other solutions, a system which aims to be independent to the extent possible of any other system or platform was developed. All data, which are added to the image, were recorded in the extended HTML. The reason for this solution was the request of long-run applicability of data, as mentioned above in the section, concerning the purposes of

digitization. The problem was solved by dividing the preparation of the description of manuscripts into three parts (Figure 3).

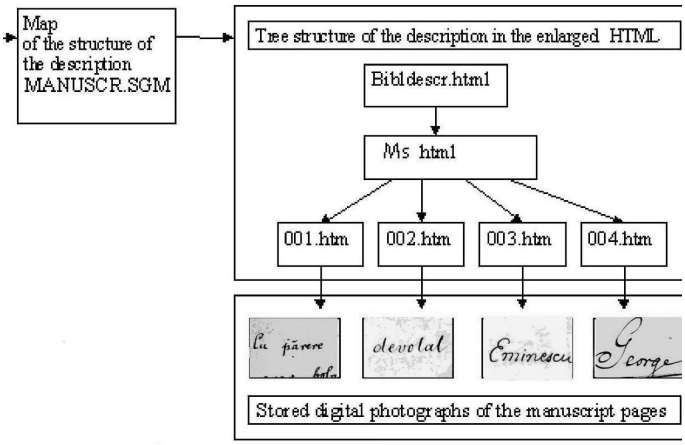

Figure 3. Tree structure of the description

First, on the base of well-known properties of the manuscripts the text – file was generated without concrete information, but in the structure, needed for this manuscript. At the same time, every future image was assigned a name with the description of the respective page. Then the experts added concrete information about the manuscript, and, if needed, even more information about individual pages was added.

When preparing the routine production of digitized manuscripts, one of the biggest concerns was the SGML (Standard Generalized Markup Language (ISO 8879:1986) platform. A trade-off between the complexity of the pure SGML and the accessibility of documents encoded in the HTML (Hypertext Markup Language) had to be found.

The problem was how to mark-up the contents and format it at the same time, while using for the basic work a normal web browser. The solution was an enlargement of the HTML DTD (Document Type Definition) with introduction of a few content-oriented elements to enable the mark-up of contents. Thus, a kind of hybrid format was created containing the formal mark-up for display together with content oriented elements. From the point of view of the user, the appearance was that of a labeled format in which each element was labeled by its name. After the digitization had been completed in June 2004, the work was continued, in accordance with Master project, with the view to completing the bibliographic descriptions and including the section entitled *Mihai Eminescu, Programme of digital access to manuscripts* into the database of the library (Figure 4).

Figure 4. Database of the library - Access to manuscripts of Mihai Eminescu

4. Technical Aspects

4.1. Two copies for collection preservation

The digitized data were stored on CD-ROM. There are always two copies of each document namely a) *Archive Masters* and b) *Distribution Masters*. The data quality control is based on regular measurement of physical and digital properties of optical media and on the self-control routine of the storage system.

The *Archive Master* in *Tagged Image File* format (TIFF) is the library's preservation means. It is meant to ensure that the library's digital assets remain valid even if technologies change over time. The archive master must be the original form of the data, prior to any modifications made to the data for distribution to users.

The second main purpose of the Archive Master is ensuring recovery in case of physical damage or hazard.

The Distribution Master is a simple copy in JPEG (*Joint Photographic Experts Group*-standard **ISO 10918-1**; 1994) format of the Archive Master, like an electronic Braille file made available to users in hard copy or in electronic format. Each image serves for production of a set of user images of lower quality that is still good for normal reading and study.

4.2. Formats

Many original manuscripts are in poor condition and sometimes it is necessary to perform activities such as sharpening, tuning the contrast or brightness, and so on. This cannot be done in the normal browser environment. The most difficult problem is usually the oversize of pages whose TIF format image is several times larger than the computer screen that makes the text illegible.

4.2.1. TIFF

The TIFF solution was particularly considered especially when digitization of the manuscripts started. It was an important solution for these manuscripts, because the original pages were usually large, but the dithering onto 1-bit image brought another problem, namely the threshold tuning in case of bad quality original. Digitization of acid-paper materials concentrated on endangered documents whose readability was rather poor in a non-negligible number of cases; therefore, it was preferred to digitize in 256 shades of grey and to let possible reduction of the image depth for some post-processing routines possibly required for fast access on the web. Thus, the degree of the application of TIF format became the basic storage format. When applying the quality factor 12 in Adobe Photoshop, the loss of information was so imperceptible that such images could be qualified as having archival value. Since the beginning we wanted to be compatible with the formats recommended for the World Wide Web, it was decided that the basic data format for manuscripts in true colors or 256 shades of grey would be TIFF.

4.2.2. JPEG

In his work the project team had to face the reluctance of certain people to accept this format for the Distribution Master because of its loss DCT *(Discrete Cosine Transform*) compression algorithm. Having performed many tests it was concluded that the manuscript specialists were rather happy with the images on which relatively tough compression had been applied and which had no high resolution. Thus a sufficiently acceptable JPEG file could have about 1 MB. The thumbnail and preview images were in GIF (*Graphics Interchange Format),* while the other images were in JPEG; the difference between them consisted not only in a looser or tougher compression ratio, but also in a modified resolution. The JPEG compression ratio should always be set up individually and its value depends on the character of the digitized original. For most manuscripts, the user image can be now acceptable and readable as it is on the screen with the resolution of 1024 x 768 dots and higher; therefore no additional tools for its manipulation are needed.

JPEG images and TIFF images of large originals remained rather voluminous for Internet transfer especially in slow speed circumstances, i.e. access from home or small institutions.

The efficiency and reliability of existing emerging solutions were compared from the point of view of their possible application in delivery of digitized library materials. The latter criterion was very important, because small files required extra work of high quality, while, when testing larger files, many problems appeared, such as: high computing for production of new compressed files, inaccurate treatment of bitmap chunks representing individual characters in a few 1-bit compression approaches, slow or even impossible display of new files, especially in the true image domain, unreliable or imperfect viewing of tools in their integration into Internet browsers.

4.3. Text as image

Text may be different from an image. If we take the text as an image, which may be very important for hand-written texts, and make a copy of analogue image reproduction, the above statement is fully true both for the original and the copies and their next generations. There will be no information loss whenever we try to preserve the integrity

of the text. Consequently the quality of copying the text, when we move the text from an information medium onto another one, is controlled by the criterion of its readability.

One of the important preservation criteria can be the stability of the medium on which the (digitally or non-digitally) encoded information has to be written. The classical media such as parchment or even good-quality paper are more stable than the media of digital information. At present, the compact disc is becoming ever more reliable. However, the problem is the standard used for the *data format*. The text domain was the first to feel this problem of a long-time based communication.

4.4. Lith and shadow

The image of any manuscript is from a physical point of view a concentrated and simplified information for our eye. Consequently, it is necessary to take the following facts into account:

- The image of three-dimensional objects in two dimensions is a reduction of primary spatial information about this object;
- The only registered physical property, i.e. the reflection of the original, does not tell much about its other properties;
- The recording of reflectance is also a reduction of primary information, contained in reflected electromagnetic undulation. This undulation, reflected from the original, is described by three digits, proportionate to the energy of reflected light in narrow-band wavelengths (RGB or CMYK);
- If the information about the properties of filters and specificity of reflected light is not stored anyhow, the interrelation between them and the properties of the original is absent;
- The image on the monitor, as well as the copy or the facsimile, creates for our eyes the flow of information, which is similar to the flow from the original under the same conditions as if it was from the recorded copy. Nevertheless, it is modified according to the properties of recording and reproducing equipment.

With regard to different character of illumination (angle, volume) it is possible to achieve various images of the same document. Even the calibration, which we have executed, we cannot eliminate the distortion of the image, caused by the characteristics of digitization technology.

4.5. Description of images

Actually, we had to elaborate a system of access to digitized documents. We elaborated a system which aims to be completely independent of any software and hardware platform. All data, which are added to the image, were recorded in the extended HTML. On the base of known properties of the manuscript the text - file without concrete information, but in the structure, needed for this manuscript was generated. At the same time, to every future image a name was added to which a description of respective page will be connected. For this purpose, the ALEPH programme has been utilized. Then the experts add the concrete information about the manuscript, and further information about individual pages.

Each composition, versions and variants, marginal notes, fragments of translation, disparate lyrics documents, etc., were classified separately, with reference to the Perpessicius edition (Mihai Eminescu, *Opere* vol. I – XVII, Bucureşti, 1939 – 2008). Since many of these compositions are to be found in different manuscripts, the cataloguer mentions all the pages of the manuscripts containing the respective work and cites incipit for each page. Thus, the composition completed, each file of the manuscript has an accurate reference. Each type of document mentioned above is accompanied by links to the manuscript images, organized as in the Perpessicius edition. The images are presented to the virtual user in two resolutions, within which the text can be read as far as the poet's writing is legible.

Efforts were made with the view to persuading other institutions to take part in the digitization programs of Eminescu's manuscripts, to contribute with their records such as The National Archives, The Museum of the Romanian Literature, The Central University Libraries of Cluj-Napoca and Iasi). It is worth noting that the interest to cooperate is larger: and the BAR team is about to reach concrete agreements for a possible national (union) database.

The catalogue plays an important role and it can really serve as a good basis for creation of a good virtual research environment in this domain. The benefits of working with colleagues were immeasurable and the relationships that have been built up will continue. This experience shows that what is needed is enthusiasm, energy, perseverance and the will to talk to associates in the same field.

5. The Facsimile Edition

In 2004, the Romanian Academy began publishing a facsimile manuscripts edition. The entire edition, coordinated by Acad. Eugen Simion, includes, according to the editorial plan (Dumitrescu, 2004), 24 volumes which contain the whole range of manuscripts, correspondence and various documents of biographical interest. Access to the manuscripts is therefore now possible revealing thus the personality of the poet, his creative process, his intellectual preoccupations in ordering text categories (by genre or otherwise) and their chronology. Each volume is provided with a brief note about the content on the title page.

The facsimile edition of excellent graphic quality is due to a team headed by Prof. Mircia Dumitrescu, of the National Academy of Arts of Bucharest, in collaboration with Gabriela Dumitrescu, the head of the Department of Manuscripts at the Romanian Academy Library. The printing is due to RA Monitorul Oficial Publishing House. In the Foreword, Acad. Eugen Simion, the coordinator of the edition, presumes this undertake could be achieved faster than the philosopher Constantin Noica (1909 -1987) had thought, several decades ago, that is five years instead of fifteen.

6. Conclusion

The aim of this paper was to inform about the structure enveloping the digital data and enabling access through WWW browsers or special software. Therefore our future work will consist in the development of a robust method taking into account specific features of such documents and specific solutions adapted to limited categories of manuscripts. Cooperation with other organizations and institutions enhancing the access possibilities through the usage of internet technology and new business models (Filip, 1996a, Filip, 1996b; Filip et al, 2001) is also envisaged.

REFERENCES

1. CERQUIGLINI, B., J.-L. LEBRAVE, **Philectre, An Interdisciplinary Research Project in Electronical Philology**. Lili. Zeitschrift für Literatur-wissenschaft und Linguistik, vol. 27(106-171p.), 1997, pp. 83-93.
2. DUMITRESCU, G., **Editorial Plan. The Eminescu's manuscripts**, Bucharest, Encyclopedica Publishing House, 2004, pp. 5-14.
3. NOICA, C., **Enciclopedia României**, http://enciclopediaromaniei.ro/wiki/Constantin_Noica, consulted on 30.10.2009
4. ExLibris ALEPH Integrated Library System http://www.exlibrisgroup.com/category/Aleph consulted on 30.10.2009.
5. FILIP F. G., **Tehnologiile informatice si valorificarea patrimoniului cultural national**. ACADEMICA Anul V, 9(69), 1996a, pp. 22-24 (in Romanian).
6. FILIP, F.G., **Information Technologies in Cultural Institutions**. Studies in Informatics and Control, vol. 6(4), 1996b, pp. 385-400.
7. FILIP, F. G., D. A. DONCIULESCU, Cr. I. FILIP, **A Cybernetic Model of Computerisation of the Cultural Heritage**. Computer J. of Moldova, vol. 9(2), 2001, pp. 101-112.
8. FRITZ, D. A., R. J. FRITZ, **Mark for Everyone**. ALA editions, Chicago, 2002
9. KNOLL, A., **Digitization of Rare Library Materials: Storage and Access to Data,** Adolf Knoll et al. Prague, National Library: Albertina icome Praha, 1997. 1 CD-ROM (Memoriae Mundi Series Bohemica).
10. MAYER, T., A. KNOLL, **Proposal of the Structure of Digitized Old Books and Manuscripts**, Vers. 1.11. July, 1996.
11. WARING, P., **Edmund Barton Digitisation Project**. http://www.nla.gov.au/nla/staffpaper/2001/waring1.html, consulted on 30. 10. 2009.