A Generic Recognition System for Making Archives Documents accessible to Public

Bertrand Coüasnon* Ivan Leplumey**

* IRISA / INRIA
Campus universitaire de Beaulieu
F-35042 Rennes Cedex, France
Bertrand.Couasnon@irisa.fr

** IRISA / INSA-Département Informatique
20, Avenue des buttes de Coësmes
F-35043 Rennes Cedex, France
Ivan.Leplumey@irisa.fr

Abstract

This paper presents annotations needed for handwritten archives document retrieval by content. We propose two complementary ways of producing those annotations : automatically by using optical document recognition and collectively by using Internet and a manual input by users. A platform for managing those annotations is presented as well as examples of automatic annotations on civil status registers, military forms (tested on 60,000 pages) and naturalization decrees, using a generic document recognition method. Examples of collective annotations built on automatic annotations are also given.

1 Introduction

French Archives, like other archives services, own millions of pages of documents with handwritten information which are difficult to access for public. Indeed a reader needs to leaf through a uge number of pages before finding the page containing an information he is looking for. In the same time a growing number of people, like genealogists, are interested in those documents. Archives have then a great challenge : how can they make a public access to millions of pages of documents with handwritten information ?

They started to scan some of those documents. It produces a digital save of paper documents, it offers the possibility of web access, of simultaneous access and of virtual leaf through.

However, scanning is not enough, as the difficulty to find a document is the same as on paper or microfilms : it is still necessary to leaf (virtually) through a uge amount of documents (images). Even if a page can be viewed by a user, the time needed to find the right page is so important that those documents can be considered as unaccessible.

Systems must be defined to allow document retrieval by content. To do so, it is necessary to associate annotations to the images of documents. With those annotations, it is then possible to make an automatic selection of images.

We present in this paper the kind of annotations needed for archives retrieval by content. We propose two complementary ways of producing those annotations : automatically by using optical document recognition and collectively by using Internet and a manual input by users. We present a platform we developed to manage collective annotations builded on automatic annotations. Those collective annotations are made by users (if they want to) when they read a document, making them available for the others to have a better access to documents.

In section 5, we show the application of this platform on various documents: civil status registers, military forms of the 19th century and naturalization decrees. For each document we present the automatic annotations we are able to produce with DMOS, a generic document recognition method we developed. We also present the collective annotations that can be added by users with the help of the automatic annotation. This platform offers a uniform interface for accessing archives documents by content.

2 Annotations

To make an image of document accessible by content, annotations must be associated to each image. We propose to consider two kinds of annotations for archives :

textual annotations: a date, a place, a name, a keyword... All kind of information on which it is interesting to make a research. This information need to be structured, to represent the logical structure existing in the original document;

geometric annotations: a position in the image like a cell, a field, a zone, represented by a rectangle or a polygon.

Of course, all textual annotation can be linked to a geometric annotation. This is important, for example, to represent the fact that a specific name is in a cell in the document. The classical problem of annotations is, before storing them, how can we produce them?

3 Producing Annotations

3.1 Automatic Annotations

Automatic annotations are produced by optical document recognition. On printed and recent documents, existing OCR

Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003)
0-7695-1960-1/03 $17.00 © 2003 **IEEE**

systems are able to recognize almost all the text which can then be used to build textual annotations for document retrieval. On archives and old documents, it is more difficult because documents were not always well preserved and they can be damaged (tears, blots, tape repairs, smudge...). The paper could have been stored in humid conditions making it warp. This is a problem because a lot of documents are digitalized with a camera without pressing the paper with a glass. Therefore, in the image, lines are transformed into curves. With time, ink crossed the paper making back side visible. Stamps can be affixed onto documents. Sometimes, sheets of papers can be pasted on documents hiding part of it...

When these old documents are printed, the characters are not well printed, making difficult their recognition. We can find some work done to access to books of the 16th century [6] and printed documents of the 19th century [7]. But when documents are handwritten, the bad quality of documents is added to the difficulty of handwriting recognition. This could explain why we were not able to find, in the literature, related work on handwritten archives document retrieval.

To produce automatically annotations on those kind of documents, it is necessary to first locate where the needed information for retrieval is in the image of document. This location allows to detect what part of the image contains handwriting and what type of information is there. For example, it is important to be able to find in the image where is a name, a date, a place...on which a search can be done. After this location, it is possible to work on handwriting recognition.

To be able to detect the position of specific handwritten text, it is important that the document is structured enough. Therefore we can work on structured documents (section 5) like forms, tables or less structured documents like only handwritten text if it is graphically structured with margins, paragraphs...

As the number of type of archives documents is important, it is not possible to develop a new recognition system for each type of document. In [1] we presented DMOS (Description and MOdification of Segmentation), a generic recognition method for structured documents. This method is made of the new grammatical formalism EPF (Enhanced Position Formalism), which can be seen as a description language for structured documents. EPF makes possible at the same time a graphical, a syntactical or even a semantical description of a document. DMOS contains also the associated parser which is able to change the parsed structure during the parsing. This allows the system to try other segmentations with the help of context to improve recognition.

We have implemented this method to build an automatic generator of structured document recognition systems. Using this generator, an adaptation to a new kind of document is simply done by defining a description of the document with an EPF grammar. This grammar is then compiled to produce a new structured document recognition system.

With this generator, we have been able to produce various document recognition systems: one on musical scores, one on mathematical formulae, one on recursive table structures. We could even use it to make a recognition system of tennis court in videos.

We present in section 5, application of DMOS on various documents to automatically produce geometric annotations on the document structure.

3.2 Collective Annotations

Some of those handwritten archives documents are really difficult to recognize with automatic methods. Indeed, to be able to produce an automatic annotation on handwritten text, it is first necessary to locate where this text is in the document. When a document is not graphically structured enough, it is quite impossible to detect this location. Moreover the handwritten text can be so badly written that a paleographic specialist is needed to read it or to propose an hypothesis of reading.

Therefore we propose to complete the automatic annotations by manual annotations. To avoid a systematic manual input which is tedious, time consuming and costly, we propose to produce collectively those manual annotations by the readers themselves. For a reader, it is not time consuming to input some annotations during his reading. All the annotations are put together, making them available for other readers to improve their access by content to documents, even if the documents are very difficult to read. As the number of readers is important, the number of annotations can grow very fast if a tool to manage them exists, and if the process is initiated with automatic annotations.

4 A platform for image document annotations

We defined a platform on Internet to consult images of archives documents, and retrieve documents by content. This platform propose a way to use and manage automatic and collective annotations.

4.1 Related Work

Related work on annotations is mainly around XML and RDF (Resource Description Framework). RDF [9] is a foundation for processing metadata; it provides interoperability between applications that exchange machine-understandable information on the Web. RDF metadata can be used in a variety of application areas. Thus, for example, it is possible to make annotations on XML documents.

Annotea [5] is a project from the W3C for shared annotations. By annotations they mean comments, notes, explanations, or other types of external remarks that can be attached to any Web document or a selected part of the document without actually needing to touch the document. When the user gets the document he or she can also load the annotations attached to it from a selected annotation server or several servers and see what his peer group thinks. They use an RDF based annotation schema for describing annotations as metadata and

XPointer for locating the annotations in the annotated document. Those annotations are well adapt to associate informations to an XML document at a precise location in a document. On images we need the same kind of functionality : to be able to associate an annotation to a precise location in the image.

Photo-RDF [3] is a project for describing and retrieving digitized photos with RDF metadata. RDF schemas have been defined or used to associate various informations to photos: title, date, camera, lens. . . The problem of this Photo-RDF is that it is not possible to associate a precise location in the image. It is only possible to associate it to the complete image.

Hunter and Zhan propose to embed metadatas in PNG files [4]. The metadatas are also described with RDF. In this schema, it is possible to define a region in the image. The region is described with an identifier, a title, some text description and the coordinates of the region. Even if this offers the possibility to associate an annotation to a precise position in the image, this region position can be seen as an attribute of a textual annotation. For document retrieval, we need to consider that a region position in a document is an annotation as well as a textual annotation and not only an attribute.

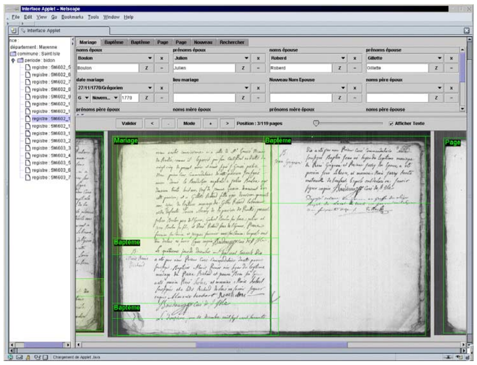
Multivalent annotations [8] offer a framework for annotations on documents of various source format: scanned documents images, HTML, DVI. . . But a position is not either considered as an annotation: lenses (geometric region annotations) give a way to transform the document under a rectangle but are not the rectangle.

4.2 Description of the platform

Therefore we propose to build a platform for archives document retrieval which could deal with textual and geometric annotations at the same level. Moreover this platform is able to create relations between textual and geometric annotations to specify that a textual information is at this specific location in the image of document. As various textual informations can be found in the same location, it is important to be able to represent as much as necessary links between textual annotations and geometric annotations. In an another way, a textual annotation can be linked to different locations in different pages of document *e.g.* in different images. So it is not possible to embed the annotations with the image. They have to be stored externally from the image like it is done in Annotea for Web documents.

We choose to develop the platform with a classical architecture: a web server (Apache) with a servlet container (TomCat); the Java servlet access to a relational database (PostGreSQL) to store annotations and send them to the client: a java applet running in a web browser (figure 1). We choose to use XML and RDF for importing or exporting annotations from the database.

To be as generic as possible on archives documents retrieval, we consider that an annotation is the smallest information that can be independently added, automatically with document recognition or manually by a reader. This smallest information is a non-structured textual annotation (a name, a date. . .) or a non-structured geometric annotation (a rectangle, a polygon. . .). Those annotations can then be structured logically (for example a birth certificate contains a name, a date, a place. . .) or physically (a register is made of images of double pages which contains two pages. . .). One or several textual annotation can be associated to one or several geometric annotation. Therefore we propose to consider an annotation as the following informations: [annotation id; creator of annotation; date of creation; type of annotation; data (a name, rectangle coordinates. . .); logical reference; physical reference; number of confirmation (increment when a reader, different from the creator, confirms this annotation)]. The allowed types of annotation are defined in a DTD configuration file which describe the structure of annotations allowed for the kind of document store in the database.



noms époux: Boulon
prénoms époux: Julien
noms épouse: Roberd
prénoms épouse: Gillette
date mariage: 27/11/1770/Grégorien (G, Novem..., 1770)
lieu mariage:
Nouveau Nom Epouse:
noms père époux:
prénoms père époux
noms mère époux
prénoms mère époux
noms père épouse
Position : 3/119 pages
☑ Afficher Texte

Figure 1. One interface to consult, annotate, retrieve archives documents

To be able to represent that a logical structure annotation (like a birth certificate) can be linked, for example, to three geometric annotations (rectangles) on three following pages, we need to store the link between annotations : [physical annotation id; creator of annotation; date of creation; logical annotation id].

With this representation of annotations and with the platform, on a web browser, a user can leaf through images of archives documents. When a page is displayed, all the associated annotations are presented on the interface: geometric annotations are drawn on the image, the textual annotations are presented in tabs for the nodes of the structure of annotations (a marriage certificate. . .) and in field boxes for the leaves (name, date. . .)(figure 1). The reader can consult annotations, add or modify annotation (if he has the right to), but is limited by the allowed annotation structure given by the DTD configuration file, according to the kind of document. The system can also store various interpretation if readers do not agree.

Structured search or full text search is possible on all the annotations whatever the way they have been produced: automatically or manually. We present in the next section examples of use of this platform on various kinds of archives documents. We show the interest of automatic annotations and the complementary of the automatic and manual annotations.

5 Examples of Archives Documents

5.1 Register of births, marriages and deaths

Automatic Annotations Those documents are really difficult to automatically annotate, due to the weak structure and the poor quality of the handwritten text. The documents are scanned by double pages. We defined a grammar in EPF describing the notion of page. With the DMOS method we have been able to produce a recognition system which detect the position of each page and produce automatic annotations. A first test has been done on 1,407 images of double pages: 99.4% have been correctly detected with 0.6% of reject.

With these page annotations, a reader can leaf through a register page by page with a zoom automatically adapted on the page, making the reading more comfortable.

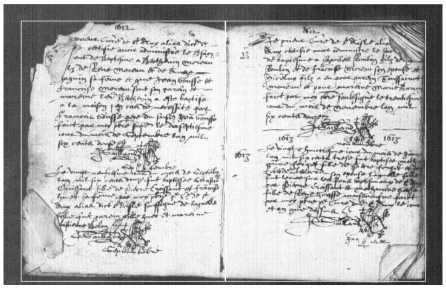


Figure 2. Automatic page annotations on a double page of a register of births and marriages

Collective Annotations On those pages, collective annotations can be added: the type of certificate (birth, marriage...); for a birth, annotations can be for example: name and surname of the child, place of birth, name of the mother, name of the father...The position of the certificate can be defined by the reader or just be associated to the automatic annotation of page. Of course there is no obligation of filling all the fields.

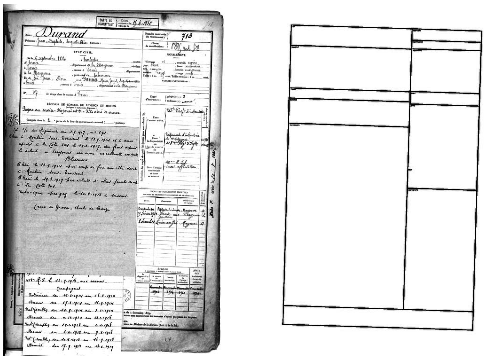
5.2 Register of Military Forms

Automatic Annotations These documents are made of quite damaged military enrollment form of the 19th century, the size of the cells change from year to year, there is a lot of pasted sheets of paper which hide the form structure, we can find stamps on it, ink can cross the paper.... We then have defined a specific EPF grammar which can take into account those difficulties.

From this specific (and small) EPF grammar, we automatically produced a new recognition system. We have first successfully tested this recognition system on 5,268 images [2]. Then we made another test, without changing the grammar, on 60,223 forms. The system rejected only 0.4% of pages (259) which are so damaged that they even can not be processed by a human operator (missing parts, wrong documents...). On the 59,964 remaining ones, the system did not make any mistake in the recognition of their structures: it allowed us to recognize correctly the complete structure of 99.6% of them, with absolutely no false recognition. Only 0.4% of forms were rejected because of their structural incorrectness (due to physical defects). We considered a structure recognition as correct when the borders of the needed cells are positioned with a precision of one millimeter. An example of the structure recognition is given in figure 3. Each cell produce an automatic annotation: a geometric annotation (the polygon of the cell) and a textual annotation (the name of the cell).

In those military forms, some cells may contain medical informations which are protected during 150 years. These informations make impossible a public diffusion of the forms. But with the automatic annotations, it is possible to automatically hide the medical cells on the platform of annotation. The documents can now be presented to the public.

As we can get the precise position of each cell we are currently working on the recognition of handwritten fields (like the name) to produce automatic annotations on them.



(a) Example of military form with the added sheets of paper

(b) Recognized structure even when it is partially hidden

Figure 3. Example of automatic annotation (using DMOS) on military forms

Collective Annotations By changing the DTD file on the platform, it is possible to specify the allowed annotations on these military forms. For example, the cell containing some birth information or the cell containing a physical description of the person, could be collectively annoted. The user will have to select the cell (an automatic annotation) to zoom on it and to associate to it some textual annotations. All those annotations could then be used for a future query by another reader.

5.3 Naturalization Decrees

Automatic Annotations These documents are from the end of the 19th century and the beginning of the 20th century. They are unique documents which are for some people the only one which can justify their French nationality. A decree is usually made of around ten pages. They are fully handwritten or sometimes fully typewritten. They are organized in tow columns with paragraphs where each paragraph concerns one person. His name is usually the first name in the paragraph. To retrieve the decree concerning one person is very tedious: the reader needs to leaf through all the pages of all the decrees.

Compares to the military forms, the structure is very weak because only made of the organization in paragraphs of the handwritten text. Due to the genericity of the DMOS method, we have been able to define an EPF grammar describing the organization of decrees in handwritten text-line, in paragraphs and columns, only with the help of the connected components detected in the image.

From this description, by compilation, a recognition system have been produced, which is able to detect the position of the name and the file number. These positions are the automatic annotations which are added in the platform. With these annotations the platform can present a table with only images of the file number and the name. This allows a really faster leaf through a decree.

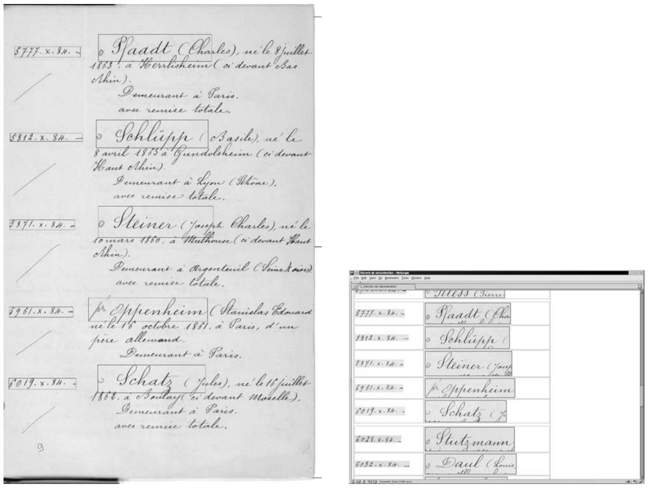
Collective Annotations When a reader found the name he was looking for, the platform presents the original page with all the existing annotations. The reader can then leaf though the original pages and if he wants, add some collective annotation like the name or the date of birth. . . .

6 Conclusion

We presented in this paper a platform to improve the access by content on archives documents with handwritten text. To make this access, annotations are needed. We showed that annotations for archives documents can be geometric or textual. The platform we propose to manage annotations, presents the interest of producing annotations in two complementary ways: automatically with document recognition and collectively with the help of the readers during their reading.

The different documents (civil status registers, military forms and naturalization decrees) on which we present the annotation platform shows the importance of a generic system for document recognition. With the DMOS method we have been able to produce new recognition systems with a minimum development work. We had to describe the documents with the EPF language to get by compilation new systems. The DMOS method has been tested on 60,000 pages of military forms. Moreover DMOS can be applied on structured documents as well as on less structured documents, like the naturalization decrees.

The platform on annotations and the complementarity of automatic and collective annotations are important to make access by the content on handwritten documents even if they are difficult to read. Depending of the difficulty of the document, the part of the automatic annotations is more or less important.



5777.x.84. – Pfaadt (Charles), né le 8 juillet 1863, à Herrlisheim (ci devant Bas Rhin).
Demeurant à Paris.
avec remise totale.

5812.x.84. – Schlüpp (Basile), né le 8 avril 1853 à Gundolsheim (ci devant Haut Rhin).
Demeurant à Lyon (Rhône).
avec remise totale.

5871.x.84. – Steiner (Joseph Charles), né le 10 mars 1860, à Mulhouse (ci devant Haut Rhin).
Demeurant à Argenteuil (Seine et Oise)
avec remise totale.

5961.x.84. – Oppenheim (Stanislas Edouard né le 16 octobre 1861, à Paris, d'un père allemand.
Demeurant à Paris.

6019.x.84. – Schatz (Jules), né le 16 juillet 1862, à Boulay (ci devant Moselle).
Demeurant à Paris.
avec remise totale.

9

(a) Automatic annotations on image: file number and name positions

(b) Table of sub-images to improve the leaf through

Figure 4. Example of automatic annotations (using DMOS) on naturalization decrees

Acknowledgments

This work have been done in cooperation with the *Archives départementales de la Mayenne*, the *Archives départementales d'Ille-et-Vilaine*, the *Archives départementales des Yvelines* and the *Archives Nationales* in France, with the support of the French Ministery of Culture, the regions *Bretagne* and *Pays de la Loire*, and the *Conseil Général des Yvelines*.

References

[1] B. Coüasnon. Dmos: A generic document recognition method, application to an automatic generator of musical scores, mathematical formulae and table structures recognition systems. In *ICDAR, International Conference on Document Analysis and Recognition*, pages 215–220, Seattle, USA, September 2001.

[2] B. Coüasnon and L. Pasquer. A real-world evaluation of a generic document recognition method applied to a military form of the 19th century. In *ICDAR, International Conference on Document Analysis and Recognition*, pages 779–783, Seattle, USA, September 2001.

[3] Describing, retrieving photos using RDF, and HTTP. W3C Note, April 2002. http://www.w3.org/TR/photo-rdf/.

[4] Jane Hunter and Zhimin Zhan. An indexing and querying system for online images based on the png format and embedded metadata. In *Proc. of the ARLIS/ANZ Conference*, Brisbane, Autralia, Sep 1999.

[5] José Kahan, Marja-Riitta Koivunen, Eric Prud'Hommeaux, and Ralph R. Swick. Annotea: An open rdf infrastructure for shared web annotations. In *Proc. of the WWW10 International Conference*, Hong Kong, May 2001.

[6] F. Lebourgeois, H. Emptoz, E. Trinh, and J. Duong. Networking digital document images. In *Proc. of the 6th ICDAR*, pages 379–383, Seattle, USA, Sep 2001.

[7] Günter Mühlberger. Automated digitisation of printed material for everyone: The metadata engine project. *RLG DigiNews*, 6(3), 2002.

[8] Thomas A. Phelps and Robert Wilensky. Multivalent annotations. In *Proc. of the First European Conference on Research and Advanced Technology for Digital Libraries*, Pisa, Italy, 1997.

[9] Resource Description Framework (RDF). Model and syntax specification. W3C Recommandation, February 1999. http://www.w3.org/TR/REC-rdf-syntax/.