IJDAR (2007) 9:193–221
DOI 10.1007/s10032-006-0030-0

ORIGINAL PAPER

DEBORA: Digital AccEss to BOoks of the RenAissance

F. Le Bourgeois · H. Emptoz

Received: 21 February 2005 / Revised: 10 November 2005 / Accepted: 28 May 2006 / Published online: 2 February 2007

Abstract EBORA (Digital AccEss to BOoks of the RenAissance) is a multidisciplinary European project aiming at digitizing and thus making rare sixteenth century books more accessible. End-users, librarians, historians, researchers in book history and computer scientists participated in the development of remote and collaborative access to digitized Renaissance books, necessary because of the reduced accessibility to digital libraries in image mode through the Internet. The size of files for the storage of images, the lack of a standard file format exchange suitable for progressive transmission, and limited querying possibilities currently limit remote access to digital libraries. To improve accessibility, historical documents must be digitized and retro-converted to extract a detailed description of the image contents suited to users' needs. Specialists of the Renaissance have described the metadata generally required by end-users and the ideal functionalities of the digital library. The retro-conversion of historical documents is a complex process that includes image capture, metadata extraction, image storage and indexing, automatic conversion in a reusable electronic form, publication on the Internet, and data compression for faster remote access. The steps of this process cannot be developed independently. DEBORA proposes a global approach to retro-conversion from the digitization to the final functionalities of the digital library centered on users' needs. The retro-conversion process is mainly based on a document image analysis system that simultaneously extracts the metadata and compresses the images. We also propose a file format to describe compressed books as heterogeneous data (images/text/links/ annotation/physical layout and logical structure) suitable for progressive transmission, editing, and annotation. DEBORA is an exploratory project that aims at demonstrating the feasibility of the concepts by developing prototypes tested by end-users.

Keywords Historical document digitization · Document image analysis · Document image compression · File format · Digital libraries

F. Le Bourgeois (✉) · H. Emptoz
LIRIS – I.N.S.A. de LYON – Bât 403, 20 Avenue A. Einstein, 69621 Villeurbanne Cedex, France
e-mail: frank.lebourgeois@liris.cnrs-lyon.fr

1 Presentation

Digital libraries (DL) provide new services such as online consultation of rare documents, improved navigation capabilities, information retrieval, and the ability to share knowledge with other readers. To make this second Gutenberg revolution possible, books that are deemed important to human knowledge need to be digitalized. Numerous research projects on digital libraries are investigating the future tools of libraries for query, retrieval, analysis, management, accessibility, usage, archiving, and preservation of information [1]. The European Community and the French Ministry of Culture support programs to digitize collections and develop networked libraries. A growing portion of digital libraries is accessible in image mode because they cannot be processed automatically by an OCR system or because users generally prefer to read rare books printed with the original typography and layout. Most users need to access documents in their original layout for specific studies investigating not only the text itself but also the appearance, the texture of the paper, the color of the ink, and the contents of the drawings, marks that can identify the manuscripts, the edition, the publication date, the editor, or the authors. It shows how important it is to consider first the users' needs to define the future functionalities of digital libraries. The main bottleneck is the reduced accessibility to digital libraries in image mode through the Internet. The size of files required to store and transmit documents images and the lack of a standard exchange file format suitable for progressive transmission of digitized books and fine querying currently limit remote access to digital libraries. Most of the existing file formats are not suited to a fine description of digitized books of several hundred pages. Digitized documents are represented by heterogeneous data (images/text/links/ annotation/physical layout and logical structure) that cannot be fully managed by current file formats. Users need a common file format suited to document description and representation that can be queried by contents, edited, annotated, exchanged, and progressively transmitted through a network. Such a file format does not yet exist because worldwide consortiums for standards have not completely taken into account libraries' needs.

Most online DL in image mode such as Numdam (http://www.numdam.org), Gallica (http://gallica.bnf.fr), and GDZ (http://gdz.sub.uni-goettingen.de) massively digitized their documents in bilevel and encoded in PDF, or compressed in Tiff or DjVu (which apply a JBIG-like compression for bilevel images). It should be noted that most digital libraries digitized microfilms and then compressed the bilevel images using compression schemes suited to color natural images such as JPEG. We also find people who have claimed to preserve original color images using a safe file format such as PDF. They generally do not realize that color and grayscale images are encapsulated in the PDF file using JPEG lossy compression. These examples show that libraries do not always understand the technology they use and image compression features. It also demonstrates the lack of technical solutions for remote access to digitized books in image mode.

The computer-assisted extraction of metadata is an important issue for the development of digital libraries. The retro-conversion process consists in converting original raw images to a reusable electronic form suitable for user needs, providing efficient functionalities such as contents queries, easy image browsing associated with the metadata, and tools for comparing book contents. These functionalities require a fine description of page contents, impossible without the assistance of image understanding systems.

Standard technologies are inefficient for ancient books, printed with rare fonts and unused typography. These documents are not supported by today's (optical character recognition) OCR or (intelligent character recognition) ICR systems. For example the European METAe project [1] has modified a commercial OCR package to read old books printed in Fraktur fonts widely used in Germany and Austria in eighteenth century. Such developments are not profitable for private companies and need public financial support because of the limitation of the market size. For economical reasons, only Latin alphabets are today efficiently processed by commercial OCR and numerous non-Latin alphabets (Arabic, Hebrew, Greek, ...) are still not supported by automatic reading systems. Moreover the layout segmentation achieved by OCR is efficient for modern documents, but the irregularity of the typographical rules for ancient documents makes difficult the layout segmentation by classical approaches. During the DEBORA project, the users asked for an efficient transcription system which can process a wide variety of old documents written in different languages (Mostly Latin, Greek, Hebrew...) and printed using different typographies. For this reason, we propose a computer assisted transcription (CAT) tool which can help to transcribe all printed documents whatever the typography used and the language or alphabet employed. Our approach is different from the classical full automated OCR approach because it is not based on pattern recognition methodology but it uses the similarity between printed patterns of characters to reduce the effort of a manual transcription. Our approach guaranties a very low error rate which is generally explained only by the characters segmentation in a poor quality image.

Historians also need to access to original images for their work on books authentication. Standard compression technologies to reduce the size of images files developed for natural images are not efficient to compress documents images. Specific compression schemes suited for digitized documents already exist but these pixel-based compression methodologies generally use a very low level interpretation of image contents. We propose an another compression scheme which uses mainly a higher interpretation of the contents of documents images. Our approach consists in decomposing the image into different layers of information, which contain a set of homogeneous objects and adapt the compression scheme to the specificities of each object.

During the project, we also found an another technological bottleneck that limits the spreading of one-line digital libraries in image mode like the lack of electronic file format suited for the browsing and the querying of digitized books. We also propose a standalone file format which represents digitized books and all their metadata.

[1] http://meta-e.uibk.ac.at/

Our main contribution for digital access to books of the renaissance (DEBORA) project consists to develop a general retro-conversion platform which allows simultaneously to extract documents metadata by image analysis, to realize an efficient compression of the images by a precise analysis of their contents, to assist the transcription of text and to manage digitized compressed books and their metadata by using an appropriate electronic file format.

1.1 Renaissance books

Sixteenth-century books interest all historians and book historians, because they were created at a key moment in European history with the invention of printing by Gutenberg. This invention profoundly modified society's relationship to knowledge, and thereby the conditions under which knowledge circulated, thus causing transformations in the places and methods of knowledge production as well as in the legitimization of the elite. The contents of these books provide useful information on the social, religious, and political evolution of the time. The contents of the books also changed during the sixteenth century with the emergence of political, scientific, cultural, recreational, and popular books, progressively replacing religious books. A genuine period of intellectual innovation followed, with the circulation, usually clandestine, of writings bearing religious, philosophical, and political criticism through counterfeited and banned books.

Books from the sixteenth century are characterized by a wide variety of forms and contents, whose authenticity must be verified frequently by specialists. It was the beginning of the rationalization and codification of texts, of typography that provided an improved reading comfort, and a progressive standardization of spelling. For example, in the printed texts of Montaigne, there are eight different ways of spelling the French verb "connaître" (to know): cognoistre, conètre, conoître, etc.

These printed documents are handcrafted and the technical constraints of the past reduced the regularity of book production (variations in spacing and margins, random alignment, etc.). These documents contain many defects due to the manufacturing process and the conditions in which these books were conserved, making computer interpretation and recognition of the characters using an OCR package difficult. But these defects are important to authenticating editions and book manufacturers. The typographical printing, which uses raised moveable types, limits the regularity of document layout. The justification at the ends of lines led printers to insert spaces or make abbreviations or contractions in order to vary the length of words as needed. To adjust all the lines on a page, the printer could vary the line spacing and the margins. The visual markers of the beginnings and endings of the parts of a book were not yet common practice. There could easily be a chapter that started in the middle of a full page of text; a line could begin in the middle of the page, rather than at the left margin. This lack of regularity makes automatic layout analysis difficult (Fig. 1). On the other hand, sixteenth century books are printed using only strokes and line-art graphics, which can be easily segmented, and they never use halftone or dithered images. These specific image features lead us to digitize documents in grayscale, and to decompose the image into the bilevel foreground (strokes, characters, line-art graphics) and the color background (image of the paper).

1.2 DEBORA project framework

The European project DEBORA [2] aims at making rare books from sixteenth century Italy, France and Portugal accessible on the Internet in image mode. This project gives access to a rich and little known heritage in the form of carefully preserved books, frequently existing in only a single copy and rarely accessible to users. The European project DEBORA has introduced a multidisciplinary approach to making Renaissance books accessible to a wider audience. The consortium of the DEBORA partners is a highly multidisciplinary group with multiple and complementary skills:

Libraries Biblioteca Casanatense, Rome (Italy), Biblioteca Geral, Universidade do Coimbra, (Portugal), Bibliothèque Municipale de Lyon (France), which have a well-known collection of rare books from the sixteenth century. Other libraries are participating as test sites, e.g., Stadtbibliothek Kulturrefat Stadtverwaltung, München (Deutschland), Universidad de Zaragoza Biblioteca, (España) and Bibliothèque Municipale et Universitaire de Gèneve, (Switzerland).

Research laboratories Lancaster University (Great Britain), which specializes in collaborative work tools, ERSICO, University of Lyon III (France), which is working in the field of analyzing users' needs and costs analysis, LIRIS at INSA in Lyon (France) and the Instituto Superior Tecnico, Lisbon (Portugal), which are working on image-understanding tools, ENSSIB, National School of Information and Library Sciences, Villeurbanne (France), for user interfaces.

Industrial partners SGBI Entreprise, Lyon (France) for web-based development and Xerox, Grenoble (France) in association with IIS, Bordeaux (France) for digitizers.

The DEBORA project has included studies on users' needs, the libraries' constraints, digitization and retro-

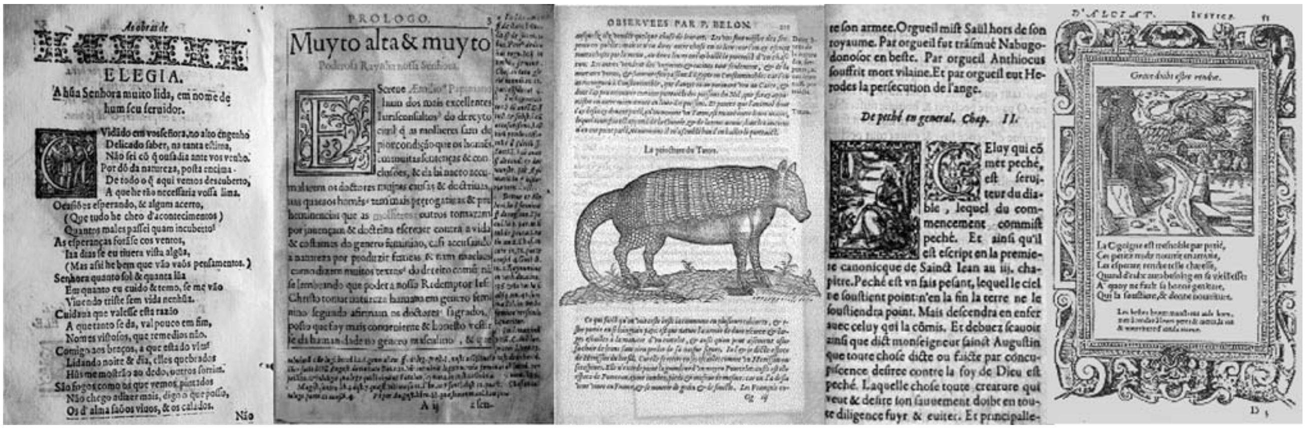


Fig. 1 Samples of images of books from the sixteenth century digitized for DEBORA

conversion costs, the requirements and limits of image understanding, metadata, and the technological bottlenecks for remote access to digital data [3]. We present in the next sections a short summary of several studies which explain our choices in terms of image processing and analysis.

1.3 User needs

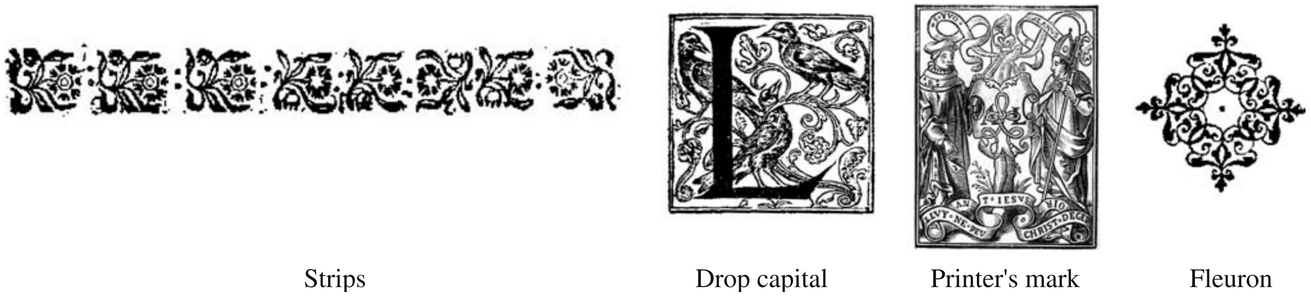
There are two main types of users: the *producers*, who create the corpuses by entering digital books in Renaissance databases (librarians, information specialists, curators), and the *final users* (researchers, professors, students, graphic designers, printers, etc.), who access the digital collection. A study of the users' needs shows that both require greater accessibility to Renaissance documents.

- *An efficient query of the document's contents:* for practical use, a very detailed description of the images' contents is needed. But these descriptions carry such a high indexation cost that it is completely unrealistic to make them available in a digital library. The images' contents can be highly detailed with image-understanding tools or a manual enrichment from a group of specialists working on a collaborative task based on an annotation system. An exchange file format, which supports annotations, is then required for this collaborative work. The file format must support the annotating of any component of the document (image zone, links, element of the physical layout, text contents, metadata extracted by image analysis, or other descriptions from another user) with any metadata (text, multimedia files, internal reference, external links URL, etc.). The users do not perceive digitized documents as sets of images. They consider them as heterogeneous data containing images, text, internal links between the text, images and annotations, the document's logical structure and physical layout. The users' interests are very different: some users work on the typography of the characters and only want to access the typeset and information on typography, others are interested in ornament authentication and want to download only the graphic elements, many of them study the text contents, the character ligatures, the printing quality, the logical structure or simply the document layout for a study of the evolution of typography. Each particular need requires detailed indexing of corresponding components in the document and a file format that can manage the associated metadata.
- *Fast access to high-quality images:* High-resolution lossless compression of the images is required for sixteenth century books to preserve the small details. The users also recommend a fully scalable compression format, which allows fast browsing by downloading only required components. Low-resolution images make reprinting, zooming, or any detailed observation impossible for serious studies.

1.4 Metadata required

Based on the preliminary conclusions of the Manuscript access through standards for electronic records (MASTER) project for access to medieval manuscript collections in image mode, DEBORA's partners defined three levels of metadata of interest to all book historians:

Level 1: book description They are identical to those typically found in library catalogues: (*Author, book title, place of publication, publication date, publisher, language, collation, reference, fingerprint, notes, secondary author, subject, etc.*). Most existing online libraries allow a query on these metadata to retrieve a particular



Strips

Drop capital

Printer's mark

Fleuron

Fig. 2 Example of metadata to retrieve in Renaissance books

book. Some websites allow a common query to access various libraries.

Level 2: book structure This indicates the logical parts of the book using a finer terminology than the concept of chapter (*Title pages, frontispiece, preliminary pieces, text pages, pull-outs, indexes, tables, final pieces*). Due to the high variability of Renaissance books, an automatic recognition of the book's logical structure is a difficult task. For instance, only specialists can achieve this description of book structures, but tools are required to edit and link the structure with the digital images.

Level 3: page contents The list of metadata at this level is long and complex but we give the most important information to retrieve sorted in decreasing order of importance:

Physical layout Main page body (*main text area without footnotes and margin*), marginal notes (*Text written at the bottom of the page or in the margin*), columns, text lines and paragraph, text alignments, typography of characters (*style, size, font*), running head (*title of the whole book or of the chapter printed at the top of each page*), caption title (*title placed in the head margin of the first page of the text of a book*), etc.

Textual information Keywords such as "incipit" (*a term designating the first words of a manuscript or a book*), signature (*letter or sign placed in the bottom margins that indicate the ordering of the pages to the bookbinder*), folio (*numbering of the sheets in a book*).

Decorations Strip (*nonfigurative decorative elements that separate the chapters and paragraphs*), fleurons (*decorative patterns employed to indicate the end of chapters*), illumination (*a small, ornamental letter illustrating the leaves of a book*), frontispiece (*illustration or engraving typically found facing the title page*), illustration (*images, engravings, maps, medallions, etc.*), drop capital (*a capital letter, often ornamental, at the beginning of paragraphs or chapters*), printer's mark, miniature (*ornamental letter colored in minimum red*)(Fig. 2).

The metadata in the pages' contents is very important to identifying, with a certain amount of accuracy, the period when a book was printed and the version of the book. The amount of information to be retrieved at the page level raises a problem in terms of use and cost. It is clear that the required information must be produced for each page. If we wish to attain a certain level of detailed description, estimating that a day's work will be necessary to cover a single book is far from exaggerating. A small collection of 1,000 books would thus take an expert 5 years; this is a completely unrealistic endeavor. Levels 1 and 2 metadata were previously described by researchers. A fine description of level 3 metadata, at the page level, is rarely done and some metadata can be recognized by document image analysis (DIA).

1.5 DEBORA objectives

In this paper we will describe the DEBORA retro-conversion system, which is based on the decomposition and the recognition of image components by a document image analysis module. The results are interpreted to enrich metadata, to improve image compression and to assist manual transcription. Our compression scheme is based on the adaptation of compression algorithms to each component of the image. We also developed a compressed file format adapted to the representation of digitized books that improves querying and navigation. In addition, we introduce the concept of CAT, which reduces manual transcription time by exploiting character pattern redundancies achieved by the compression stage.

The overall scheme of the DEBORA system is described in Fig. 3. The decomposition into segmented components, the physical layout, and metadata extraction are described in Sect. 2. Image compression is presented in Sect. 3, computer-assisted transcription in Sect. 4, and the file format in Sect. 5.

Due to the limited duration of the project, we did not extract all required metadata and we focused our work on compression and file format specification. The file format allows the progressive development of future tools to segment and interpret new components of the images. When a new module of image understanding is

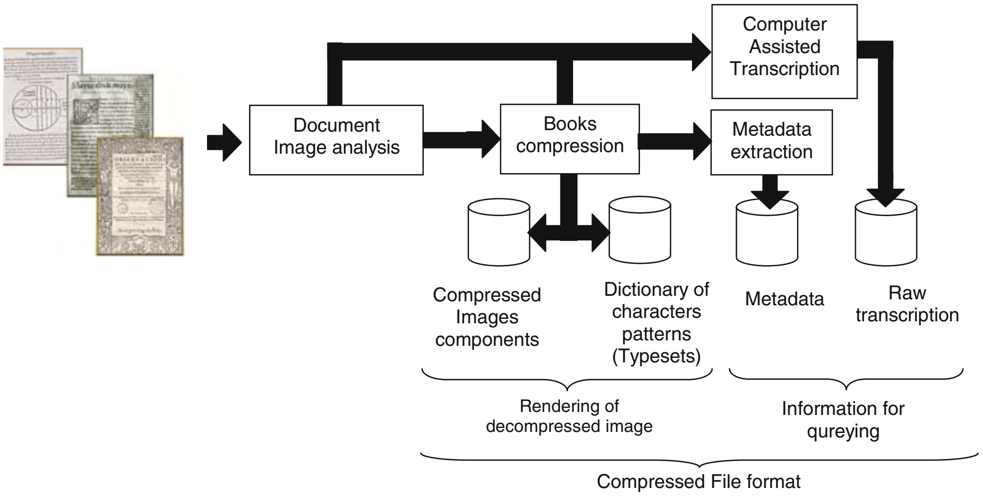

Document Image analysis
Books compression
Computer Assisted Transcription
Metadata extraction
Compressed Images components
Dictionary of characters patterns (Typesets)
Metadata
Raw transcription
Rendering of decompressed image
Information for qureying
Compressed File format

Fig. 3 General overview of the DEBORA scheme

introduced, the results of the analysis are simply added as new metadata in the compressed file.

2 Document image analysis

Document image analysis is the main part of the system. It decomposes images into individual elements that can serve either to improve compression or to enrich the book's metadata. The efficiency of the segmentation has different impacts on the system. Segmentation errors are not critical for the image compression stage but are important for metadata enrichment and critical for computer-assisted transcription. Our main contributions are the design of a multistage segmentation scheme that simultaneously separates text from graphics and adapt the threshold methods to the image content. The image processing required for all needed tasks (text from graphics separation, recto from verso segmentation, text and graphics correct segmentation and denoising) cannot be achieved separately and sequentially. We group all dependent tasks in a general multistage segmentation scheme which optimizes the segmentation efficiency.

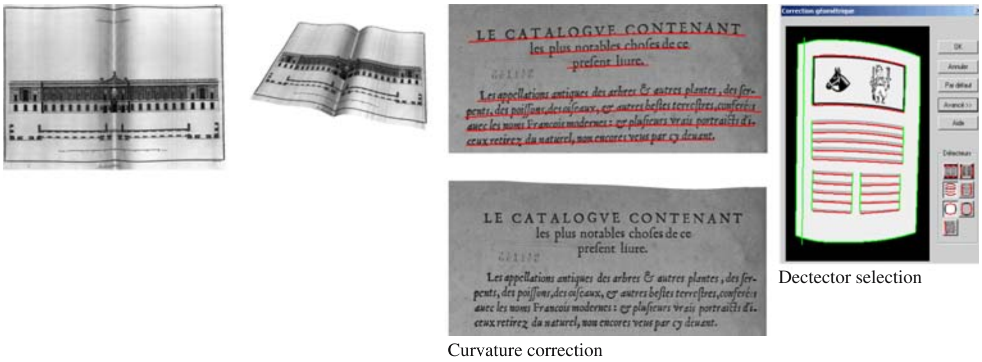
2.1 Image preprocessing

To reduce the impact of the defects on image segmentation, we recommended librarians capture and store original high-resolution images (minimum 300 dpi) in grayscale levels or in true colors, using lossless compression if possible. High-quality images should be digitized directly from original books and not from facsimiles or microfilms. Color or grayscale images, even in low resolution, may be properly segmented using the color depth information. All partners respected these minimal specifications so that image analysis systems could be used today and in the future. All the libraries chose to digitize the books at the grayscale level, because the color is rarely used by books manufacturers in sixteenth Century in order to reduce the cost of printing and only precious books of the Renaissance are printed in color like the famous Gutenberg Bible.

We corrected the skew and the curvature due to the book binding using Bookrestorer software, which was developed in collaboration with our laboratory [4] (Fig. 4).

2.2 Multistage segmentation

We define the foreground the original print matrix, and the background the paper medium. The background/ foreground separation is a critical step of the process, which is rarely done by simple thresholding. In most instances, the verso appears on the recto and stains and defects of the paper are also segmented into the foreground. The direct use of an adaptive thresholding method such as Niblack's or Sauvola's is not accurate with normal parameters for Renaissance books because all the paper's defects and the verso are segmented. The best results for the foreground/background were found by using a multistage algorithm that adapts the parameters and the thresholding method for each particular element of the image (Fig. 5). This approach is best because the thresholding method depends on the image's contents. Our background/foreground separation depends on the segmentation of text and graphic elements. A specific separation between the recto and the verso is achieved only for textual parts, because character segmentation is a very important step for the physical layout and computer-assisted transcription.



Curvature correction

Dectector selection

Fig. 4 Input curved image and the 3D model generated by image analysis

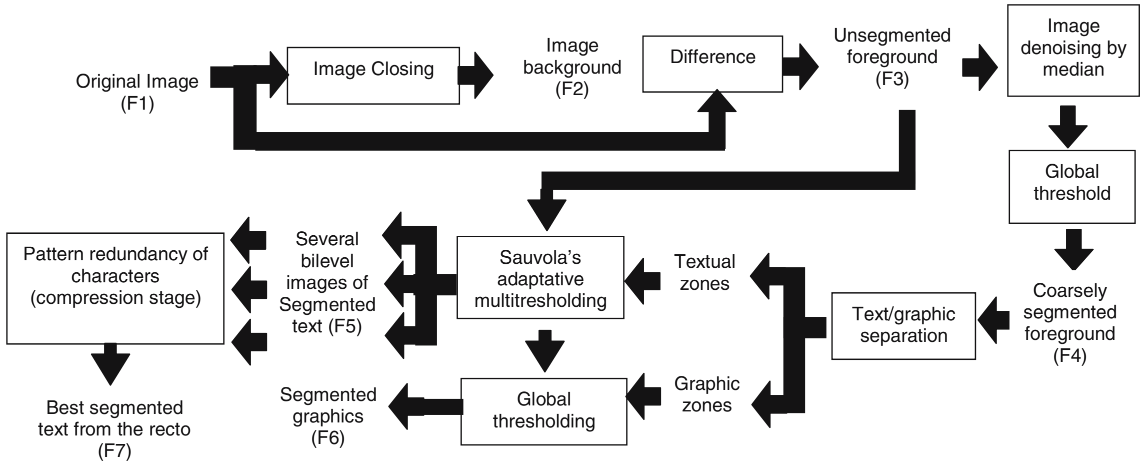

Original Image (F1)
Image Closing
Image background (F2)
Difference
Unsegmented foreground (F3)
Image denoising by median
Global threshold
Coarsely segmented foreground (F4)
Text/graphic separation
Textual zones
Graphic zones
Sauvola's adaptative multitresholding
Global thresholding
Several bilevel images of Segmented text (F5)
Segmented graphics (F6)
Pattern redundancy of characters (compression stage)
Best segmented text from the recto (F7)

Fig. 5 DEBORA's multistage segmentation scheme

In the first step, we apply a morphological top hat transform to the original Image I (*F1*) in order to correct the illumination change near the book binding or for nonplane pages [29]. We first apply a morphological closing of the image with a large structural element B to remove all strokes and printing elements (Fig. 6). The result of the closing (*F2*) is kept for the compression stage to encode the background. Then we differentiate the background and the original image to find the unsegmented foreground (*F3*). The result of this process, also called "top hat transform", is a grayscale image whose intensity measures the relative darkness of the pixels independently of the illumination variation. The unsegmented foreground contains the recto, the verso, paper defects, and noise.

Structurant element $B = \left\{(u,v) \in Z^2/|U| + |v| \leq R\right\}$

Image dilation : $D^B(I) = \text{Max}\{I(x{+}u, y{+}v) \quad \forall(u,v){\in}B\}$

Image erosion : $E^B(I) = \text{Min}\{I(x{+}u, y{+}v) \quad \forall(u,v){\in}B\}$,

Image closing : $C(I) = E^B\left(D^B(I)\right)$

Top hat transform $C(I) - I$

The structural element B uses a radius R equal to the half of the maximum width of the thicker strokes contained in the image. The radius R depends on the size of the original books, the thickness of the font used for the printing and the resolution and scale defined during the digitization process. All the collections of DEBORA has been digitized in 300–400 dpi and processed with the same radius $R = 4$. We choose a 8-connectivity element to define B because it's more suited for the curviness of the Latin alphabet.

The foreground is processed by a median filter with a high radius $\rho = 5$ to remove the noise, smooth the contours, and run together the broken strokes. The first stage uses the Fisher criterion to threshold the foreground coarsely and separate graphical parts from text areas. The text/graphic separation is detailed in Sect. 2.3. We note that the specific parameters k and R of Sauvola's algorithm [5] accurately separate recto and verso. We tried to find a good criterion to automatically select these parameters from the local contrast, and the problem was solved by [6] at the end of the project. For Debora, we selected Sauvola's parameters by using the redundancy rate of character patterns returned by the compression

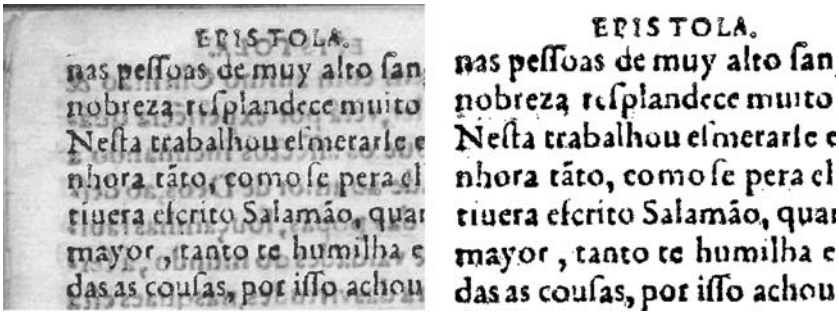
original image I (F1)

Morphological closing (I) = Erode^B (Dilate^B (I)) Background (F2)

Top Hat transform C(I) – I Unsegmented foreground (F3)

Fig. 6 Illumination correction

Fig. 7 Original image (1) Best segmented text from the recto (7)



stage. Our approach is slightly different from the previous work [25] which uses other criterion. The lack of pattern redundancy of characters in a printed text denotes a major problem of text segmentation and in particular in separating recto and verso. We found this criterion very efficient in most cases. The foreground is thresholded by using different parameters of Sauvola's algorithm and the compression stage selects the best segmented image, which has the highest redundancy of character patterns (Fig. 7).

Pattern redundancy analysis is achieved in the compression stage and is detailed in Sect. 3.4. The multiplicity of cascade processing and the numerous feedback loops increase the computation complexity, but it guaranties good results and a constant segmentation quality.

The graphical zones are thresholded again against the unsegmented foreground *(F3)* using a simple Otsu threshold (Fig. 8b). This additional computing is necessary because the median filtering, which removes noise and smooths character patterns for their segmentation, also degrades the drawings (Fig. 8a). The segmentation of the graphical parts such as ornaments, strips, illustrations, and drop capitals has no impact on text interpretation, but the fidelity of the rendering is an important issue for historians who seek to authenticate books. End-user partners of the DEBORA consortium have suggested segmenting the graphical part with a maximum fidelity without any image restoration *(F6)*. Some image defects provide important information for book authentication. This explains why we did not try to remove the noise or separate the verso around graphic parts of the images.

For image compression, we call the segmented foreground, the addition of the segmented graphics *(F6)*, and the segmented text *(F7)*. The temporary images *(F3)*, *(F4)*, and *(F5)* are deleted at the end of the segmentation.

2.3 Separating text from non text

As we do not have a precise model of the layout, we chose a classical bottom-up segmentation (data-driven), which does not require prior knowledge. First we localized all the Connected Components (*CCs*), and then gradually merged them into higher interpreted elements, such as text, drawings, frames, etc. We separated text from graphics using geometrical measures and the regularity of the *CC* placement along a baseline. We noticed that the average size of the connected components from the entire book was a good estimation of the character size and the average height of the main font. We classify *CCs* as text or nontext by comparing the size of each *CC* to the average size of the book's characters. We statistically estimate the average size of characters of all *CCs* segmented on several pages randomly chosen in

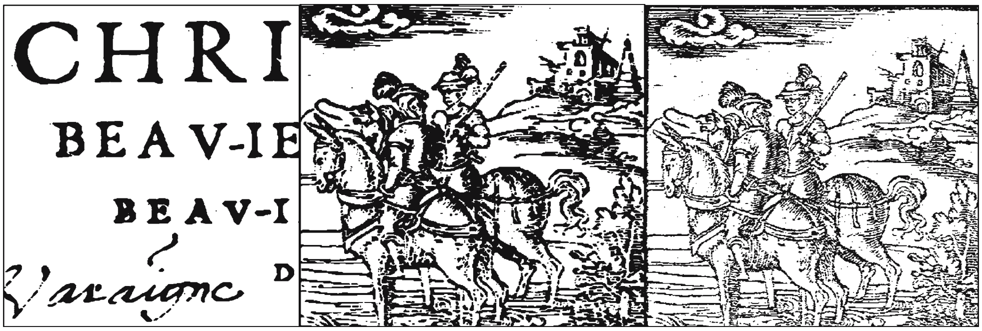


Fig. 8 **a** Coarsely segmented foreground text and graphics (F4). **b** Re-segmented graphics (F6)

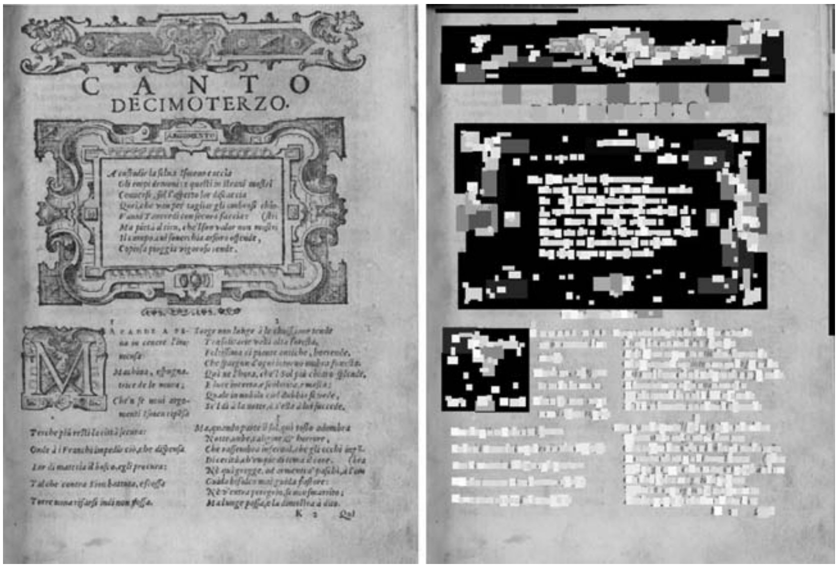


Fig. 9 Classification text/nontext of CCs. **a** Original image, **b** text probabilities $P(x)$ of each *CCx*

the book. Then we compute a text probability *P(x)* for each *CC x* based on the normalized differences between the size of the component and the average text symbol size :

$$P(x) = 1 - \frac{|\text{Size}(x) - \text{Average Size}|}{\underset{\text{For each object } y}{\text{Max}} |\text{Size}(y) - \text{Average size}}$$

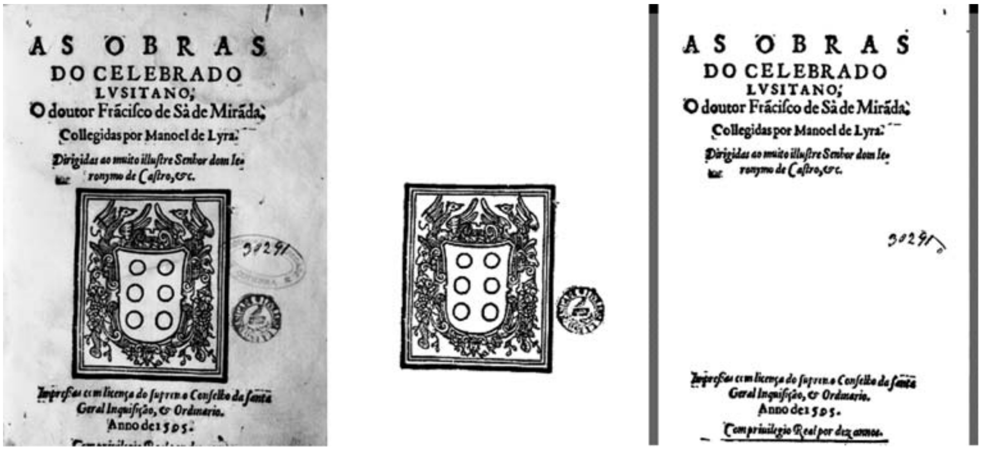
A high-value display in light gray (Fig. 9b) indicates that the component has a high probability of being a text symbol. A simple automatic threshold of the distribution of the *CC* sizes roughly separate text from nontext (Fig. 10).

The *CCs* having probability *P(x)* inferior to the threshold are considered graphics, the others are classified as text. The neighboring *CCs* classified as graphics are progressively merged into larger graphical areas. The graphical elements considered as text lines can be eliminated and false detections rejected by studying the coherence of *CC* alignment. We give two options that the user can set independently for each page of the book. The first option *(alignment constraints)* allows the removal of a false detection based on the *CC* alignment. The second option *(inclusion constraints)* removes all *CCs* classified as text into a larger graphical element.

The *CC* alignment criterion is not always suitable for a particular image. For example, strips are made of separate floral decorations that are perfectly aligned and detected as text lines (Fig. 11). Other images having decorated frames or text inside illustrations cannot be correctly segmented if we apply the inclusion constraint (Fig. 12).

2.4 Segmentation of the main text body and left and right margins

The location of the main text body is a very important issue because the classification of objects may change according to their position in the main document body. For example, a text zone is classified as annotations if it is located outside the main body and regular text if it is inside. The main body may be delimited by a frame, but


AS OBRAS
DO CELEBRADO
LVSITANO;
O doutor Fráciſco de Sà de Miráda;
Collegidas por Manoel de Lyra.
Anno de 1595.

Original image (Los Luisades) *Graphic layer* *Textual layer*

Fig. 10 Separation graphics/text

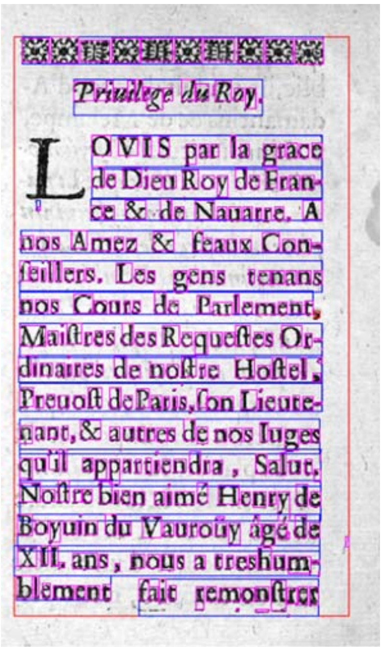

Priuilege du Roy.
LOVIS par la grace de Dieu Roy de France & de Nauarre. A nos Amez & feaux Conſeillers. Les gens tenans nos Cours de Parlement, Maiſtres des Requeſtes Ordinaires de noſtre Hoſtel, Preuoſt de Paris, ſon Lieutenant, & autres de nos Iuges qu'il appartiendra, Salut. Noſtre bien aimé Henry de Boyuin du Vauroüy âgé de XII. ans, nous a treshumblement fait remonſtrer

Fig. 11 Impact of the alignent constraint on the text/graphics separation

the book's borders and figures that might be considered as frames could introduce confusion. We chose to detect the main body by locating the text, even though the text is not always justified, nor does it always fill the entire page.

We sum the text probability *p(x)* values horizontally and vertically along each *CC* to build X–Y profiles for the entire image (Fig. 13). An automatic threshold is computed for each X–Y profile. The text bodies, for two pages of one document, are found by taking the maximum limits of the profile coordinate having values lower than the computed threshold. This approach works most of the time, even on manuscripts [7], but the text bodies can be oversized in the case of large annotation areas in the margin (Fig. 14)

2.5 Physical layout segmentation

The bottom–up segmentation under typographical constraints merges the *CCs* into higher interpreted elements, e.g., columns or paragraphs. Figure15 shows the results of the physical layout segmentation by a bottom–up grouping method, which progressively agglomerates the *CCs* into characters, characters into words, and words into columns and paragraphs. This method requires threshold parameters that define minimal spaces between lines of the same paragraph, words of the same text line, and the minimal and maximal size of each interpreted component (word, line, column, paragraph). These parameters are automatically computed from the sizes of characters printed with different fonts and all the various spaces between components on several pages. In the DEBORA project, we use this aggregative approach to find the physical layout. This approach is not novel; many authors have previously developed layout segmentation based on the analysis of CCs [30]. This bottom-up classical approach is frequently used for document having different layouts and complex structures which are difficult to modelize. Thus, the lack of regularity of old printed documents and the use of rectangular boxes still limit the performance of the segmentation. Other methods continue to be studied to improve the results of document layout segmentation.

The segmentation of the text lines is generally accurate, but some errors occur with touching text lines

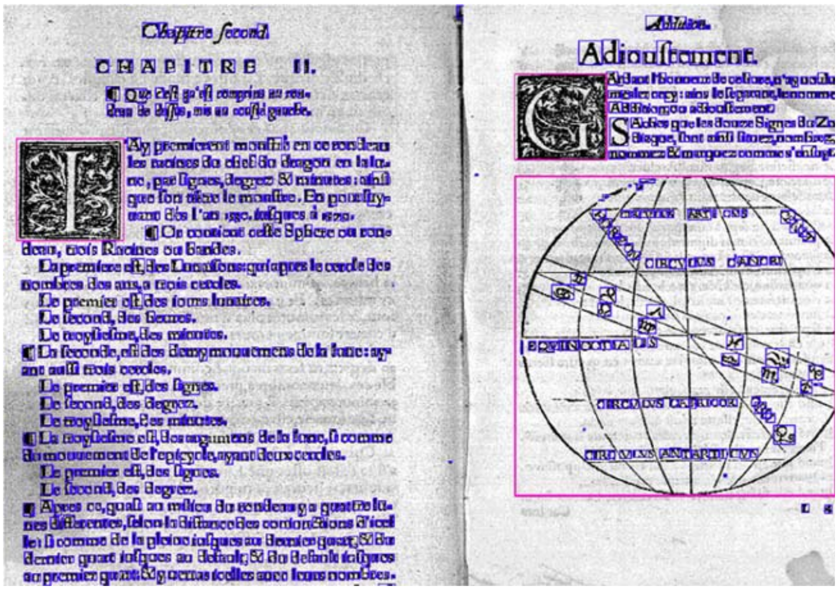


Fig. 12 Impact of the inclusion constraint relief on the text/graphics separation

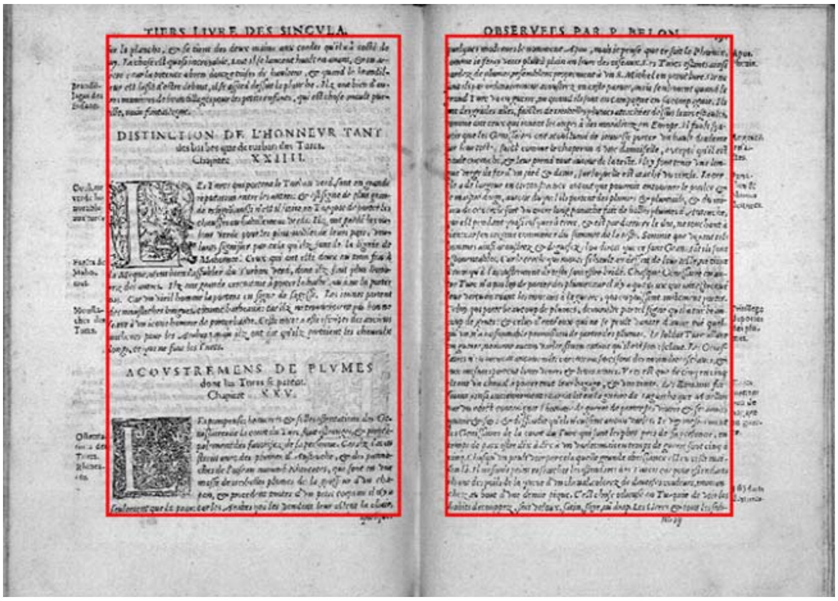
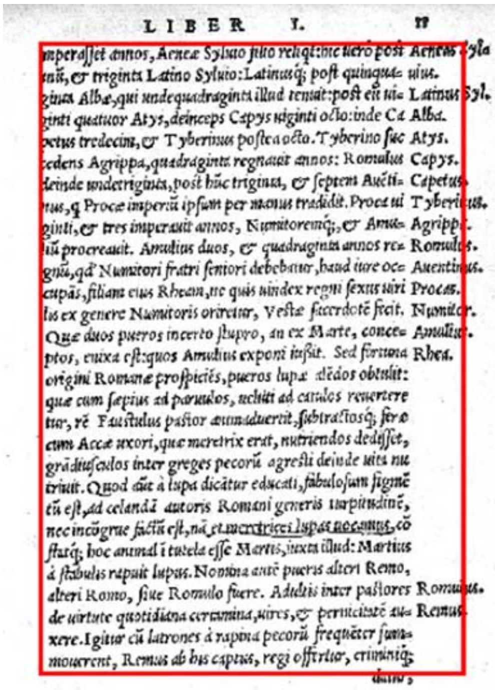


Fig. 13 Main body segmentation using X–Y projections of P(x)

or because of the lack of prior knowledge for the correct segmentation of tables and index pages, which have a more complex layout. The segmentation of the characters is critical for the computer-assisted transcription and compression, which is based on character pattern redundancy. We concentrate on character segmentation and particularly on the over-segmentation of broken characters; ligature characters, for instance, are not processed and remain connected. Character segmentation depends on the quality of the printing and the foreground segmentation. Most of the digitized books of the DEBORA project that have been digitized in high quality generally give very few errors on character segmentation without considering the originally printed ligated characters frequently found in Renaissance books (Fig. 16).

On the contrary, there are many errors in the segmentation of words because of the random shift of the character location along the text line. Many old documents show more spaces between characters than between words (Fig. 17). This major problem makes it difficult to segment words by only taking spaces into account, without any linguistic analysis. This explains why the raw transcription shows many errors in word delimitation. The lack of word segmentation influences text readability more than querying by keywords; text search engines today are very accurate in a raw text without word delimitation.

Finally, the segmentation of columns and paragraphs was not evaluated, because these physical layout elements are not studied by end-users in the prototype



LIBER I. 17

mperaſſet annos, Aeneæ Syluio filio reliq͂t: hic uero poſt Aeneas Sylu-
nũ, & triginta Latino Syluio: Latinusq; poſt quinqua- uius.
ginta Albæ, qui undequadraginta illud tenuit: poſt eũ ui- Latinus Syl.
ginti quatuor Atys, deinceps Capys uiginti octo: inde Ca Alba.
petus tredecim, & Tyberinus poſtea octo. Tyberino ſuc Atys.
cedens Agrippa, quadraginta regnauit annos: Romulus Capys.
deinde undetriginta, poſt hũc triginta, & ſeptem Aucti- Capetus.
nus, q Procæ imperiũ ipſum per manus tradidit. Procæ ui Tyberinus.
ginti, & tres imperauit annos, Numitoremq;, & Amu- Agrippa.
liũ procreauit. Amulius duos, & quadraginta annos re- Romulus.
gnũ, qd Numitori fratri ſeniori debebatur, haud iure oc- Auentinus.
cupās, filiam eius Rheam, ne quis uindex regni ſexus uiri Procas.
lis ex genere Numitoris oriretur, Veſtæ ſacerdotē fecit. Numitor.
Quæ duos pueros incerto ſtupro, an ex Marte, conce- Amulius.
ptos, enixa eſt: quos Amulius exponi iuſſit. Sed fortuna Rhea.
origini Romanæ proſpiciēs, pueros lupæ alēdos obtulit:
quæ cum ſæpius ad paruulos, uelut ad catulos reuertere
tur, rē Fauſtulus paſtor animaduertit, ſubtractosq; fero
cum Accæ uxori, quæ meretrix erat, nutriendos dediſſet,
grādiuſculos inter greges pecorũ agreſti deinde uita nu
triuit. Quod aũt à lupa dicātur educati, fabuloſum figmē
tũ eſt, ad celandā autoris Romani generis turpitudinē,
nec incōgrue factũ eſt, nā et meretrices lupas uocamus, cō
ſtatq; hoc animal i tutela eſſe Martis, iuxta illud: Martius
à ſtabulis rapuit lupus. Nomina autē pueris alteri Remo,
alteri Romo, ſiue Romulo fuere. Adultis inter paſtores Romulus.
de uirtute quotidiana certamina, uires, & pernicitatē au- Remus.
xere. Igitur cũ latrones à rapina pecorũ frequēter ſum-
mouerent, Remus ab his captus, regi offertur, criminiq;
datus,

Fig. 14 Oversized main body segmentation error

evaluation. The text line and character segmentation is evaluated by the users through the raw transcription obtained by the computer-assisted transcription (Sect. 4). For a correct rendering of the transcription, we need a good segmentation of characters and text lines but also a correct ordering of characters along the text lines. The performance of the layout segmentation depends essentially on the correct segmentation of CCs. The main errors reported by users are all explained by the loss of connectivity due to the bad quality of the printing or to the document itself. Errors are found for broken or touched characters, overlapped text lines and words having insufficient spacing. These errors depend entirely of the quality of printing and book preservation.

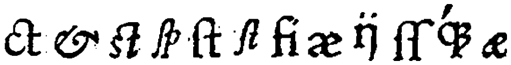
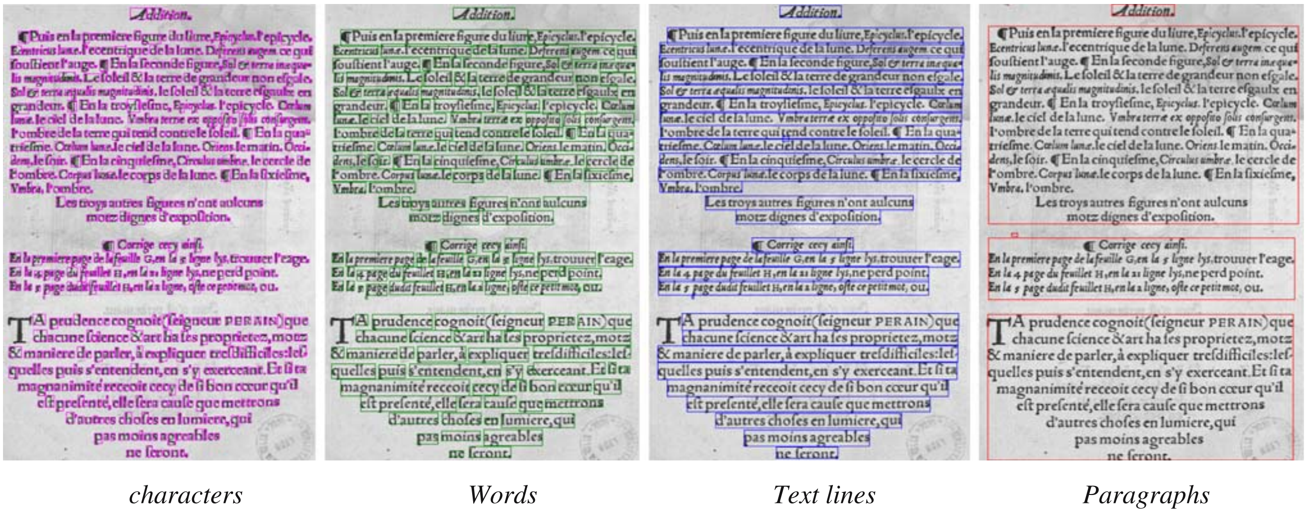


Fig. 16 Set of joined characters as originally printed

2.6 Metadata extraction

We did not focus the work on the recognition of all metadata described previously, as other partners have done this work for DEBORA and we cannot report their results in this paper. We classified graphical elements into three classes: strip, drop capital, illustration. The recognition of drop capitals and strips is based on a fixed model that uses information on location within the page, object size, and justification with surrounding text zones. Strips are large graphical elements generally located at the top of the page that cover the main text body of the page. Drop capitals are square graphical elements that are surrounded by text lines at the right and the bottom. Illustrations are large graphical elements in the main text body that can be surrounded by text lines. Simple decision rules have been developed to show the feasibility of the metadata extraction at the page level (Fig. 18).

The user interface and the file format describe in Sect. 5 makes it possible to browse the metadata of an entire book. A simple query on the drop capital shows a few errors due to the decision rule (Fig. 19). By studying the page corresponding to the false detection, we



Addition.

Puis en la premiere figure du liure, Epicyclus. l'epicycle. Ecentricus lunæ. l'ecentrique de la lune. Deferens augem ce qui ſouſtient l'auge. En la ſeconde figure, Sol & terra inæqualis magnitudinis. Le ſoleil & la terre de grandeur non eſgale. Sol & terra æqualis magnitudinis. le ſoleil & la terre eſgaulx en grandeur. En la troyſieſme, Epicyclus. l'epicycle. Cœlum lunæ. le ciel de la lune. Vmbra terræ ex oppoſito ſolis conſurgens. l'ombre de la terre qui tend contre le ſoleil. En la quatrieſme. Cœlum lunæ. le ciel de la lune. Oriens le matin. Occidens, le ſoir. En la cinquieſme, Circulus umbræ. le cercle de l'ombre. Corpus lunæ. le corps de la lune. En la ſixieſme, Vmbra. l'ombre.

Les troys autres figures n'ont aulcuns
motz dignes d'expoſition.

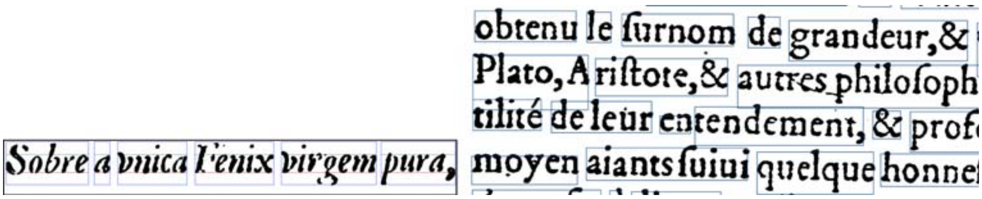
Corrige cecy ainſi.

En la premiere page de la feuille G, en la 5 ligne lys, trouuer l'eage.
En la 4 page du feuillet H, en la 21 ligne lys, ne perd point.
En la 5 page dudit feuillet H, en la 2 ligne, oſte ce petit mot, ou.

TA prudence cognoit (ſeigneur PERAIN) que chacune ſcience & art ha ſes proprietez, motz & maniere de parler, à expliquer treſdifficiles: leſquelles puis s'entendent, en s'y exerceant. Et ſi ta magnanimité receoit cecy de ſi bon cœur qu'il eſt preſenté, elle ſera cauſe que mettrons d'autres choſes en lumiere, qui pas moins agreables ne ſeront.

characters *Words* *Text lines* *Paragraphs*

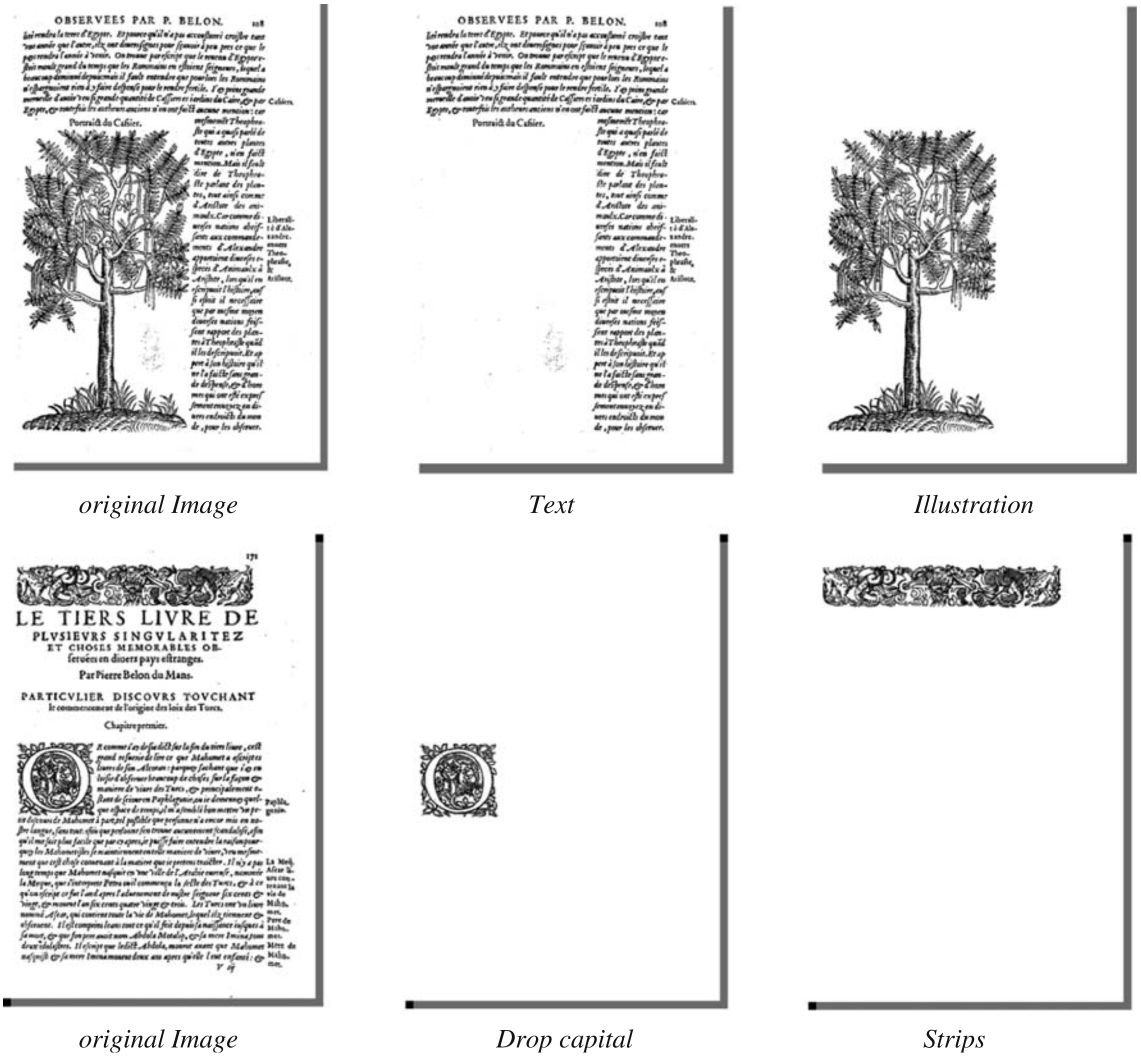
Fig. 15 Results of the segmentation of the physical layout by a bottom–up approach



Sobre a vnica Fenix virgem pura,

obtenu le ſurnom de grandeur,&
Plato,Ariſtote,& autres philoſoph
tilité de leur entendement, & prof
moyen aiants ſuiui quelque honne

Fig. 17 Spacing variability making it difficult to segment words



original Image *Text* *Illustration*

original Image *Drop capital* *Strips*

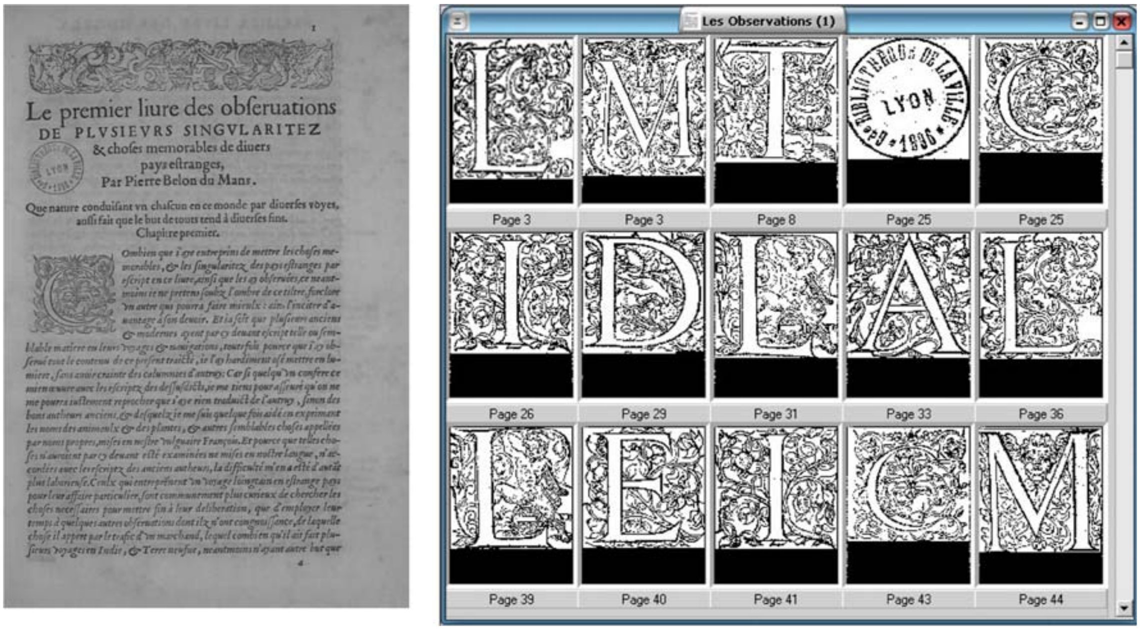
OBSERVEES PAR P. BELON.
Portraict du Caſſier.
LE TIERS LIVRE DE
PLVSIEVRS SINGVLARITEZ
ET CHOSES MEMORABLES OB-
ſeruées en diuers pays eſtranges.
Par Pierre Belon du Mans.
PARTICVLIER DISCOVRS TOVCHANT
le commencement de l'origine des loix des Turcs.
Chapitre premier.

Fig. 18 Classification of graphical elements of the physical layout by using decision rules

note that the segmented stamp has as an aspect ratio corresponding to a square and is also surrounding by text lines. Other particular false detections have been reported because of exceptional cases that can be found in various books from different origins. It shows that a priori fixed decision rules are limited in labeling elements of the physical layout without prior knowledge.

3 Document image compression

The end-users recommended the use of high-resolution images for their studies. The size of images files makes their remote access difficult through a limited bandwidth network such as the Internet. Fast access to a digitized book via the Internet has been one of the main issues



1

Le premier liure des obſeruations
DE PLVSIEVRS SINGVLARITEZ
& choſes memorables de diuers
pays eſtranges,
Par Pierre Belon du Mans.

Que nature conduiſant vn chaſcun en ce monde par diuerſes voyes, auſſi fait que le but de touts tend à diuerſes fins.
Chapitre premier.

Ombien que i'aye entreprins de mettre les choſes memorables, & les ſingularitez des pays eſtranges par eſcript en ce liure, ainſi que les ay obſeruées, ce neantmoins ie ne pretens ſoubz l'ombre de ce tiltre, forclore vn autre qui pourra faire mieulx : ains l'inciter d'auantage à ſon deuoir. Et ia ſoit que pluſieurs anciens & modernes ayent par cy deuant eſcript telle ou ſemblable matiere en leurs voyages & nauigations, toutesfois pource que i'ay obſerué tout le contenu de ce preſent traicté, ie l'ay hardiment oſé mettre en lumiere, ſans auoir crainte des calumnies d'autruy: Car ſi quelqu'vn confere ce mien œuure auec les eſcriptz des deſſuſdictz, ie me tiens pour aſſuré qu'on ne me pourra iuſtement reprocher que i'aye rien traduict de l'autruy, ſinon des bons autheurs anciens, & deſquelz ie me ſuis quelque fois aidé en exprimant les noms des animaulx & des plantes, & autres ſemblables choſes appellées par noms propres, miſes en noſtre vulgaire Françoys. Et pource que telles choſes n'auroient par cy deuant eſté examinées ne miſes en noſtre langue, n'accordées auec les eſcriptz des anciens autheurs, la difficulté m'en a eſté d'autãt plus laborieuſe. Ceulx qui entreprenent vn voyage loingtain en eſtrange pays pour leur affaire particulier, ſont communement plus curieux de chercher les choſes neceſſaires pour mettre fin à leur deliberation, que d'employer leur temps à quelques autres obſeruations dont ilz n'ont congnoiſſance, de laquelle choſe il appert par le trafic d'vn marchand, lequel combien qu'il ait fait pluſieurs voyages en Indie, & Terre neufue, neantmoins n'ayant autre but que

a

Fig. 19 Example of classification error for a drop capital

for the DEBORA project. Image compression concerns only their remote access; compression is not advised for the long-term preservation of images. The original raw images are actually stored in each library in a long-conservation format with no processing or compression.

Only lossy compression with an acceptable perceptible loss of information can reduce the amount of data required to transmit the documents images. The end-users had to decide between accurate image rendering, how much information loss they were prepared to accept, and Internet access speed.

The originality of this work resides in

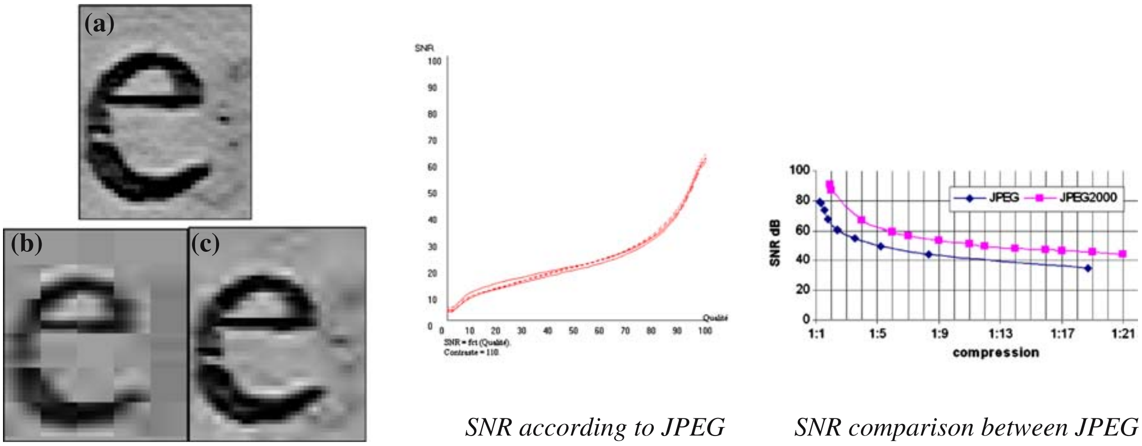
- The use of document image analysis, which decomposes images into separate objects that are compressed differently by the most suitable compression method.
- A pattern-based compression, which performs better than pixel-based compression methods.
- A compression at the book level based on the redundancy of characters patterns from the entire book, which is more efficient than an image compression of each page individually.
- A lossless compression of the residuals (difference between the original and compressed image) based on the property of the discrete pattern matching and substitution algorithm and the digitization theory.

The compression prototype developed for DEBORA is compatible with the Joint Bi-level Image Experts Group(JBIG) recommendations and can be modified to generate JBIG2 files. However, for DEBORA we implemented the compressed files into our file format described in Sect. 5.

3.1 JPEG compression

The JPEG standard produced by the Joint Photographic Expert Group is a very effective compression scheme suitable for color and grayscale images. It provides a high compression rate, varying between 5:1 and 500:1 depending on the quality factor set by the user. The rendering quality of the decompressed image is very high for natural scenic pictures but it becomes low on particular images such as images of documents, which essentially contain black lines over a uniform background. JPEG is not the best compression scheme for images of documents because they show numerous grayscale transitions around the contours of shapes and they are made of lines and strokes that build complex patterns such as characters and drawings. For these particular images, the high-frequency filtering during the discrete cosine transform (DCT) quantification distorts transitions between black ink and the white paper and finally randomly destroys the contours of the characters' shapes. JPEG is not suitable for document image compression, image readability, and image processing.

To measure the effects of JPEG compression on document images, we measured the signal-to-noise ratio (SNR) according to the compression quality factor on different images of printed documents from the sixteenth century (Fig. 20). We can see that JPEG modifies



SNR according to JPEG quality factor on real printed documents.

SNR comparison between JPEG and JPEG2K according to the compression ratio.

Fig. 20 Impact of JPEG compression on document images. **a** Uncompressed image, **b** JPEG 1:12, **c** JPEG2000 compressed 1:12

the image information even with 100% quality factor, disproving the belief that JPEG compression can achieve a lossless compression by using the highest quality factor. The compression ratio is not a linear curve; the main compression is achieved for a quality JPEG factor between 100 and 70%. With a lower quality factor, the compression is not more efficient. It is impossible to use a quality factor under 70% for document images without getting visible artifacts. With an average quality parameter used by default, the loss of information is still high and the 1:12 compression rate is not sufficient to solve the problem of remote accessibility.

JPEG2000 compression based on wavelets shows better rendering than the Discrete Cosine Transform, for the same 1:12 compression ratio. JPEG2000 better preserves the complexity of shapes, but this compression is not more efficient in reducing the size of the images, with a maximum compression ratio of 1:20 with acceptable perceptible deformation. This study definitively proves that generic compression schemes are not suitable for the specificity of document images in terms of quality and compression rate.

New compression schemes such as DjVu or Tiff-FX developed by Xerox separate images into different layers that are compressed layer by layer with an adapted algorithm. These new compression schemes, suitable for digitized documents, reach compression rates higher than 1:100 (compare to 1:10 on average for JPEG and JPEG2000).

3.2 Document image compression survey

Most digital libraries use binary images, which can describe most of the textual and graphical information from various documents. Bi-level images have been the focus of a great deal of research for fax technology, which requires efficient compression algorithms to speed up document transmission. The International Telegraph and Telephone Consultative Committee (CCITT) was founded for that purpose in the early 1980s. This committee proposed lossless compression algorithms where information is preserved using minimum space by encoding frequent information with a shorter code [8]. Encoding has been improved by using prior knowledge on frequent pixel sequences commonly found in document images [9]. The compression algorithms developed by this committee, called CCITT G3 and G4, are both incorporated in TIFF format. The Joint Bi-level Image Experts Group (JBIG), founded in 1993, produced a new bi-level image compression standard called JBIG based on graphical information redundancies that outperformed the CCITT G4 standard [10]. JBIG provides a lossy compression of binary images, which use template models and adaptive arithmetic coding to encode predictions [11]. The JBIG compression scheme is mainly a pixel-based compression, which is not originally suited for progressive transmission [12]. To make JBIG scalable, the original image is divided into layers computed at different resolutions. JBIG loses fidelity when a small-scale structure is essential, as in the case of text documents. Moreover, it is impossible to process the compressed bit stream directly because of the direct use of the arithmetic coding on pixels. Soft pattern matching [13], which uses the prototype image from a dictionary as a template to ensure more accurate coding, remains a pixel-based approach. However, the JBIG1 compression scheme was not widely applied because of the lack of a standard format. The improved JBIG2 compression has a better chance of being widely used because it is going to be adopted as an ISO standard [14]. Moreover, JBIG2 offers new capabilities by processing halftone images and regions of line art, figures, or graphs. Today, several companies are experimenting with transforming the JBIG2 compression scheme into usable data formats, e.g., Xerox Corporation with XIFF (extended TIFF) similar to TIFF-FX proposed by Scansoft and AT&T with DjVu [15]. These companies implemented variation in the original JBIG standard in order to process color documents. The compression is achieved by separating the colored background (paper texture, images, colored frames) from the foreground (ink strokes, characters, line art, graphics), and each layer is compressed differently with an appropriate algorithm. By assuming that the foreground contains redundant patterns, it is compressed using the JBIG scheme. The color background generally contains no textual information and can be heavily compressed with a generic compression algorithm suited for color images, e.g., JPEG or JPEG2k. This approach provides a high compression rate and preserves image readability and quality for automatic interpretation as well.

Pattern-based compression is a better alternative for document images than generic pixel-based approaches because it is well-suited to the repetition of similar patterns of characters frequently found in any printed document. Also called the pattern matching and substitution method (PMS), this approach is a dictionary-based compression (also called token-based compression), extended to two-dimensional images. It stores all different binary patterns in a dictionary and transmits only the dictionary entry for each occurrence of a similar pattern image. This method, introduced by Asher and Nagy in the 1970s for document image compression [16] and developed by IBM [17,18], was not used to its full potential until Witten's work [19,20]. The main difficulty of any pattern-based compression is to define the best similarity measure between the two-dimensional patterns and the growth of the dictionary size. A pattern-matching and substitution approach is essentially a lossy compression scheme because most of patterns from the image are substituted with an inexact prototype found in the dictionary. Witten describes a lossless compression based on pattern matching by encoding residuals left after inexact replacement by the prototype and significant differences between the prototype pattern from the dictionary and the original symbol within the image. The difficulty of directly accessing any components for information indexing and the lack of intermediate representation are the main drawbacks of Witten's work [12]. Nevertheless, compression based on pattern redundancy must be applied to a layer, which contains very frequent shapes like characters.

3.3 Compression by pattern matching and substitution

There are several problems that limit the usage of the pattern matching and substitution approach to compressing bilevel text images:

- **Image comparison**
  Some variation in character patterns can occur because of the random shift placement of the sampling grid during the digitization process. This noise, randomly located along shape contours, is amplified by the irregularity of printing in early Renaissance printed books. Using exact matching, the number of different patterns increases in the dictionary. To get a better compression rate, the comparison of images can tolerate large pattern deformation and each pattern class can represent the different letters that appear visually almost similar, e.g., 1, l, I, j, t. In this case, the size of the dictionary is reduced and the compression rate is high. The rendering error due to character substitution is corrected by the compensation mechanism described in the next section. AT & T and Xerox have chosen to use an automatic clustering in a feature space or the direct Hausdorff[2] distance between patterns, respectively. These approaches give frequent substitution errors, which are corrected by the compensation during the image decompression and rendering.
- **The reduced size of the patterns dictionary**
  The size of the pattern dictionary can increase indefinitely. To limit the size of the dictionary, they generally reinitialize the dictionary for each new image. In our case, each image from a printed book has similar characters and fonts. We must use the redundancy of patterns in the entire book to improve the compression rate and make it possible to use this information for the computer-assisted transcription described in Sect. 4.
- **Compensation**
  Compensation corrects the decompressed image by adding the differences (residuals) to the original image. The image of the residuals contains pixel runs randomly located around contours of the substituted patterns (Fig. 21). The residuals are more complex to compress than the direct compression of the original image. The encoding of the residuals has been

[2] The Hausdorff distancemeasures the radius of the biggest difference between two patterns

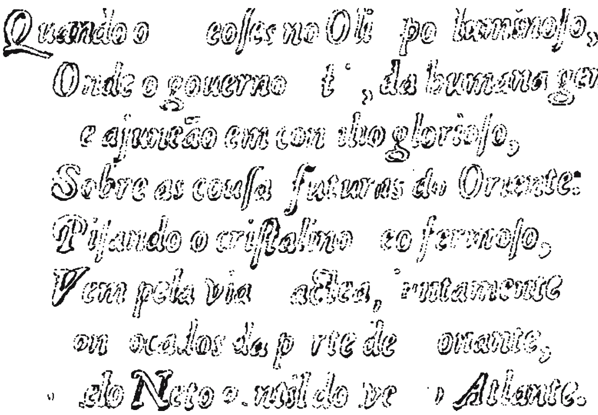


Fig. 21 Maps of residuals

the main issue of the compression by pattern redundancy. Solutions have been found with soft pattern matching and substitution [11] and the run-length coding along the character contour [12].

3.4 Computation of character pattern redundancy

We define the redundancy rate as the ratio between the number of different shapes of characters N_p found in the pattern dictionary and the number of characters N_c found in the text.

$$\text{Character redundancy rate} = 100 \times \left(1 - \frac{N_\mathrm{p}}{N_\mathrm{c}}\right),$$

$$\textit{with} \begin{cases} N_\mathrm{p} = \textit{Number of different characters patterns} \\ N_\mathrm{c} = \textit{Number of characters} \end{cases}$$

The redundancy rate changes according to the quality of character printing. For poor document quality, noise can distort character shapes and artificially increase the number of different character patterns. Broken characters or merged characters also increase the dictionary size and reduce the compression rate. Character distortions, due to the geometric correction of the book curvature or for the image deskewing also affect the compression rate. Another specific problem due to old document production in the sixteenth century involves character patterns, potentially completely different from one folio to another and the redundancy rate can drop when changing folios in a book. This can only be explained by modifications during the book's production or by changes in workshops or printing materials. In the sixteen century it was common to use several workshops with workers of different skill levels to print different parts of a single book. For these reasons, the study of pattern redundancy of characters provides important information about the production of a book for its authentication. In spite of these problems, the redundancy rate of character patterns is high enough for efficient compression.

We used a pattern matching and substitution (PMS) algorithm to compress text areas without substitution errors for later use of pattern redundancy. The pattern comparison must be precise and tolerant to noise. The pattern matching algorithm described in [21] has been slightly modified to match distorted character patterns by increasing tolerance to image noise.

According to the digitization theory [22,23], two similar digitized patterns are randomly shifted by ±1 pixel and the direct difference of these digitized patterns shows a random noise of ±1 pixel along the contour. To make the pattern comparison robust to noise, we must ignore the pixels of the difference that are connected to the contour of the patterns. The erosion of the symmetric difference $A \Delta B = (A - -B) \cup (B - A)$ between two patterns A and B is not the solution, because it indifferently removes all pixels from the contour of the difference $A \Delta B$. To remove only the pixels from the differences that are connected to the pattern contours, we define the following similarity measure S between two patterns A and B:

$$\begin{aligned} S(A,B) = {} & \mathrm{Area}((A - B) - \mathit{Dilate}(B)) \\ & + \mathit{Area}((B - A) - \mathrm{Dilate}(A)) \end{aligned}$$

The morphological dilation uses a 3×3 square structural element to remove pixels from the contour using 8-connectivity. This similarity measure preserves the differences between similar patterns such as ('u', 'n') ('h', 'b') and ('c', 'e') and ignores the random noise around character contours.

To avoid the problem of random pattern placement due to the random phase of the digitization grid, we computed the best similarity measure between A and B for all horizontal and vertical translation by ±1 pixel, producing nine measures. The proposed similarity measure is more precise than the Hausdorff distance, which actually measures the maximal thickness of the symmetric difference.

In order to increase comparison speed, we organized partial tests in a decision tree (Fig. 22). Let *fg* and *fd* be the binary shape of two characters. We define $\Delta g = fd - fg$ and $\Delta d = fg - fs$ as the nonsymmetric difference between shapes *fg* and *fd*. We store the template $\Delta\, g$ and $\Delta\, d$ in each node. For each unknown pattern X, we compared $X \cap \Delta\, g)$ and $(X \cap \Delta d)$ to find the appropriate direction in the decision tree. The pattern Y found in the leaf of the tree is compared to the input pattern X. We automatically updated the tree

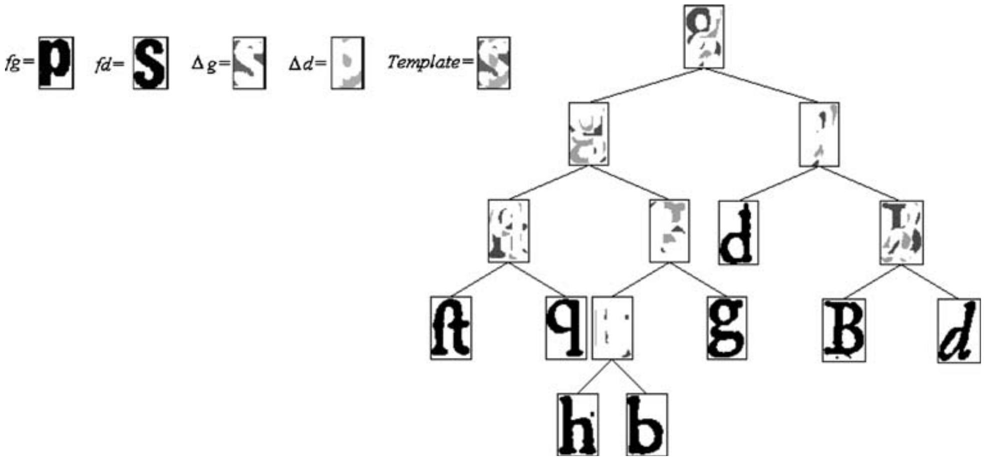
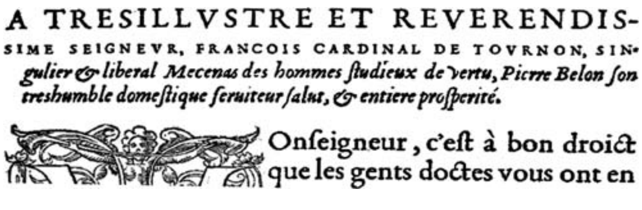


Fig. 22 Decision tree for the efficient organization of tests to find the similar pattern

node if the new input pattern X had a significant difference $S(X, Y) > \varepsilon$ with the pattern Y. We replace the leaf by a node containing the differences $\Delta g = X - Y$ and $\Delta d = Y - X$ and store two final leaves containing in the left the pattern X and the right the pattern Y. We define several decision trees according to different character sizes found in the document.

The decision trees grow rapidly at the beginning of the document. The introduction of new character patterns decreases progressively after several readings of documents from a single source. On a precise example, from text printed in several typefaces (Fig. 23), we observe a progressive rise in the redundancy rate according to the number of pages processed (Table 1).

The redundancy rate started around 80% for the first pages and reached an average of 98% for the entire book depending on the quality of the printing. This supports our choice of creating a dictionary of character patterns for the entire book in spite of the lack of regularity in its production and the variable quality of the printing.


A TRESILLVSTRE ET REVERENDISSIME SEIGNEVR, FRANCOIS CARDINAL DE TOVRNON, SINgulier & liberal Mecenas des hommes ſtudieux de vertu, Pierre Belon ſon treshumble domeſtique ſeruiteur ſalut, & entiere proſperité.

Onſeigneur, c'eſt à bon droict que les gents doctes vous ont en

Fig. 23 Sample of the test pages using several fonts

3.5 Compression of residuals for compensation

The end-users required a lossless compression of the textual layer to preserve details and get the exact reproduction of the original text. As soft pattern matching and substitution [11] does not completely guarantee the preservation of details around character shapes, we developed our own compression of the residuals. Our lossless compression of the residuals cannot outperform any lossy compression proposed in the literature. For DEBORA, we simply wished to reasonably decrease the size of the residuals to make the compression of the textual layer more efficient, without loss of information.

We calculated a compensation layer (Fig. 24) with the symmetric difference (XOR) between the decompressed textual layer (Fig. 25) and the original textual layer (Fig. 26). The compensation layer details small differences between the original and the decompressed

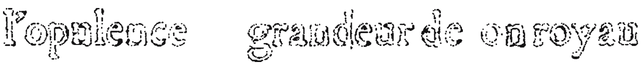

l'opnlence grandeur de on royau

Fig. 24 Compensation layer (residuals)

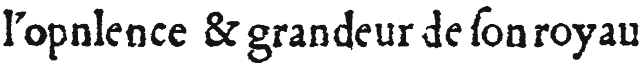
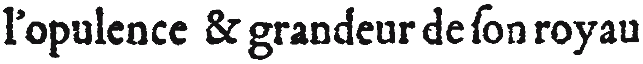

l'opnlence & grandeur de ſon royau

Fig. 25 Decompressed textual layer

Table 1 Compression rate and substitution errors for first pages

Number of pages	1	2	3	4	5	6	7	8
Cumulated number of prototypes	555	915	1,209	1,485	1,678	1,870	2,083	2,262
Cumulated number of characters	2,327	4,245	6,681	8,681	11,159	13,589	16,141	18,028
Pattern redundancy rate(%)	76	78	81	82	84	86	87	88
Cumulated number of substitutions	6	6	6	7	8	8	11	12
Cumulated substitution rate (%)	0.25	0.14	0.09	0.08	0.071	0.058	0.068	0.066



l'opulence & grandeur de ſon royau

Fig. 26 Original textual layer

image along character contours or for a substitution error (the letter ' *u*' from the word ' *opulence*'). To reconstruct the exact copy of the original image, we only need to make an XOR between the residuals and the decompressed textual layer. Consequently, substitution errors during the matching do not affect the final rendering of the decompressed image. But the substitution error during the pattern matching is critical for the future usage of the pattern redundancy.

To compress the residuals efficiently, we made several observations (Fig 27a) that support our compression method:

Consequence of the digitization theory:

A1: All residual pixels are always connected to the contour using 8-connectivity.
Consequences of the symmetric difference to compute the residuals:

A2: Runs of residuals along the contour frequently have a similar distribution.

A3: Residuals are exclusively inside or outside the reference pattern.

By using observation A3, we can entirely define the distribution of residuals along the contour with a unique code K with values in [0..7]. For each pixel of the contour, we find the code K corresponding to the position of residuals by using the table Fig. 27c. We chose *K* code with seven different values because our similarity measure *S* and its threshold ε never provide residuals far from three pixels from the contour in 300 dpi. We tested a larger window size of 11 to encode residuals with a distance of 5 from the contour without changing the compression results. For images in higher resolution, the threshold ε and the size of the window must be changed.

The Code *K* exploits the correlation between successive residual runs along the contour according to observations A1 and A2 (Fig. 27b). We encode the successive configuration of residuals along the contour using a window of size 7 in the orthogonal direction from the pattern. The sequence of codes $K_1 K_2, \ldots, K_n$ is highly correlated and efficiently compressed by an arithmetic coding with a context limited to the two previous values. The compressed chain codes that represent the residuals are stored for each character independently to render a part of the image on demand.

The compensation layer compressed without loss of information requires between 20 and 60 kb of additional data. The dictionary of character patterns is also lossless compressed using CCITT-G4 and is indexed according to character position on each page in order to transmit only patterns that are necessary to display a required page. The compressed textual layer is reduced to the bounding box co-ordinates of each character and the associated link to the character pattern in the dictionary. This data structure makes it possible to partially decompress an image around a region of interest selected by the user.

3.6 DEBORA's compression scheme

We introduced a specific compression based both on an interpretation of information contained in the image by the DIA stage and the redundancy of character patterns in the entire book. The originality of our compression scheme compare to DjVu compression scheme, Kia works and JBIG specification resides in:

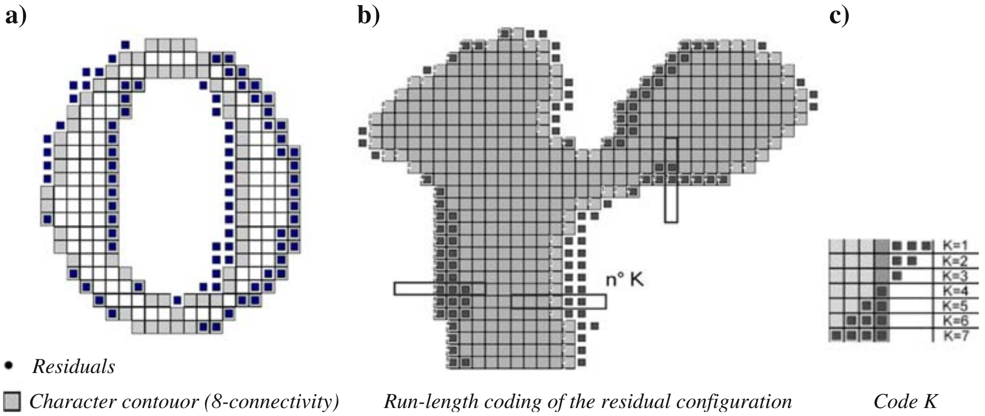

a)
b)
c)
n° K
K=1
K=2
K=3
K=4
K=5
K=6
K=7
● *Residuals*
□ *Character contouor (8-connectivity)*
Run-length coding of the residual configuration
Code K

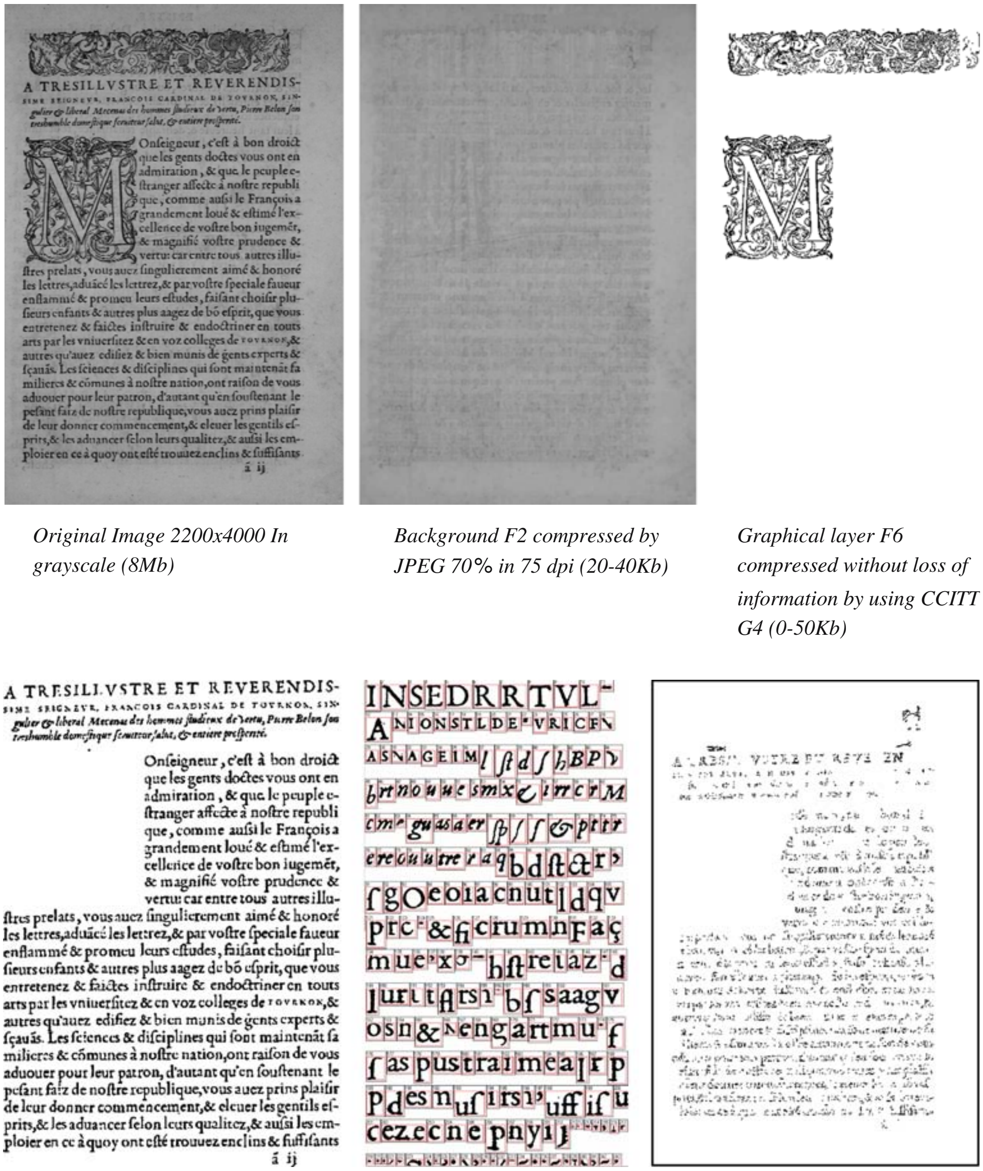
Fig. 27 Placement of the residuals along the contour and residuals encoded using code K

- Adaptation of compression methods to image contents by exploiting the DIA results. Other compression methods use more basic segmentation algorithms which are not specific to the documents. The use of specific segmentation algorithms suited for printed books of the Renaissance allow a better performance in the image compression. In the other hand, our approach cannot be applied to other documents images like old manuscripts or modern printed documents without important modification of the segmentation algorithms.
- Precise pattern matching, which reduces the substitution errors by taking into account the random noise around shape contours described in the digitization theory. This is the main difference with DjVu and Kia approaches. The precision of the characters patterns comparison allow the development of a Computer Assisted Transcription described in Sect. 4.
- Lossless compression of the residuals based on the observation of residual placement along the pattern contours. This approach is completely different from DjVu and JBIG compression methods which mainly use the loss of information of the residuals to increase the compression performance of the textual layer.
- Pattern redundancy throughout the book provided by the unlimited dictionary. This feature increases the performance of our compression scheme. In the other hand, we loose in the simplicity of the page coding and we must encode separately the patterns dictionary of the book and the compressed stream for individual pages. The characters patterns from the dictionary which are necessary for the rendering of a single image of a page must be transmitted to the user.

These features improve the compression rate and provide users with new services such as assisted transcription or document authentication by studying printing regularity. The compression of an entire book makes better use of pattern character redundancies than the compression of each page independently. The DIA stage decomposes the images into homogeneous components that can be compressed more efficiently by the appropriate compression method. We separate the document image into four different layers (Fig. 28):

- *Background layer* (grayscale or color image): This layer represents the printing medium without characters or drawings. This layer is the exact copy in low resolution of the image *(F2)* obtained after the separation of background and foreground as described in Sect. 2.2. This grayscale or color image contains no useful information and is highly compressed with a heavy loss of information using JPEG 70% quality applied at 75 dpi low resolution. The amount of storage required for the background layer is on average between 20 and 40 kb. The efficiency of the JPEG compression for this layer is explained by the homogeneity of the background, which contains no complex shape and no transition of colors that produce high frequencies.
- *Graphical layer* (bilevel image + 1 color/object): This layer contains all the graphical elements that are not characters and have no pattern redundancies (drop capitals, markings, ornamentation, engravings, etc.). This layer is created by the image (*F6*) after the text is separated from the graphics as described in Sect. 2.2. We compress this layer without loss of information using a CCITT-G4 compression scheme. The size of this layer depends on the number of elements and the complexity of shapes. The compressed size of this layer does not generally exceed 50 kb for rare pages having several ornaments or drop capitals.
- *Textual layer* (bilevel image + color of each character): This layer contains only characters, which have the highest rate of pattern redundancy. This layer is a copy of the image (*F7*) produced by the segmentation stage described in Sect. 2.2. It is compressed using pattern matching and a substitution scheme, as described in Sect. 3.4, with an unlimited dictionary for the entire book. The layer only contains a list of coordinates and a number entry in the dictionary of character patterns. To render the textual layer, the character patterns from the dictionary must be stored in the file or transmitted through the network.
- *Compensation layer or residuals* (bilevel image): This layer contains the bitmap difference between the decompressed textual layer and the original textual layer (*F7*). These residuals allow the reconstruction of the exact original foreground and correct small differences stemming from the substitution by the prototype found in the dictionary. The compensation layer makes our compression of the foreground lossless. It is compressed by run-length encoding of code K along the contour of characters as described in Sect. 3.5.

Possible segmentation errors during the text/graphics classification do not modify the final rendering because the textual layer is added to the graphical layer and both are losslessly compressed.



A TRESILLVSTRE ET REVERENDISSIME SEIGNEVR, FRANCOIS CARDINAL DE TOVRNON, SINgulier & liberal Mecenas des hommes ſtudieux de vertu, Pierre Belon ſon treshumble domeſtique ſeruiteur, ſalut, & entiere proſperité.

Onſeigneur, c'eſt à bon droict que les gents doctes vous ont en admiration, & que le peuple eſtranger affecte à noſtre republique, comme auſsi le François a grandement loué & eſtimé l'excellence de voſtre bon iugemēt, & magnifié voſtre prudence & vertu: car entre tous autres illuſtres prelats, vous auez ſingulierement aimé & honoré les lettres, aduācé les lettrez, & par voſtre ſpeciale faueur enflammé & promeu leurs eſtudes, faiſant choiſir pluſieurs enfants & autres plus aagez de bō eſprit, que vous entretenez & faictes inſtruire & endoctriner en touts arts par les vniuerſitez & en voz colleges de TOVRNON, & autres qu'auez edifiez & bien munis de gents experts & ſçauās. Les ſciences & diſciplines qui ſont maintenāt ſamilieres & cōmunes à noſtre nation, ont raiſon de vous aduouer pour leur patron, d'autant qu'en ſouſtenant le peſant faiz de noſtre republique, vous auez prins plaiſir de leur donner commencement, & eleuer les gentils eſprits, & les aduancer ſelon leurs qualitez, & auſsi les emploier en ce à quoy ont eſté trouuez enclins & ſuffiſants

ã ij

Original Image 2200x4000 In grayscale (8Mb)

Background F2 compressed by JPEG 70% in 75 dpi (20-40Kb)

Graphical layer F6 compressed without loss of information by using CCITT G4 (0-50Kb)

Textual layer F7 compressed by pattern matching and substitution (<10kb)

Dictionary of patterns lossless compressed by CCITT G4 (243 different character patterns)

Residuals compressed by run-length contour encoding (20-60Kb)

Fig. 28 Different layers and the associated compression method

Figure 29 shows a global view of the compression scheme used for DEBORA. We give a greater importance to information, which improves navigation capabilities, than to the compression itself. This information concerns the hierarchical structures of the book, the typography, character position and redundancies, drawing locations, as well as information on the physical layout and logical structure. Our goal is not only compression but also the transmission of useful information to satisfy most users' needs.

3.7 Compression results

We statistically analyzed the contents of six compressed Renaissance books (Table 2). For each book, we provide the number of pages, the size of images before and after compression, the amount of data required to encode the pattern dictionary for the entire book and the number of patterns in the dictionary. The loss of topology or connectivity of character patterns decreases the performance of the compression. Nevertheless, the

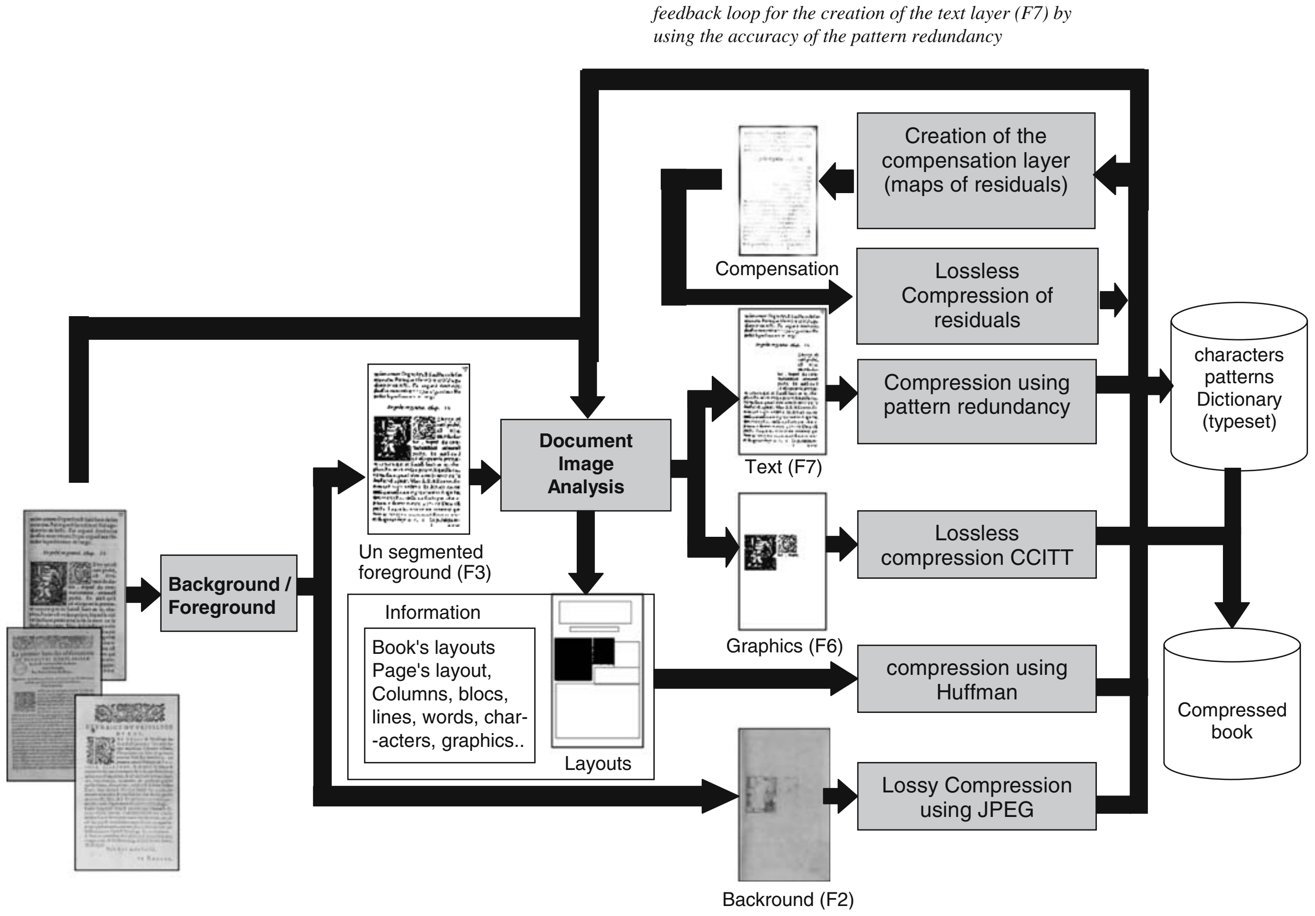

feedback loop for the creation of the text layer (F7) by using the accuracy of the pattern redundancy
Creation of the compensation layer (maps of residuals)
Compensation
Lossless Compression of residuals
characters patterns Dictionary (typeset)
Compression using pattern redundancy
Text (F7)
Document Image Analysis
Un segmented foreground (F3)
Background / Foreground
Lossless compression CCITT
Graphics (F6)
Information
Book's layouts Page's layout, Columns, blocs, lines, words, char- -acters, graphics..
Layouts
compression using Huffman
Compressed book
Lossy Compression using JPEG
Backround (F2)

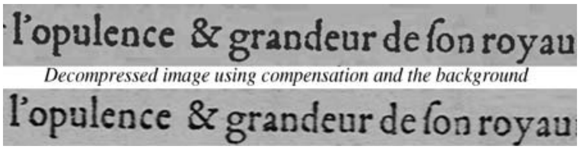
Fig. 29 Overview of the compression scheme designed for DEBORA

Table 2 Detailed results of the DEBORA compression

Book reference	Number of pages	Size of uncompressed images (A) (kb)	Size of compressed data (B) (kb)	Size of the pattern dictionary (kb)	Number of patterns in the dictionary	Compression ratio (A):(B)	Size of the book files including tags (mb)
BML373177	98	1,224,412	18,625	592	5,127	65:1	20
BML341145	452	3,679,280	68,914	1,440	13,458	53:1	79
BML355882	93	1,034,112	15,456	495	4,588	66:1	17
COI-RB-23-6	173	1,508,673	27,507	1,580	12,939	54:1	32
COI-RB-32-4	373	1,015,200	21,226	1,183	11,176	47:1	30
Côte 15992	363	770,633	10,537	2,025	17,565	73:1	14

compression rate for images of books from the sixteenth century is an average of 60:1, which corresponds to 50–160 kb of required storage for a single page in 300–400 dpi with 20 kb of additional information on the physical layout in a tagged file, as described in Sect. 5. This performance makes it possible to store 20 books on a single CD-ROM instead of six CD-ROMS for one book without compression. The file format described in Sect. 5 allows the progressive transmission on demand. There are four different modes of display for the final users (M1: text without compensation, M2: graphics, M3: text with the compensation, M4: colored background). The user can select each mode independently and can browse rapidly only text or graphics without color information and compensation. The textual layer is rendered very quickly, especially when the patterns from the dictionary have already been downloaded and stored in temporary files. Some delay is necessary to retrieve the compensation layer and render the precise textual layer of the foreground.

The reconstructed image compared to the original image (Fig. 30) appears a bit synthetic because of the



l'opulence & grandeur de ſon royau

Decompressed image using compensation and the background

l'opulence & grandeur de ſon royau

Fig. 30 Original image

juxtaposition of the different layers and the separation between the foreground and the background. But the information of the foreground is preserved in the compressed image.

Table 3 gives the compression rate compute in average for all the six test books with default parameters for DjVu and Jpeg. The compression rate obtained by our method outperforms JPEG compression. The comparison with DjVu is not easy because the DjVu compression rate starts from 1:17 in lossless mode and reaches 1:120 using the aggressive mode, which compresses with visible loss of information (Fig. 31). Using the regular lossy mode, DjVu gives a better compression rate. We obtain a compression rate of 1:80, equivalent to DjVu in the most aggressive mode, by compressing the residuals with a loss of information. We achieve this performance by averaging the chain codes $K_1, \ldots, K_n$ to increase the correlation. The rendering of the averaged chain codes smooths character contours in the decompressed image. This experiment shows that the main problem is the compression of the residuals that occupy a large space in the compressed bitstream. Using a lossy compression of the residuals or image smoothing of the text, we considerably augment the compression rate. But the partners of the DEBORA consortium did not accept any loss of information in the textual layer since it was important for studying the book. Compared to DjVu, the main drawback of the DEBORA compression is the lack of generality. Because our compression scheme is based on document image analysis, it cannot work properly for unexpected images from documents other than Renaissance books. The methods used are specific to printed books of the Renaissance, but the DIA employed is sufficiently robust to process a wide range of printed books from XVI to nowadays. But unfortunately, we must greatly modify the DIA stage in order to process other printed materials like newspapers or dictionaries because of their specific layout which uses multi-columns.

The evaluation of compression scheme must be done visually by users. Objective figures for the evaluation of the compression do not make sense because we can tune the parameters of the different compression scheme (Quality factor for Jpeg, compression factor for DjVu) to obtain several compression rates and different visual results.

4 Computer-assisted transcription

Renaissance books are printed with rare fonts and generally unused typography, which are not supported by today's OCR systems. Character patterns, document layouts, language, and vocabulary are totally different. The Renaissance was a period of high creativity in typography and document layout. A great number of fonts were created to reproduce the diversity of the scripts found in medieval manuscripts. Moreover, the fonts of the Renaissance show many characters that no longer exist today (Fig. 32).

This explains why OCR packages suitable for modern printed documents cannot be used for old printed documents, even those in Latin. The development of an OCR for the Renaissance is not within the limits of the DEBORA project and we have proposed an alternative called CAT for computer-assisted transcription. CAT uses information on the character pattern redundancies computed during the compression process to increase the speed of manual transcription. Manual labeling of characters from the pattern dictionary allows for the reconstruction of the entire text in only a few hours. The number of different character patterns to edit manually depends on the regularity of the printing. Usually, the number of different character patterns represents 2% of all the characters of the book, corresponding to an average of 10,000 patterns from the dictionary. The

Table 3 Average compression rate comparison computed for the six tested books

Compression scheme	JPEG (Q = 75%)	DjVu lossless	DjVu lossy	DEBORA
Compression rate	1:11	1:17	1:98	1:59

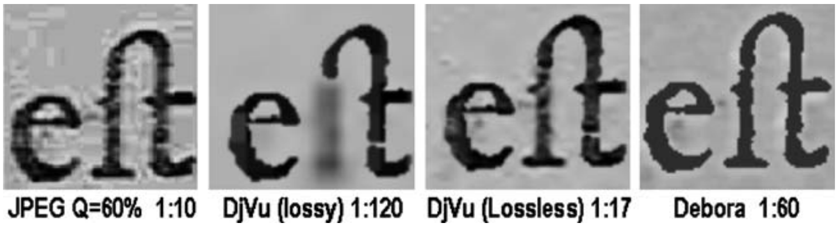
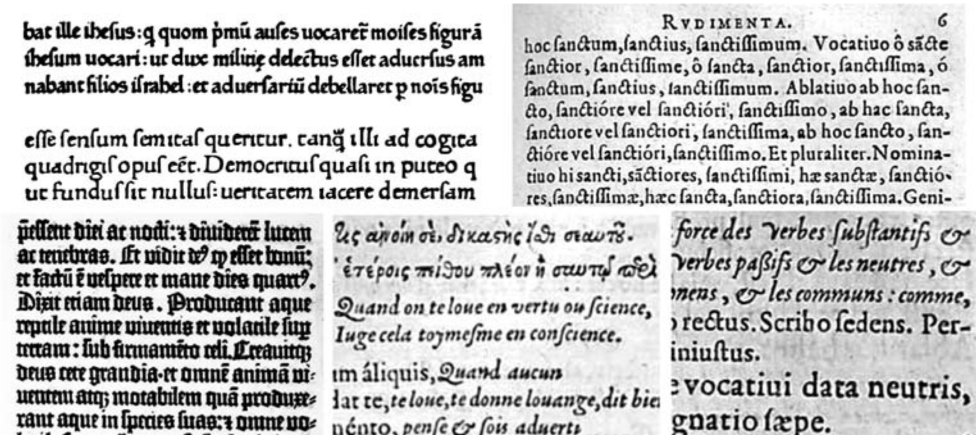


Fig. 31 Visual comparison between different compression scheme for a sample of text



bat ille ihesus: q quom pmū auſes uocaret moiſes figurā
ihesum uocari: ut dux militię delectus eſſet aduerſus am
nabant filios iſrahel: et aduerſariū debellaret p nois figu

eſſe ſenſum ſemitaſ queritur. tanq̃ illi ad cogita
quadrigiſ opuſ eēt. Democrituſ quaſi in puteo q
ut funduſ ſit nulluſ: ueritatem iacere demerſam

RVDIMENTA. 6
hoc ſanctum, ſanctius, ſanctiſſimum. Vocatiuo ô sāčte
ſanctior, ſanctiſſime, ô ſancta, ſanctior, ſanctiſſima, ó
ſanctum, ſanctius, ſanctiſſimum. Ablatiuo ab hoc ſan-
cto, ſanctióre vel ſanctióri, ſanctiſſimo, ab hac ſancta,
ſanctiore vel ſanctiori, ſanctiſſima, ab hoc ſancto, ſan-
ctióre vel ſanctióri, ſanctiſſimo. Et pluraliter. Nomina-
tiuo hi ſancti, sāctiores, ſanctiſſimi, hæ sanctæ, ſanctió-
res, ſanctiſſimæ, hæc ſancta, ſanctiora, ſanctiſſima. Geni-

pſſent dici ac nocti: & diuiderēt lucem
ac tenebras. Et vidit de⁹ φ eſſet bonū;
et factū ē veſpere et mane dies quart⁹.
Dixit etiam deus. Producant aque
reptile anime viuentis et volatile ſup
terram: ſub firmamēto celi. Creauitq;
deus cete grandia. et omnē animā vi-
uentem atq; motabilem quā produxe-
rant aque in ſpecies ſuas: & omne vo-

ὡς αἰνοίη σε, δικαστὴς ἴσθι σεαυτοῦ.
ἑτέροις πείθου πλέον ἢ σεαυτῷ περὶ

Quand on te loue en vertu ou ſcience,
Iuge cela toymeſme en conſcience.

im áliquis, Quand aucun
dat te, te loue, te donne louange, dit bie
nénto, penſe & ſois aduerti

force des verbes ſubſtantifs &
verbes paßifs & les neutres, &
mens, & les communs: comme,
rectus. Scribo ſedens. Per-
iniuſtus.
vocatiui data neutris,
gnatio ſæpe.

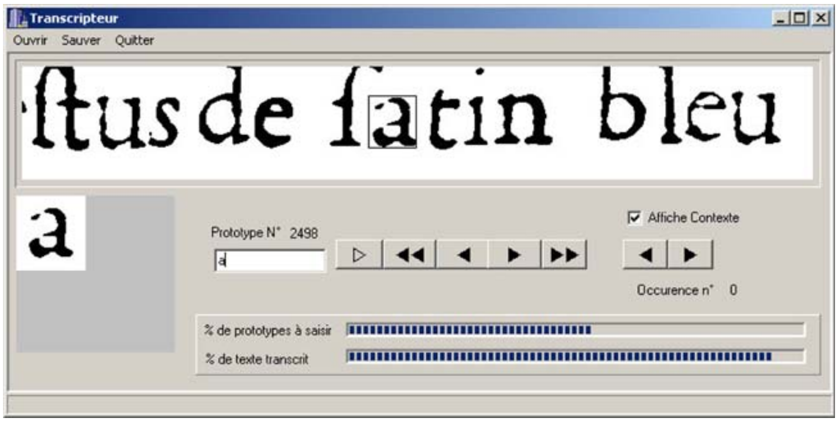
Fig. 32 Diversity of fonts and obsolete characters

complete transcription of a book can be achieved in 6 h for a book of 200 pages with 2,000 characters per page by manual labeling of all the patterns from the dictionary. These figures have been coarsely evaluated by observing the transcription of several books by different users. Assisted transcription generally increases the transcription speed by 98%, which is the average redundancy rate measured on Renaissance books. It means that only 2% of all characters must be manually entered to obtain the transcription of the entire book. This approach is possible because we have used a precise pattern comparison for the entire book, which does not tolerate substitution errors. Links between full-text words and their location in the image are also stored in the compressed file. Moreover, 50% of the pattern dictionary concerns rare character patterns that occur only once in the entire book. They generally represent degraded characters or rarely ligated or broken characters. If we manually transcribe the 50% of the pattern dictionary corresponding to the first 5,000 patterns from the dictionary, which describe the most frequent character patterns, we obtain a correct transcription of 80% of the entire book in 3 h. This rate is sufficient for a query with a search text engine that accepts missing characters. The encoding of obsolete characters from the Renaissance is not solved because there is still no standardized Unicode description for these characters. Therefore we use substitution letters to code these special characters with a loss in readability of the transcription. Moreover, the lack of precise segmentation of words influences the transcription readability but not the query by keywords. We have observed the users behavior during the transcription by using the CAT system, and we have designed the user interface to simplify the manual constraints. Because users cannot recognize an isolated character pattern without the context, we display all text lines which contain the current character pattern (Fig 33).

A number unexpected applications were found during the DEBORA project. For example, we noticed that our compression scheme based on character pattern redundancies is sensitive to imperceptible variation of the typesets used by the typographer. The variations in the character pattern redundancy rate provide interesting information on printing regularity and book manufacturing. It describes the successive corrections of the books and provides information to compare different editions of the same book. A future project will use our character pattern comparison algorithm to study the different typesets used in France and authenticate the different publisher and book origins during the Renaissance.

5 Versatile format suited for digitized books

The DEBORA project is an exploratory project that has attempted to develop new ideas to improve accessibility to digitized books. In this context, we studied all existing formats that can manage digitized books as heterogeneous data (structures, text, images, links, annotation) containing a high density of interconnected information. The file format must also allow querying and editing any element of the book content. Image formats such as TIFF, XIFF, PNG, and JPEG2K, etc., printing formats such as PDF and PS, and edition formats such as RTF/Word, LaTeX, SGML/TEI, XML are not suitable for simultaneously managing images, text, annotations, links, and structures. On the other hand, XML METS is an appropriate file format that can describe documents in both image and text mode as well as book structure. But this file format has several drawbacks. The high density of information in the description of a book at the word level provides very large XML description files of several hundred megabytes, without the images. The repetition of tags for each element and



Prototype N° 2498

a

☑ Affiche Contexte

Occurence n° 0

Fig. 33 Software interface for the assisted transcription

the high number of hierarchy levels gives description files that can be larger than the images. The parsing of large XML description files makes rapid browsing and querying difficult. Moreover, XML-METS manages books by separate files: the description files and image files. Retaining the coherence between image descriptions and image files is difficult. A standalone file format that embeds image and metadata is easier to maintain. The encoding and transmission of digitized documents are important issues for the development of digital libraries, which open new research fields for the DIA community.

For the DEBORA project, we developed a tagged file format, similar to XML METS, but encoded in a binary file, to reduce storage and increase management speed [24]. Like RTF/Word, our format uses a highly compressed standalone file. A single file can contain only a few elements, several images of pages, or several books. Accessing the documents, editing the document's physical layout and logical structure, managing images and their descriptions, and editing annotations are all improved. The file format can be decomposed for progressive transmission and display for remote consulting. Figure 34 gives an example of a book's description using

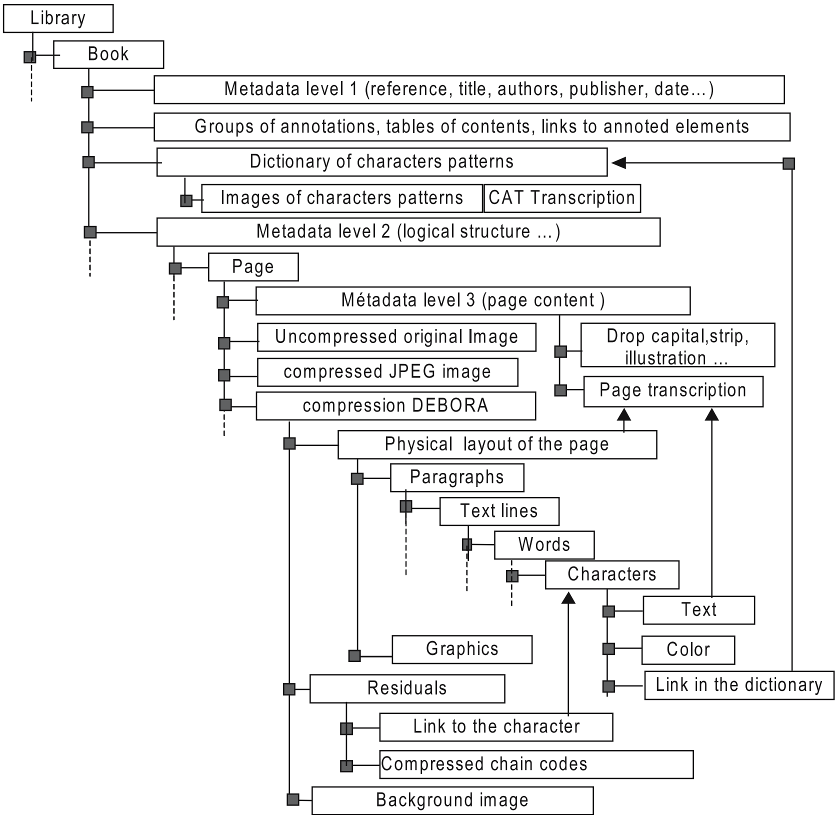

Library
Book
Metadata level 1 (reference, title, authors, publisher, date…)
Groups of annotations, tables of contents, links to annoted elements
Dictionary of characters patterns
Images of characters patterns
CAT Transcription
Metadata level 2 (logical structure …)
Page
Métadata level 3 (page content)
Uncompressed original Image
compressed JPEG image
compression DEBORA
Drop capital,strip, illustration …
Page transcription
Physical layout of the page
Paragraphs
Text lines
Words
Characters
Text
Color
Link in the dictionary
Graphics
Residuals
Link to the character
Compressed chain codes
Background image

Fig. 34 Partial representation of the book's structure in DEBORA format

our tagged file. Note that the format can manage both compressed and uncompressed images. The same page can be represented by several different images stored in different image formats. A page can be represented by different image qualities using different compression schemes. When images are directly stored using the JPEG, TIFF, or PNG format, the information on the layout and the transcription is empty. After the DEBORA compression, a new [compression Debora] field is created and the metadata at the page level is automatically completed. Only images that are compressed by our compression scheme allow the query of metadata at the page level.

The pattern dictionary is located on the book level; when we merge pages from different books, the required patterns from the different dictionaries are automatically added to build a new dictionary. During computer-assisted transcription, the users manually enter the transcription for each pattern of the dictionary in the [CAT transcription] field. This field is automatically duplicated in the [Text] field of corresponding characters. When a substitution error occurs, the correction is reported only at the character level in the [text] field of the character. The transcription uses the information on the physical layout to reconstruct text lines and the [text] field of characters. The [page transcription] field is automatically created to avoid the reconstruction of the transcription for each query.

The proposed image format was designed to satisfy most users' needs and has following features:

Decomposability The file format defines all needed containers for all information handled by the system. A tag codes each container that describes a single component of the document and its hierarchical property. The decomposition of books into independent elementary components guarantees full scalability of the format. A tag can describe objects located at various scales. For example, any physical element can be accessed with the following hierarchy (set of books►book►set of pages ►page►layers►text ►column ►paragraph ►line►word ►character ►prototype of pattern). If a full-text transcription exists, we can also access a textual component in both image mode and full-text mode. The logical structure is manually provided for some books and incorporated into the file format. Users can download books and work on them without being connected to the server. All functions described below remain accessible when users are offline.

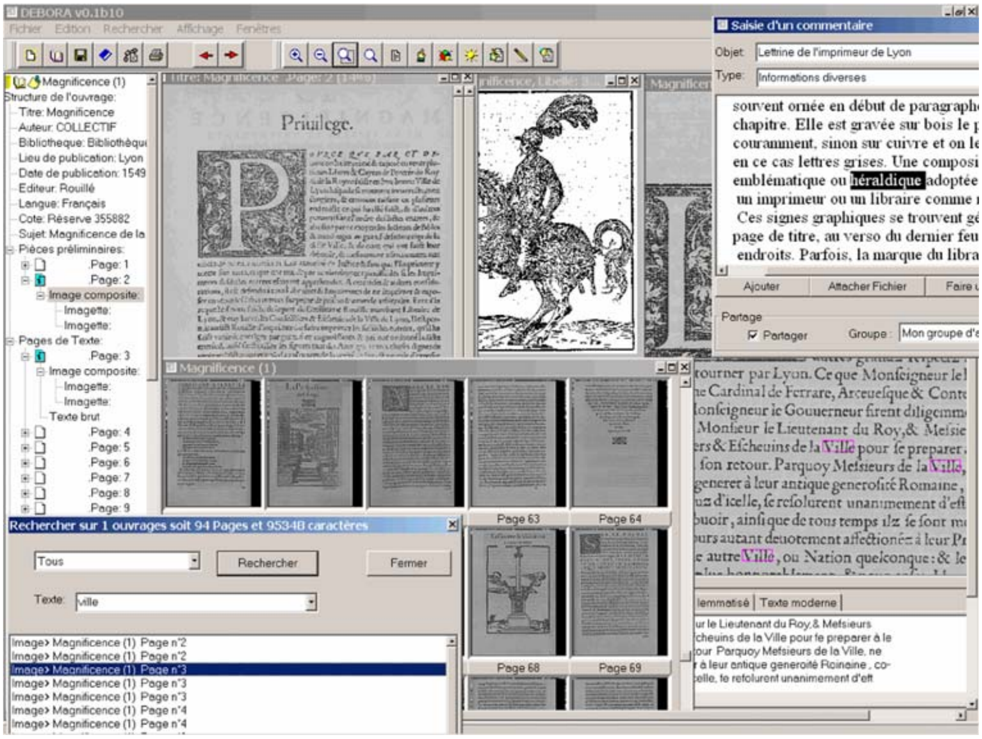
Annotation management The tagged file supports annotations of any element (physical or logical) at any level of the hierarchy with any media (rich text, image, sound, video, executables, or any existing files). Annotations are also referenced by author, date, subject, annotated elements, and authority level for improved access with internal research tools. As annotations may increase the size of files, a user can download only relevant annotations according to precise criteria. When the user is online, new shared annotations are received and transmitted through the server, which compares versions.

Editing The file format allows the user to recompose any parts of several books into a personal document with all or a part of the related annotations. This function is useful for researchers who wish to produce a new document with the components that interest them. This private document is saved in the same format and can be shared with other users.

Search tool capabilities The users need standardized interrogation, browsing, and search tools that are easily mastered and that run in a relatively homogenous manner. The proposed file format allows complex queries for any downloaded elements. The search tools today available in our browser allow queries on annotations and on any component of the logical structure and the physical layout (Figs. 35, 36).

Manageability As we cannot describe all books with the same logical structure, the proposed format takes advantage of XML capabilities to describe a document with a user-defined structure.

Table 2 in Sect. 3.7 gives the final sizes of standalone compressed files for each book. It shows that a compressed book required an average of only 20 Mb. We can evaluate the size of descriptions (tags and metadata) by comparing the final size of the tagged file and the size of the compressed images in Table 2. The tags and the metadata increase the size of the final file by 10–50%, depending on the complexity of the book contents and the quantity of metadata. Figures 35 and 36 show some response to preprogrammed queries such as the search for graphical elements and keywords. Other queries are also available such as the search of pages containing right alignments, having a given number of text lines or an ornament. Statistics can be also programmed such as the redundancy rate page by page to find the insertion of new pages. But we did not develop a language to achieve these queries and they must be implemented directly in the program. The precision of the textual query is not good because of the variation of the syntax and the frequent substitution between the letters 'u', 'v' and 'i', 'j' in sixteenth century. A more efficient textual search engine that tolerates variation of syntax and missing characters must be added in the end-user browser software. A tree-view allows the user to navigate using both the physical and the logical structure of the book. Drag-and-drop functions make it possible to edit the logical structure of books.



Magnificence (1)
Structure de l'ouvrage:
Titre: Magnificence
Auteur: COLLECTIF
Bibliotheque: Bibliothèqu
Lieu de publication: Lyon
Date de publication: 1549
Editeur: Rouillé
Langue: Français
Cote: Réserve 355882
Sujet: Magnificence de la
Pièces préliminaires:
Page: 1
Page: 2
Image composite:
Imagette:
Imagette:
Pages de Texte:
Page: 3
Image composite:
Imagette:
Imagette:
Texte brut
Page: 4
Page: 5
Page: 6
Page: 7
Page: 8
Page: 9

Objet: Lettrine de l'imprimeur de Lyon
Type: Informations diverses

souvent ornée en début de paragraph
chapitre. Elle est gravée sur bois le p
couramment, sinon sur cuivre et on le
en ce cas lettres grises. Une composi
emblématique ou héraldique adoptée
un imprimeur ou un libraire comme
Ces signes graphiques se trouvent gé
page de titre, au verso du dernier feu
endroits. Parfois, la marque du libra

Partage
☑ Partager Groupe : Mon groupe d'e

tourner par Lyon. Ce que Monſeigneur le
ne Cardinal de Ferrare, Arceueſque & Cont
Monſeigneur le Gouuerneur firent diligemm
Monſieur le Lieutenant du Roy,& Meſsie
ers & Eſcheuins de la Ville pour ſe preparer
ſon retour. Parquoy Meſsieurs de la Ville,
generer à leur antique generoſité Romaine,
us d'icelle, ſe reſolurent unanimement d'eſt
buoir, ainſi que de tous temps ilz ſe ſont m
urs autant deuotement affectionéz à leur P
e autre Ville, ou Nation quelconque: & le

ur le Lieutenant du Roy,& Mefsieurs
ſcheuins de la Ville pour ſe preparer à le
our Parquoy Mefsieurs de la Ville, ne
r à leur antique generoité Roinaine , co-
elle, fe refolurent unanimement d'eft

Rechercher sur 1 ouvrages soit 94 Pages et 95348 caractères
Tous
Texte: ville

Image> Magnificence (1) Page n°2
Image> Magnificence (1) Page n°2
Image> Magnificence (1) Page n°3
Image> Magnificence (1) Page n°3
Image> Magnificence (1) Page n°3
Image> Magnificence (1) Page n°4
Image> Magnificence (1) Page n°4

Fig. 35 Browser for DEBORA compressed files (http://rfv6.insa-lyon.fr/debora/client.html)

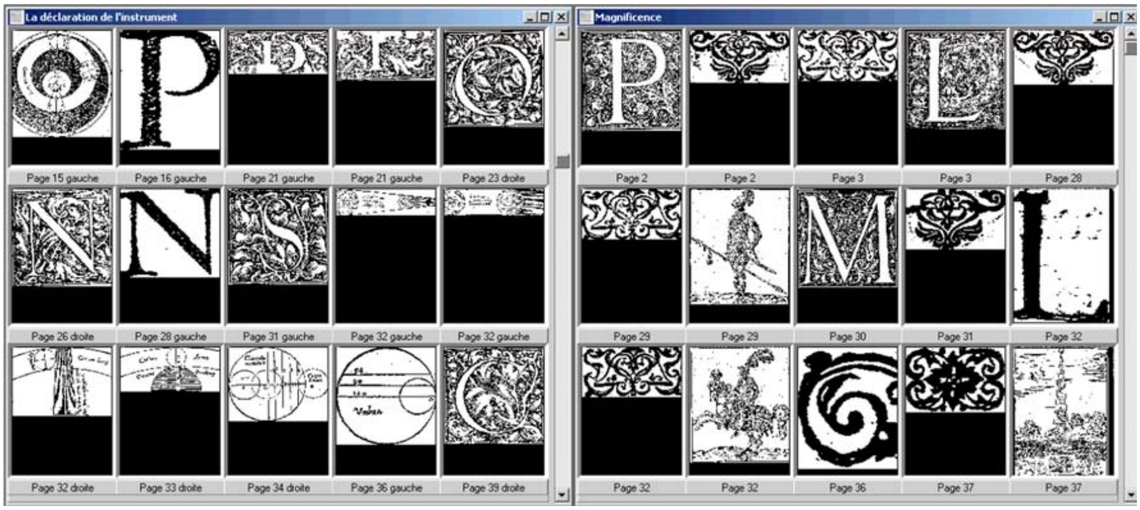


Fig. 36 Query for some graphical elements in two different books: BML355882, BML 373177

6 Conclusion

The tools developed were evaluated by end-users and Renaissance specialists. The prototype DEBORA has shown numerous advantages such as reduction in the size of digitized books, portability, computer-assisted transcription, access to typographical information, and in particular the pattern dictionary and the physical layouts, as well as the possibility to study the regularity of book manufacturing by analyzing the compression ratio of the textual layer. Several negative points have been highlighted such as the text search engine that cannot tolerate syntax variation and the lack of metadata on graphical elements (*Fleurons, illumination, frontispiece, illustration, printer's mark, miniature, etc.*) and on textual elements *(folio, signature and other marks)*. The recognition of these elements requires intensive development that must use several features and generic knowledge on Renaissance books. However, the recognition of these elements remains difficult because of the variability in their specification. The recognition of the page reference *(folio and signature)* is an important issue that is currently under investigation. In spite of the limitation of the DEBORA prototype, end-users have positively received the advances and the concepts developed for this project. The DEBORA prototype has also developed new needs in terms of image analysis, most particularly extending the comparison of character patterns to a similarity measure between graphical elements so as to objectively identify the editor and manufacturer of a book. A large content-based image retrieval system is needed to find typographical elements used by each manufacturer automatically. The concepts developed in DEBORA are currently being used in other research projects such as AGORA in collaboration with the CESR (Centre d'Études Supèrieures de la Renaissance) of Tours for the recognition of metadata in Renaissance books. In the project Agora, we currently work on the optimization of the pattern matching algorithm to reduce the number of prototype of characters patterns and increase the transcription speed. We also work on Latin contraction and the unicode encoding of old characters patterns with Andre [26,27]. Because of the lack of unicode for ancient characters of the Renaissance, both DEBORA and AGORA systems cannot display obsolete characters and we still replace them by special ASCII codes which make the transcription difficult to read. We also work with the ATILF laboratory (*Analyse et Traitement Informatique de la Langue Française*) of Nancy to analyze the typography of early printed dictionaries and in particular the dictionary of Trevoux [28]. On the other hand, they provide us electronic dictionary of old French language of the Renaissance which allow the development of a syntactical stage. This stage can speed up the transcription step by replacing automatically missing characters for unambiguous words at the end of the process.

References

1. Le Bourgeois, F., et al.: Document images analysis solutions for digital libraries. In: Proceedings of first International Workshop on Document Image Analysis for Libraries (DIAL'04). Palo Alto, California, pp. 2–24, 23–24 January 2004
2. http://debora.enssib.fr
3. DEBORA: European project, on-line book, 171p. http://rfv6.insa-lyon.fr/debora (2000)
4. Trinh, E.: De la numérisation à la consultation des documents anciens : Elaboration de procédures de numérisation, de traitements de restauration et proposition d'une plate-forme de consultation, PhD, INSA de Lyon, Villeurbanne France, 212 p, 3 April 2003
5. Sauvola, J., et al.: Adaptative document binarization. In: Proceedings of the 4th International Conference on Document Analysis and Recognition, ICDAR'97, vol. 1, Ulm, Allemagne, pp. 147–152 (1997)
6. Wolf, C.: Text localization enhancement and binarization in multimedia documents. In: Proceedings of the ICPR'02, vol 2, August 11–15 2002, Québec, Canada, pp. 1037–1040
7. Le Bourgeois, F., Kaileh, H.: Automatic metadata retrieval from ancient manuscripts. In: Proceedings of International Workshop on Documents Analysis Systems (DAS2004), Florence, 8–10 September 2004
8. Hunter, R., Robinson, A.: International digital facsimile coding standards. Proc. IEEE **68**, 854–867 (1980)
9. Bodson, D., Urban, S., Deutermann, A., Clarke, C.: Measurement of data compression in advanced group 4 facsimile system. Proc. IEEE **73**, 731–739 (1985)
10. JBIG Committee: ISO/IEC JTC1/SC29/WG1 (ITU-T-SG8) WD 14492, (1998)
11. Pennebaker, W., Mitchell, J., Langdon, G., Arps, R.: An overview of the basic principles of the Q-coder adaptative binary arithmetic coder. IBM J Res. Dev. **32**, 717–726 (1988)
12. Kia, O.E.: Document image compression and analysis. Ph.D. of the university of Maryland, 1997, p. 191 (1997)
13. Howard, P.: Lossless and lossy compression of text images by soft pattern matching. In: Proceedings of the IEEE Data compression Conference, pp. 210–219 (1996)
14. Howard, P., Kossentini, F., Martins, B., Forchhammer, S., Rucklide, W., Ono, F.: The emerging JBIG2 standard. IEEE Trans. Circ. Syst. Video Technol. **8**(5), 838–848 (1998)
15. Bottou, L., Haffner, P., Howard, P.G., Simard, P., Bengio, Y., LeCun, Y.: High-quality document image compression with DjVu. J Electron. Imaging, **7**(3), 410–428 (1998)
16. Asher, R., Nagy, G.: A means for achieving a high degree of compaction on scan-digitized printed text. IEEE Trans. Comput. **23**, 1174–1179 (1974)
17. Wong, K., Casey, R., Wahl, F.: Document analysis system. IBM J Res. Dev. **26**, 647–656 (1982)
18. Mohiuddin, K., Rissanen, J., Arps, R.: Lossless binary image compression based on pattern matching. Proceedings of the International Conference On Computers, Systems and Signal Processing, pp. 447–451 (1984)
19. Witten, I., Bell, T., Emberson, H., Inglis, S., Moffat, A.: Textual image compression: two stage lossy/lossless encoding of textual images. Proc. IEEE **82**, 878–888 (1994)
20. Inglis, S., Witten, I.: Compression-based template matching. Proc. of the IEEE Data Compression Conference, pp. 106–115 (1994)
21. LeBourgeois, F., Emptoz, H.: Document Analysis in gray level and typography Extraction using character pattern redundancies. In: proceedings of the 5th ICDAR, Bangalore India, pp. 177–180, 20–22 (1999)
22. Gross, A., Latecki, L.J.: Digital geometric methods in document image analysis. Pattern Recogn. **32**, pp. 407–424 (1999)
23. Sarkar P., et al.: Spatial sampling of printed patterns. IEEE Trans. Pattern Anal. Mach. Intell. 20: pp. 344–351 (1991)
24. Le Bourgeois F., et al.: Networking digital document images. Proceedings of the ICDAR, Seattle, pp. 379–383 (2001)
25. O'Gorman, Binarization and multi-thresholding of document images using connectivity. Comput. Vis. Graph. Image Process. J. Graph. Models Image Process. **56**(6), 494–506 (1994)
26. Hersch, R., André, J., Brown, H.: Electronic publishing, artistic imaging, and digital typography. Springer, Berlin Heidelberg New York (1998)
27. André, J.: Numérisation et codage des caractères de livres anciens. J. Doc. Numér. **7**(3), 127–142 (2003)

28. Turcan, I.: L’édition scientifique d‘ouvrages anciens sur support électronique: éthique méthodologique du traitement numérique des ornements et marques typographiques des dictionnaires dans le programme de numérisation des collections d’ouvrages anciens du laboratoire ATILF, actes de la XIVe Conférence Européenne TeX (EuroTeX’2003), Retour à la typographie. Brest, 24–27 juin 2003

29. Bres, S., Jolion, J.M., Le Bourgeois, F.: Traitement et analyse des images numériques. Paris Hermès Lavoisier. ISBN 2-7462-0741-9, 408 p (2003)

30. Nadler, L.: A survey of document segmentation and coding techniques. Comput. Vis. Graph. Image process. **28**, 240–262 (1984)