DocExplore: Overcoming Cultural and Physical Barriers to Access Ancient Documents

Pierrick Tranouez, Nicolas Stéphane, Dovgalecs Vladislavs, Burnett Alexandre, Heutte Laurent, Liang Yiqing, Guest Richard, Fairhurst Michael

▶ **To cite this version:**

Pierrick Tranouez, Nicolas Stéphane, Dovgalecs Vladislavs, Burnett Alexandre, Heutte Laurent, et al.. DocExplore: Overcoming Cultural and Physical Barriers to Access Ancient Documents. ACM Document Engineering, Sep 2012, France. pp.NA. hal-00730420

HAL Id: hal-00730420
https://hal.science/hal-00730420v1

Submitted on 10 Sep 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

DocExplore: Overcoming Cultural and Physical Barriers to Access Ancient Documents

Pierrick Tranouez, Stéphane Nicolas, Vladislavs Dovgalecs, Alexandre Burnett, Laurent Heutte
University of Rouen, LITIS EA 4108
BP 12 – 76801 - Saint-Etienne du Rouvray – France
FirstName.SurName@univ-rouen.fr

Yiqing Liang, Richard Guest, Michael Fairhurst
School of Engineering and Digital Arts, University of Kent
Canterbury, CT2 7NT – UK
{Y.Liang, R.M.Guest, M.C.FairHurst}@kent.ac.uk

ABSTRACT
In this paper, we describe DocExplore, an integrated software suite centered on the handling of digitized documents with an emphasis on ancient manuscripts. This software suite allows the augmentation and exploration of ancient documents of cultural interest. Specialists can add textual and multimedia data and metadata to digitized documents through a graphical interface that does not require technical knowledge. They are helped in this endeavor by sophisticated document analysis tools that allows for instance to spot words or patterns in images of documents. The suite is intended to ease considerably the process of bringing locked away historical materials to the attention of the general public by covering all the steps from managing a digital collection to creating interactive presentations suited for cultural exhibitions. Its genesis and sustained development reside in a collaboration of archivists, historians and computer scientists, the latter being not only in charge of the development of the software, but also of creating and incorporating novel pattern recognition for document analysis techniques.

Categories and Subject Descriptors
H.3.7 [**Digital Libraries**]: Collection, Dissemination, Systems issues
J.5 [**Arts and Humanities**]

Keywords
cultural heritage, historical documents, manuscripts, document image analysis, document indexing, word spotting, authoring system, virtual book.

1. OBSTACLES AND SOLUTIONS

1.1 Obstacles
Many institutions hold within their depths unique artifacts of times past, concealed from the public's eye. For instance, the Municipal Library in Rouen stores a document called the Ivory Book, an aggregation of manuscripts ranging from the 10th to 13th century, bound in a magnificent ivory plated cover. It is precious and fragile, therefore seldom accessed except by a few scholars and archivists. Another such document is a 20-meter long roll of scroll, which adds its cumbersomeness to the previously mentioned fragility. They would be worthy subjects of exhibition, as they are both beautiful and interesting, but are instead stored carefully for preservation.

Digitalization is a first step to breach this physical barrier preventing a general access, but it is not yet sufficient. Indeed, ancient documents, for example medieval manuscripts, can only be understood or appreciated if inserted into a coherent cultural context. The writing and the language are the first obvious obstacles, but a global understanding of the background of the document can be necessary to reach a culture so distant in time it may seem alien to a general contemporary observer.

1.2 Digitally overcoming these difficulties
Scientific and technological endeavors to help our written heritage enter the digital age are a field of rapidly growing importance. The number of conferences and workshops on digital humanities has exploded in the last couple of years. We can cite for example the HIP '11[1], ESF Digital Paleography[2], or DISH[3] conferences.

Some of these contributions aim at the scholar community: tools for transcribing or translating, annotations, simple text recognition (much more difficult in ancient documents than in modern productions) [1]. Others are supposed to be used by the archivists, helping them organize masses of digitized documents, often through the use of metadata and adapted indexable file formats [2]. Finally, others try to make the documents accessible to a more general public by developing software interfaces or simply viewers of digitized media [3].

The DocExplore project intends to overcome the obstacles of the previous section through the creation of augmented documents. As we will describe, we borrow from the three preceding fields, allowing scholars to pour their knowledge into textual and

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

DocEng'12, September 4–7, 2012, Paris, France.

Copyright 2012 ACM 978-1-4503-1116-8/12/09...$15.00.

[1] http://www.comp.nus.edu.sg/~hdocp/

[2] http:/www.zde.uni-wuerzburg.de/veranstaltungen/digital_palaeography/

[3] http://www.dish2011.nl/

multimedia metadata that enhance the digitized document. We also provide tools for building exciting visualizations extracted from the augmented documents. Furthermore we capitalize on our past experience in handwriting recognition to try to provide tools for searching text in images of variable quality and from exotic handwriting. We want to spot words [4] without first segmenting the image into lines and words, thus letting us apply similar techniques to spot various kinds of patterns in the document image: details of illuminations, faces, crowns, colors etc.

2. OUR PROJECT

2.1 An Interreg IVa project

The DocExplore research project was funded by Interreg IVa France (Channel) England. It brought together on both sides of the Channel historians, archivists and computer scientists. The English side is centered on the University of Kent and the Canterbury Cathedral Archive, while the French side includes the University of Rouen and Rouen Municipal Library. It is headed by the LITIS laboratory.

Our project aims to investigate and implement an IT-based system for the exploration of historical documents, providing solutions that enhance the interaction with and understanding of documents and associated metadata.

The strong multi-discipline partnership ensures that DocExplore is not a pure theoretical project, but an effective effort aiming at satisfying at the same time computer science research scientist, historians, and archivists or librarian.

2.2 The DocExplore Software Suite

We are building a software suite that lets users build augmented documents, as well as multimedia presentations of these documents: 3D models of the codex, textured with high-resolution photographs of the digitized manuscript, as well as selected parts of the *augmentation* of the documents, relatively to the emphasis of the presentation. This suite is composed of three applications: (1) one application providing advanced computer-aided functionalities to manage and annotate collections of digitized manuscripts (the Manuscript Management Tool), (2) one application to easily build multimedia presentations of augmented documents (the Authoring Tool), and (3) one application to visualize and interact with multimedia presentations (the Viewer). The viewer is the front-end application for the general public, the authoring tool is more dedicated to the archivists and the professionals of cultural heritage preservation and promotion, and the manuscript management tool allows the scholars and researchers to study and enrich the digitized sources. Thus our software suite makes it possible to gather those, professional and general public, producers and consumers, who are concerned by cultural heritage discovery. The general architecture of our system is depicted on the Figure 1.

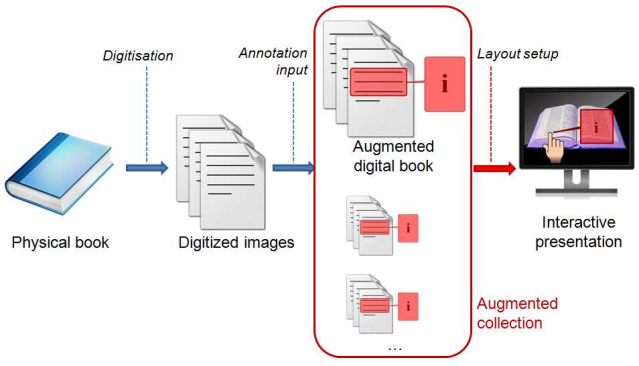

Figure 1. General architecture

2.2.1 *A Manuscript Management Tool (MMT)*

The MMT may be the most complex application of the three: it allows to aggregate pictures of manuscript in codices, it provides tools for text and pattern recognition or spotting, and allows historians or archivists to enhance the document with any knowledge they deem useful: transcriptions, annotations, comments, pictures, movies, hyperlinks (including inside the document). The user interface of the MMT is illustrated on figure 2.

2.2.1.1 *Building virtual codices*

The initial digitization leads to the creation of masses of document images. For other uses the images can then be embedded in a metadata containing file format (e.g. ALTO/METS), but in our suite we suppose we just have raw high-definition images (e.g. TIFF).

The first step of building an augmented document is aggregating a selection of these images into a virtual codex. It can be the same codex containing the original manuscripts, but it can also be a selection of some folios, a compilation etc.

Virtual libraries of such codices can be assembled.

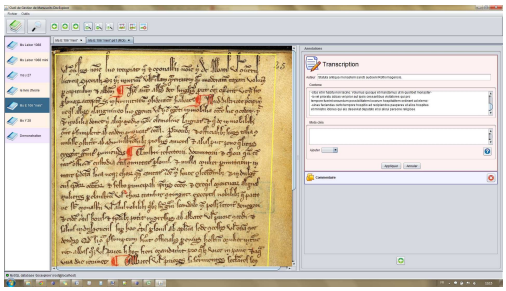

Figure 2. The Manuscript Management Tool

2.2.1.2 *Augmenting the document*

Data and metadata can be added to the virtual codex. It can be related to the whole document, to one page or to a selection of this page called a Region Of Interest (ROI). ROIs can have any polygonal shape: they could delimit a face in a painting, a part of a text, an illumination etc.

Textual data linked to a ROI, a page or a book can be metadata such as keywords or tags, but also a transcription of text or any type of comment. The data can more generally be multimedia (movies, animations, pictures), or a mix of all these types.

The ROIs can have attributed hyperlinks indicating families of illuminations, variants on the scripting of a same word etc. The same links can point to other augmented documents of the same virtual library.

2.2.1.3 *Indexing, Recognition or spotting: facilitating the navigation*

One of the strengths of a digital document over a mundane one is its ability to be automatically indexed and searched. It is quite easy to provide this functionality to the added textual data.

We also try to address this difficulty in the actual text as depicted in the images of the manuscripts. On recent documents OCR allows a good automatic transcription that can be afterwards indexed. This is generally not the case on ancient documents: manuscripts are often hand written, as their name implies; they may have had an imperfect conservation trajectory, leading to degradations not fixable at the image manipulation level; they can have complicated scripts, with a very dense interweaving of ascenders and descenders. OCRs often fail spectacularly on this kind of document [5].

If full recognition is often not possible, word spotting may be more efficient [6].

Figure 3. Query word (left) and retrieval result (right)

Completely new perspectives in digital document exploration are opened if word and general graphical pattern search capabilities are added to the existing system. Users are able to highlight a general object of interest such as a word (see Figure 3) or a pattern (see Figure 4) and the system returns the digitized pages containing the requested object and its location. Collection-wide search using this approach offers an efficient work tool for historians trying to discover regularities in the studied documents.

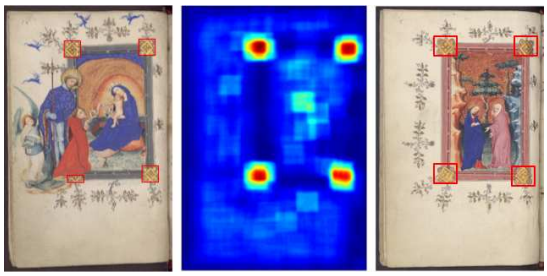

Figure 4. Four user selected queries (left), heat-map for pattern detection (middle) and four detected regions (right)

The challenge lies in the fact that digitized documents are of very complex structure and exhibit variable writing styles due to different authors or ages amongst other issues. This advanced search capability will naturally extend the software suite enabling exploration and search functionality. At the same time, historians can use the same approach for novel and unexplored manuscript annotation and study.

2.2.2 An authoring tool

The aim of the authoring tool is to allow building multimedia presentations, by defining a course through a selection of pages and annotations within a collection of augmented documents. This course can be linear like a book, but not necessarily - the presentation could also be like a graph of hyperlinked pages. The authoring tool interacts with the MMT, and allows the editor of the presentation to select the data and the associated metadata he wants to show in his presentation. This authoring tool is designed for use by anyone and requires no particular skills in programming, text encoding or data processing. The user has simply to provide a selection of the documents from a collection he wants to include in his multimedia presentation as well as supplementary metadata as attachment. The user can also define the order of the included data in the presentation easily by drag and drop operations, or the relationships between them in case of a non-linear presentation with links. The appearance of the user interface of the authoring tool is quite similar to that of the MMT in order to facilitate the user experience. The authoring tool is depicted on the Figure 5.

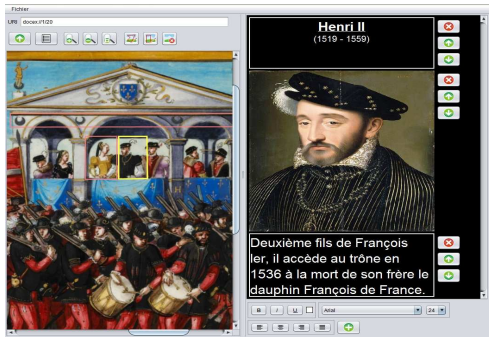

Figure 5. Multimedia presentation publishing using the authoring tool

The authoring tool bridges the gap between the knowledge database fed by the scholars through the MMT and the multimedia presentation restituted to the general public thanks to a multimedia viewer. This is the component that allows integration of the data and the metadata in a multimedia presentation, encompassing both the physical and cultural obstacles discussed in the introduction. Authoring tools available for the publishing of cultural heritage sources are few. Among the recent commercial solutions we can cite Apple's iBook Author software. However this solution is dedicated to the publishing on tablets, and not to the production of multimedia presentation for museums or libraries. Furthermore, the strength of our solution relies in the possibility to interact with the knowledge database provided by the MMT.

2.2.3 A viewer

The viewer is a front-end component of the system intended to present and allow interaction with the document presentations built in the authoring tool, for use by the general public in exhibition locations such as libraries or museums. This viewer exploits 3D modeling and textured rendering of the manuscripts (see Figure 6), in order to enhance the discovering experience of the users, and give them as much as possible the impression of handling a real book by proposing natural interactions, such as page flipping using touchscreen or gesture interactions.

Figure 6. 3D page rendering in the viewer

This viewer looks like some other digital book viewers, such as the Turning the Pages viewer[4] (a review of some viewing solutions dedicated to cultural heritage content is provided in [3]), and it enables the user to browse in a natural way the digitized content provided inside by flipping the pages. But it allows much more by offering access to supplementary material provided by the researchers, scholars, or any editor of the content. This material can be textual information such as transcriptions, translations, annotations, commentaries, as well as images, audio, video or hyperlinks to some other content. Therefore, the viewer is much more than a simple “page flipping app” for digitized

[4] http://www.turningthepages.com/

content: this is a viewer dedicated to the visualization of and interaction with augmented documents (see Figure 7).

Figure 7. Viewing the metadata attached to a ROI in the viewer

Furthermore, it works in a client-server mode, and is able to connect to a database to provide the user with access to several multimedia presentations available on the server. Thus the viewer is not dedicated to one particular digital book, but can be used to browse a large variety of multimedia presentations. For instance, it can be installed on multiple consulting stations in a library or a museum, allowing a centralized distribution and management of the diffused content. The inherent problems involved by the loading time of the data are obviously dealt with by the viewer.

From a technical point of view, our viewer is implemented using Java and OpenGL technologies. This choice is motivated by the fact that Java and OpenGL are supported by most platforms, including web browsers.

2.3 First uses of the DocExplore Software Suite

2.3.1 Manuscript Management Tool

Our historian partners have made use of the MMT to annotate and transcribe more than a thousand pages of medieval manuscripts. Hundreds of those transcriptions were aligned with the pictured words. They have thus built a learning base for our pattern recognition algorithms.

Our developers are currently addressing their bug reports and overall usability remarks.

2.3.2 Authoring and viewer tools

We joined our partner of Rouen Municipal Library at an ancient books trade fair called Salon du Livre Ancien in 2010, 2011 and 2012, where we demonstrated augmented documents created and presented with our suite. Attendant feedback was very positive[5].

This experience gave us a list of improvements needed for those tools, which we've implemented. A new version of the suite has been given to our partners in early January, so as to lead to a new cycle of improvements.

2.3.3 Advanced Search Tool

In our first part of preliminary experiments with word spotting in ancient documents we've obtained promising results by returning highly relevant and precise locations of the requested words. In the second part of our experiments we managed to obtain precise detection of requested objects in spite of natural visual variability, which is due to drawings performed by hand.

Our precision and recall need to be more formally evaluated and improved, but our first results are very encouraging.

[5] http://klog.hautetfort.com/archive/2012/04/01/un-feuilleteur-pour-valoriser-les-documents-patrimoniaux.html

3. DISCUSSION

The primary objective of the DocExplore Software Suite is to let archivists and scholars easily augment digital facsimiles of valuable and rare manuscripts, an application of this augmentation being the production of rich multimedia creations. The use of the suite does not imply advanced knowledge of data representation formats (SGML, XML, TEI schema), while allowing the definition of various rich annotations and interpretations.

The augmented multimedia creations can be used in cultural heritage exhibitions, by enhancing interactions with ancient documents, through natural user interfaces (currently touch and gestures) and multimedia content (images, audio, video, interlinks, hyperlinks). We are building an exhibition, scheduled for September 2013, on the theme of "Writings from the Middle Ages to Present Time in the Anglo-Norman Region". The DocExplore Software Suite will be an essential part of this setup.

We are currently working on the integration of pattern recognition functionalities, to complement full text search in annotations and transcriptions, at first in the form of *word spotting*: multiple occurrences of the requested word can be returned upon user's request. Finally, an extension of existing tools and approaches will allow searching for free form graphical patterns in a seamless manner.

4. ACKNOWLEDGMENTS

The authors would like to thank Interreg IVa for its funding of the DocExplore Project, as well as our partners, Catherine Richardson and Alixe Bovey from the University of Kent, Cressida Williams from Canterbury Cathedral Archives, Elisabeth Lalou, Alexis Grelois and Cécile Capot from GRHis laboratory of the University of Rouen, and Vincent Viallefond and Maïté Vanmarque from Rouen Municipal Library.

5. REFERENCES

[1] Romero, V, Serrano, N., Toselli, H. A., Sanchez, A. J., and Vidal, E. 2011. Handwritten Text Recognition for Historical Documents. In *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage*.

[2] Constantopoulos, P., Doerr, M., Theodoridou, M., and Tzobanakis, M. 2002. Historical documents as monuments and as sources. In : *Proceedings of Computer Applications in Archaeology 2002*.

[3] Cauchard,J.R., Ainsworth, P.F., Romano, D.M., and Banks, B. 2006. *Virtual Manuscripts for an Enhanced Museum and Web Experience - "Living Manuscripts".* 12th International Conference on Virtual Systems and Multimedia, 18-20 October 2006, Xi'an, China.

[4] Leydier, Y., Lebourgeois, F. and Emptoz, H. 2007. Text search for medieval manuscript images. *Pattern Recognition* no. 40, pp. 3552-3567.

[5] Govindaraju, V., Cao, H., and Bhardwaj, A. 2009. Handwritten document retrieval strategies. In *Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data* (AND '09). ACM, New York, NY, USA, 3-7.

[6] T.M. Rath and R. Manmatha. Word spotting for historical documents. *International Journal on Document Analysis and Recognition,* vol. 9, no. 2, pp. 139-152, 2007