Journal of Universal Computer Science, vol. 14, no. 18 (2008), 2912-2935
submitted: 22/9/08, accepted: 29/9/08, appeared: 1/10/08 © *J.UCS*

A Generic Architecture for the Conversion of Document Collections into Semantically Annotated Digital Archives

Josep Lladós
(Computer Vision Centre – Computer Science Department
Universitat Autònoma de Barcelona, Spain
josep@cvc.uab.es)

Dimosthenis Karatzas
(Computer Vision Centre – Computer Science Department
Universitat Autònoma de Barcelona, Spain
dimos@cvc.uab.es)

Joan Mas
(Computer Vision Centre – Computer Science Department
Universitat Autònoma de Barcelona, Spain
jmas@cvc.uab.es)

Gemma Sánchez
(Computer Vision Centre – Computer Science Department
Universitat Autònoma de Barcelona, Spain
gemma@cvc.uab.es)

Abstract: Mass digitization of document collections with further processing and semantic annotation is an increasing activity among libraries and archives at large for preservation, browsing and navigation, and search purposes. In this paper we propose a software architecture for the process of converting high volumes of document collections to semantically annotated digital libraries. The proposed architecture recognizes two sources of knowledge in the conversion pipeline, namely document images and humans. The Image Analysis module and the Correction and Validation module cover the initial conversion stages. In the former information is automatically extracted from document images. The latter involves human intervention at a technical level to define workflows and to validate the image processing results. The second stage, represented by the Knowledge Capture modules requires information specific to the particular knowledge domain and generally calls for expert practitioners. These two principal conversion stages are coupled with a Knowledge Management module which provides the means to organise the extracted and acquired knowledge. In terms of data propagation, the architecture follows a bottom-up process, starting with document image units, called *terms*, and progressively building meaningful *concepts* and their *relationships*. In the second part of the paper we describe a real scenario with historical document archives implemented according to the proposed architecture.

Keywords: Software Architectures, Document Image Analysis and Understanding, Digital Libraries, Document Mining.
Categories: H.3.1, H.3.2, H.3.7, J.4, J.5, M.0, M.4

1 Introduction

The conversion of paper-based document collections into an electronic form is a much revisited subject over the past decades. The various reasons why libraries and archives undertake such activities include preservation, increasing accessibility, indexing and retrieval, research etc. It can be easily appreciated that depending on the purpose of the conversion exercise, the requirements and the logistics can be very different.

Libraries are generally interested in mass-digitisation and transcription of their book collections. Numerous past and current research projects (e.g., the EU project IMPACT [IMPACT, 08]) focus on improving the automatic aspects of digitisation, document image analysis and recognition techniques. The economy of mass-digitisation dictates keeping human involvement to a minimum, therefore the ultimate goal in this case is to achieve full automation.

In other cases, as for example the Bibliothèques Virtuelles Humanistes project [BVH, 08] or the Codex Sinaiticus project [Sinaiticus, 08], the objective is to showcase specific documents of high value, provide access to a wider audience and assist scholarly research activities. Such projects can afford a much higher degree of human involvement in the conversion process, generally calling for experts in the field to interpret aspects of the information and provide semantic annotations.

In the case of archives the main purpose is indexing and retrieval of information. In such cases a considerable amount of human input is generally required, in order to ensure that the paper based information is accurately converted and tagged. Examples include the digitization of 2nd World War documents through the EU project MEMORIAL where specific interfaces were created for the editing and validation of automatically obtained transcriptions [Antonacopoulos, 04b], or the conversion of archived specimen cards from the UK Natural History Museum where Web based interfaces were used for the same reason [Downton, 07].

Independently of the end purpose, the conversion pipeline from paper documents to semantically annotated digital libraries comprises three distinct stages. Initially paper documents are scanned and converted into digital images. Typically, during this stage basic metadata identifying the document are manually or semi-automatically appended to the digital record. The subsequent stage of document analysis focuses on the extraction of information from the document image. More often than not, this refers to the automatic extraction and recognition of the textual content. Finally, the extracted information is semantically annotated and organised in order to facilitate text-based interaction with the digital collection.

While the first stage of mass-digitisation is well addressed by industry, the subsequent stages of information extraction and true knowledge creation are typically addressed by ad-hoc solutions, tailored to each individual document collection. Through this paper we attempt to address these later stages of document conversion in a generic way.

A generic framework for the conversion of document collections to digital archives should comprise a variety of distinctly different modules to assist the expert practitioner in the conversion process. Such a framework should expose automatic methods for document analysis, the ability to validate the results of automatic methods, methods to deposit knowledge resting with expert practitioners in the digital

archive and a flexible knowledge representation system. In detail, the minimum functionality expected of such a framework should be:

i. To provide the means to automatically extract information from document images. This should cover automatic document image analysis and understanding methods that can be run in a batch process fashion over selected parts of the collection.
ii. To expose interfaces allowing the human operator to invoke automatic processes, visualise, check and validate their results.
iii. To anticipate the involvement of certain groups of expert practitioners in the conversion process and supply the means for them to provide semantic annotations and otherwise deposit their knowledge about the collection in the digital archive. Expert practitioners in this sense are not necessarily the archivists themselves, but depending on the collection could be historians, art critics, geographers, sociologists etc.
iv. To define a structured way to represent the information extracted from the documents and the knowledge captured from experts. Knowledge representation should be expandable to incorporate new knowledge, but structured around specific blocks that correlate to the different physical and informational aspects of the document collection.

In this paper, we present a generic architecture, for the conversion of document collections into semantically annotated digital archives. Document conversion is a process for creating knowledge. The suggested architecture recognizes both documents and humans as possible sources of knowledge, specifies how and at which stage of the conversion process they can be introduced in the digital archive and provides a generic way to represent this knowledge, compatible with existing standards. In addition to image-based information extraction, the architecture proposes a flexible way to accommodate human input in different stages of the conversion process as deemed necessary. Furthermore, it places an emphasis on easy portability to different types of collections, achieved through a modular approach.

Following the detailed description of the architecture, we use a particular real-life scenario, the "Expedients de Frontera" ("Border Records") collection residing at the Historical Archive of Girona, Spain, to show how this architecture can be used to implement a system for the conversion of a historical document collection.

In the next section we give an overview of the state of the art in the related fields. Section 3 presents a detailed description of the architecture proposed. Section 4 shows the particular implementation for the scenario chosen while section 5 concludes the paper.

2 Overview – State of the Art

Indisputably, different document collections share different characteristics that necessitate different techniques for their conversion into semantically annotated digital archives. Nevertheless, the conversion process at the functional level can be described as a series of common tasks. Various systems have been proposed in the literature for the conversion of document collections, generally offering ad-hoc solutions to specific problems.

In the case of administrative document analysis, Hamza et al propose a system for the automatic processing of invoices generated in an industrial environment [Hamza, 08]. Their system follows an automated process, where for each new document the nearest processing experience is retrieved and used to analyse and interpret the document. Subsequently, the document is added to the document database following an incremental learning process.

The Brazilian project Nabuco [Mello, 02] aims to re-assemble historical documents, i.e., document image analysis and recognition processes are applied to generate synthetic documents looking like the original ones. It is used for broadcasting, browsing and preserving purposes.

One of the few approaches to define a generic document analysis and interpretation system is the DocMining platform [Clavier, 03, Adam, 04]. They propose an architecture for document mining, a general platform for document interpretation. It is a plug-in oriented architecture that allows an incremental integration of heterogeneous software *components* for document processing. The combination of such processing units with the domain-dependent knowledge defines an application *scenario*. Scenarios are described by *contracts*, which consist of descriptions of processing flows, the desired input and output. *Control* modules interpret scenarios by processing the associated contract. All data (documents, scenarios, contracts, etc.) are represented in XML, to facilitate data manipulation and communication inside the platform.

In [Sánchez, 04] a similar platform was developed and it can be considered an early stage of the one proposed in this paper. It consists of three components. First, a repository of modules able to extract descriptors that combined with domain-dependent knowledge and recognition strategies allow the interpretation of a target document. Second, a representation model based on a graph structure that allows to hierarchically represent the information of the document at different abstraction levels. Finally, the third component implements a sketching interface that allows the user to provide feedback to the system.

Another generic approach to document analysis and understanding is smartFIX [Dengel, 02]. SmartFIX is a requirements-driven system that permits the processing of documents ranging from fixed format forms to unstructured letters. Generally speaking, the SmartFIX architecture can be seen as a pipeline in three steps, namely input, analysis and output. Thus, the *Importer* process transfers document images from a predefined source into the *Control Database*, together with possibly known information. The *analyzer* module performs operations such as image processing, classification, information extraction, etc. Finally, the *Exporter* module transfers finished documents out of the system into archive systems. In addition, a Verifier process allows human interaction providing correction and validation tasks. The whole process is driven by a *Document Manager* module.

It is generally accepted that human input is necessary during the conversion process to ensure the quality of the results of any automated processes. In the case of DocMining, The xmillum (XML illuminator) framework is used for cooperative and interactive analysis of documents [Hitz, 00]. In the same mindset, smartFIX caters for human agents, called verifiers, who provide services such as the manual completion of missing data.

Apart from this kind of technical human input, there is a further stage in the conversion process where human input of a different kind is desirable. The efficient conversion of any document collection into an electronic form is usually impossible without some external knowledge related to the structure of the collection and its specific knowledge domain. Therefore, the underlying concept for the efficient conversion of document collections should be the integration of two sources of knowledge, namely the documents themselves, but also any expert practitioners that possess relevant knowledge to the interpretation of the collection, or otherwise external knowledge complementary to the documents' contents.

The main function of expert practitioners is the provision of semantic information for the documents that would aid their analysis and interpretation. The involvement of expert practitioners in the process is not a novel idea, but it has been put in practice in the past in various forms: from one-off information provision to active involvement in the conversion process.

In the particular case of historical documents a prominent example has been the EU project MEMORIAL, where historians were asked to actively participate along with archivists to the conversion process [Antonacopoulos, 04a]. Historians were of key importance for the particular case of MEMORIAL, as the topic at hand was 2nd World War documents pertinent to concentration camps.

In another example, the conversion of specimen cards residing at the British Natural History museum was informed by the curators of the museum, who were asked to make an one-off contribution in terms of a template describing the logical structure of the specimen cards [Downton, 07].

Finally, document visualization and browsing is also an important task in relation to human interaction activities. Although each collection has different requirements in terms of visualisation and browsing, a few generic approaches have been suggested in the literature. As an example, the Multivalent Browser [Phelps, 01] is an architecture that allows creating different browsers for different kinds of documents. To do that the system comprises three main parts: a flexible document tree data structure, a set of extensions to add functionalities and an extensions manager that allows defining a set of functionalities specific to each document type. These three parts along with an easy way to define a browser by means of an XML file makes this system a good basis for building ad-hoc document visualisation and browsing interfaces.

3 Architecture for the Conversion and Annotation Of Information in Documents (ArCAnOID)

The Architecture for the Conversion and Annotation Of Information in Documents (ArCAnOID) presented here assumes that a semi-automatic process is followed for the conversion of the document collection. The expected result should be extracted knowledge in terms of physical and logical document structure, and the contents of documents expressed in *concepts*, *terms* and *relationships* within a given knowledge domain specific to the document collection and supplied by expert practitioners.

The ArCAnOID follows a modular paradigm and addresses the following issues:

i. Specifies the functionality of the individual conversion modules
ii. Defines how the modules are interconnected

iii. Describes how data propagate between the modules
iv. Explains how modules can be specialised to tackle application-specific tasks.

The flexibility a modular paradigm offers is a particularly desirable property, as the architecture needs to be adaptable to a range of distinctly different document collections both in terms of their physical structure and organisation as well as their information content.

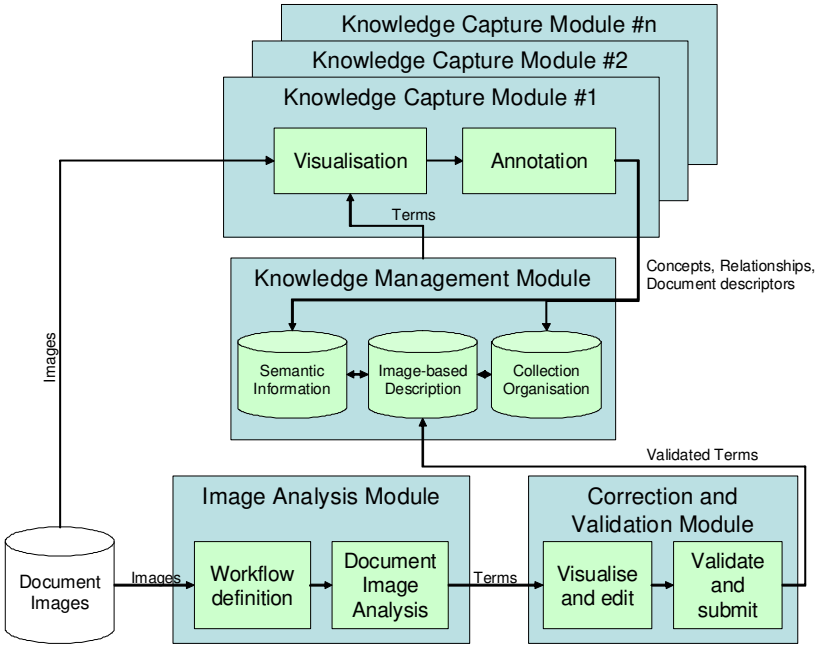
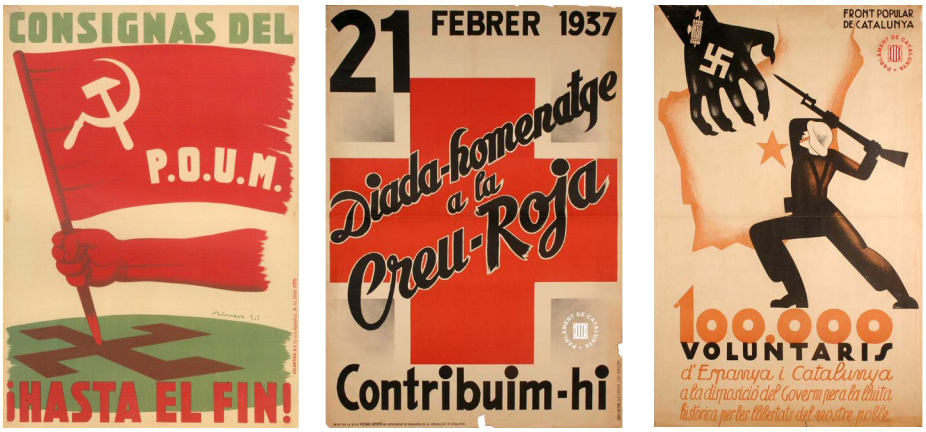


Figure 1: Overview of the Document Collection Conversion Architecture

An overview of the architecture proposed here is shown in Figure 1. The four modules comprising the architecture address the main groups of processes essential to the conversion process. Specifically there are two instances of human input during the conversion process. The Image Analysis module and the Correction and Validation module cover the initial conversion stages where the input required is of a technical nature, and involves the definition and supervision of the automatic part of the conversion process. The focus here is on the extraction of information from the document images, and the definition of the *terms*, the image-based units of information, that will be subsequently used to define higher level, semantically important *concepts*. Human input here could be completely avoided, assuming that the Image Analysis module achieves perfect results. In reality, the state of the art in document analysis methods does not render this a realistic option.

The second stage where human input is required is the process of the semantic annotation of the documents, where the extracted terms give rise to meaningful *concepts* and their *relationships*, and other external knowledge (not explicit in the document images) can be deposited in the digital archive. While the first stage is principally of a technical nature, this second stage requires information specific to the particular knowledge domain and generally calls for expert practitioners. These two principal conversion stages are coupled with a Knowledge Management module which provides the means to organise the extracted and acquired knowledge.

In terms of data propagation, the architecture follows a bottom-up process, starting with document images and progressively building meaningful *concepts*. The four modules of the architecture are described in detail next. To facilitate the description, we will employ here a running example with historical and social overtones. The images presented in Figure 2 are scanned propaganda posters from the collection of posters of the Spanish civil war, kept by the Catalan Government. We will use these images to illustrate the process of building up semantically important information from such a collection using the ArCAnOID.



(a) (b) (c)

Figure 2: Poster images from the collection of propaganda posters of the Spanish civil war.

3.1 Image Analysis Module (IA)

The Image Analysis module (IA) is responsible for performing automatic analysis of the document images in a batch fashion. The basic functionality of the IA module is the following:

- Provide interfaces for the definition of the workflow of automatic processes
- Classify document images into separate processing batches
- Invoke automatic processes in a batch fashion

Depending on the document collection, the IA module may or may not necessitate human input for the continuous monitoring of the process. Typically

though this is considered to be a define-once run-many type of module, where the series of algorithms to be applied are defined once and then run in a batch fashion throughout the collection. An obvious improvement is the categorisation of the documents into different processing classes, in terms of their specific processing needs. Again this might be a human-assisted or completely automatic process.

In the running example set above, this module could comprise colour segmentation, region classification and further region-based processing depending on the identified region type (e.g. OCR for text regions, symbol spotting for graphical areas etc). Thus, for the images of Figure 2 a conceivable output of this module would be physical structure data, recognised text and graphics.

It is important to note that the data produced by the IA module should not be expected to be perfect in any sense, as this module follows largely an automated processing scheme. The trade-off here is between accuracy and mass processing, and the design choice is towards the latter. The subsequent Correction and Validation module, as we will see next, is responsible for ensuring the accuracy of the results by building the first meaningful structures from the output produced by the IA module.

3.2 Correction and Validation Module (CV)

The Correction and Validation module (CV) is responsible for advancing the conversion process from mass data of questionable quality provided by automatic processes, to high-quality validated image-based data. The basic functionality of the CV module is the following:

- Provide interfaces to correct the results produced by the IA module
- Provide quality feedback to the IA module
- Allow the execution of specific analysis algorithms on demand
- Construct a description of each document image in the form of a collection of validated *terms*

The main issue at this conversion stage is to describe all the content of a document image as a set of descriptive *terms*. By *terms* we refer to image-based units of information, such as connected components, labelled page regions, segmented words or graphics etc. No attempt is made at this point to interpret the content of the document images; the objective is to obtain a complete, although still image-based, description of the document.

The exact definition of the *terms* will depend on the collection at hand, while more than one type of *terms* might be necessary. For the running example of the propaganda posters, a good choice for descriptive *terms* would be segmented regions corresponding to (a) single words and (b) graphical symbols, as these would be the most descriptive image-based units of information that we can use to fully describe such documents.

The automatic processing performed by the IA module should aim into building such *terms* automatically, but, as explained before, due the nature of the IA module we should not assume a perfect output. The minimum requirement of the CV module is the validation of the IA produced set of *terms*, while more often than not, some degree of correcting and editing will be necessary. The CV module is the first instance when significant human input is introduced in the conversion process. The

human input at this stage is of a technical nature, meaning that the user is not supposed to have any expert knowledge on the collection.

The functionality of the CV module should be implemented by providing interfaces to assist the user and accelerate the process. The synergies between the IA and the CV modules are important, as this stage could be a potential bottleneck of the conversion process. For example, it has to be anticipated that the user might require running selected image analysis algorithms on-demand for parts of the image (e.g. OCR on a newly segmented paragraph, or symbol recognition on a newly segmented graphical area).

Many different types of interfaces can be considered for the implementation of the CV module depending on the specific needs, including pen-based and voice-driven interfaces. This particular module can also take advantage of distributed proofreading approaches, for example the "Distributed Proofreaders" initiative [PGDP, 08] that work closely with the Project Gutenberg [Gutenberg, 08]. In the case of historical document collections, where commercial OCR is generally unreliable, reCAPTCHA technology is an alternative used to achieve fast transcription of huge document collections [von Ahn, 08]. The idea behind this is to show segmented words to online users packaged as a CAPTCHA test, and require the user to provide the word transcription as the test's solution.

3.3 Knowledge Capture Modules (KC)

Up to this point in the conversion process, all the potential image-based information has been identified within the document images, and structured as a set of *terms*. No specific knowledge about the context of the document collection has been assumed up to now. Nevertheless, the description of a document as a set of *terms* is only useful in order to give rise to its interpretation. A second stage of important human input is now due, aiming to capture external knowledge into the digital archive. The Knowledge Capture modules (KC) provide interfaces for expert users to deposit their domain specific knowledge into the digital archive. The basic functionality of a KCmodule is summarized below.

- Provide interfaces to define *concepts* building upon existing document *terms* when this is possible
- Provide the means to define *relationships* between *concepts* and qualify them by using document *terms* when this is possible
- Provide interfaces to introduce other external knowledge, that does not emanate from the document image, into the document collection

There are two kinds of external knowledge that generally need to be captured into the archive. The first kind is knowledge that can be directly derived from the document images. As such it relates to the interpretation of the contents of a document image, and can be seen as attaching a meaning to individual *terms* or combinations of *terms*. We call *concept* any higher level piece of semantic information that can be built upon existing *terms*. In the example of the propaganda poster collection, such *concepts* might be governing systems (e.g. democracy, communism, fascism, anarchism), political parties, social groups (e.g. workers, soldiers, politicians) etc. *Concepts* stem from the existence of certain *terms* in the document such as the acronym of a political party or a symbol with certain

connotations given the knowledge domain (e.g. the swastika, the sickle and hammer or the red cross in the documents of Figure 2).

Moreover, in the process of interpreting a document image, it is often necessary to indicate how certain *concepts* relate to each other. As an example a specific political party might be promoting a specific governing system, while actively opposing another one. Such *relationships* could also be indicated, and qualified if necessary, by specific document *terms*. For the example of Figure 2(a) the opposing *relationship* between the communist party of "P.O.U.M." and fascism emanates from the graphical representation of the communist flag stabbing the swastika. This combination of graphical areas provides the qualifying *terms* for the above *relationship*.

The second kind of external knowledge is external knowledge that lies with expert practitioners but cannot be derived directly from the contents of a document image. For example a *relationship* between a political party and a governing system might be known to a political scientist, but not emerging by any *terms* in the document image. Another example, stemming from the image of the paradigm, is the meaning of the acronym "P.O.U.M.", standing for Partido Obrero de Unificación Marxista, (Workers Party of Marxist Unification) or for that matter, even the fact that it is a political party in the first place. In the same category of external knowledge not directly derived from the image falls information such as the artistic style of a work of art, the period it represents, archival information for a document etc.

The KC modules cover a family of specialised modules that provide interfaces for expert practitioners. Typically, more than one type of KC module would be necessary for the semantic annotation of a document collection. In the running example above, KC modules would be required for historians, social scientists and archivists, possibly exposing different interfaces aiming at capturing information covering different knowledge domains.

The KC modules encode the captured knowledge in terms of *concepts*, *relationships*, archival information and arbitrary notes, and should establish links to the image-based *terms* when this is possible. In a similar fashion to the previous modules, different interface paradigms may be considered for the implementation of KC modules.

Archival information is a special case for which a KC module could be used. Contrary to other knowledge domains, in the context of archival information certain industry standards exist that could be implemented. For example, the Open Archival Information System (OAIS) ISO standard addresses a full range of archival information preservation functions [OAIS, 02]. In the case OAIS compatibility is sought, a specific implementation of the KC module could provide the interface for the submission of OAIS Submission Information Packages (SIP).

3.4 Knowledge Management Module (KM)

The desired result of the conversion effort, as defined before, should be extracted knowledge in terms of physical and logical document structure, and the contents of documents expressed in *concepts*, *terms* and *relationships* within a given knowledge domain specific to the document collection and supplied by expert practitioners. This extracted knowledge should be stored and managed in an efficient manner. The

Knowledge Management module (KM) provides the knowledge management backbone of the ArCAnOID. The basic functionality of KM is summarized below.

- Maintain organizational information (e.g., archival data) for the collection
- Support a *term* based description of the collection at the page level
- Define and store higher level representations in the form of *concepts* and *relationships*
- Allow the explicit linking of *concepts* and *relationships* with existing *terms*

As the extracted knowledge differs greatly between collections, it is important for the KM module to be designed so that it can be easily adapted to the particular requirements. To implement the functionality described above the KM module is structured as three self-contained information repositories. These information repositories and their connections are specified in Figure 3. The three repositories reflect the different types of information obtained throughout the document conversion process, and could be individually implemented using different technologies if necessary.

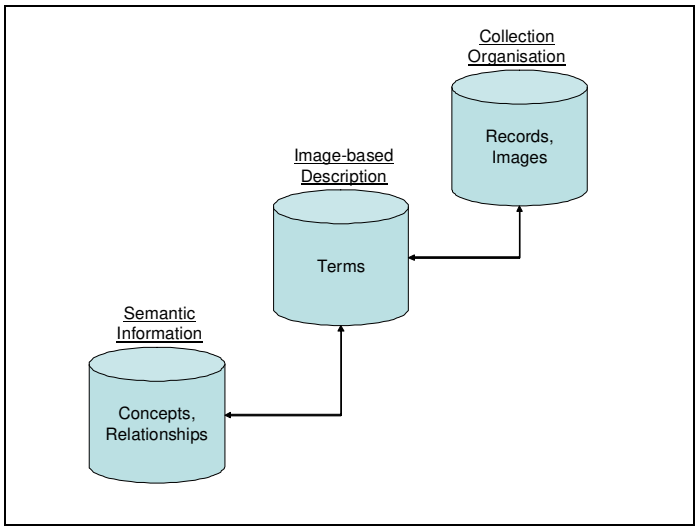


Figure 3: Structure of the Knowledge Management module

The archival metadata and other information related to the physical organisation of the collection are managed within the *Collection Organisation* repository. The type of organisational information stored here varies depending on the collection. In the running example of the posters collection, such organisational information could reflect their physical storage in the archive, or provide thematic or date groupings. The only design requirement imposed by the ArCAnOID is that the KM module should be able to expose organisational information at the level of individual images. This is so that extracted *terms* can be linked to the organisational information at the image level. Similarly to the KC module, this repository of the KM module could be

implemented based on the OAIS standard, if such compatibility is sought, by structuring submitted information as Archival Information Packages (AIP) [OAIS, 02].

The second information repository: *Image-based Description*, deals with the representation and storage of *terms*. As a reminder, *terms* are image-based units of information defined as required by the nature of the document collection. The minimum amount of information that the ArCAnOID requires to be stored is the geometric description of each *term*. The geometric description could be bounding boxes, connected components or pixel-level information, depending on the nature of the *terms* selected to describe the specific collection. For the running example of the posters collection, the *terms* chosen were segmented regions corresponding to individual words and graphical symbols. These could be described as bounding boxes or polygonal regions.

Finally, the *Semantic Information* repository manages all the higher level-constructs based on combinations of individual *terms*, namely: *concepts* and their *relationships*.

As mentioned before, each repository can be implemented in a different way. In terms of archival information, library standards such as the OAIS ISO standard can be employed. The *Image-based Description* and *Semantic Information* repositories can be implemented based on relational databases, XML etc depending on the application requirements. Generally the structure of the KM module is defined once before the conversion process and gradually filled in with information.

4 Case Study: The Archive of Border Records of Girona

The main purpose of this paper is the introduction of the ArCAnOID as a generic solution to the problem of the conversion of document collections to semantically annotated electronic data. It is not within the scope of this paper to examine various specific implementations of each of the four modules, as these depend on the characteristics of each particular collection. Nevertheless, for completeness it is deemed necessary to present an example to illustrate how the ArCAnOID can be implemented to address a particular scenario.

In this section we employ a collection of typewritten documents, the “Expedients de Frontera” from the Historical Archive of Girona, Spain, and we demonstrate how each of the ArCAnOID modules can be implemented to create a system for the conversion of the archive documents.

The “Expedients de Frontera” is a document collection of particular social and historical relevance, as it comprises records and documents related to people crossing the Spanish-French border following the Spanish Civil War. It consists of 93 meters of printed and handwritten documents collected between 1940 and 1976. The documents are physically organised in personal bundles which are in turn kept in boxes, and the archive possesses more than 800 such boxes of documents. For each bundle there is a cover page with the names of people whose information is contained in this record (an older attempt to manually organise this information). Each bundle contains diverse documentation, e.g., safe-conducts to cross the border, arrest reports, information about professional activities, prisoner transfers to labour camps, medical reports, correspondence with consulates, telegrams etc. This documentation is of great

importance in the studies of historical issues related with the Spanish Civil War and the 2nd World War. Figure 4 shows some sample document images. The collection we were presented with was already scanned and stored in binary raw image files.

Contrary to libraries, archives usually contain collections of information that is related in some way. Usually, an archive consists of a number of document categories (e.g. forms, letters, certificates, etc.) with hundreds of instances belonging to each one. In the process of converting an archive to a digital format, the extraction of metadata usually involves repetitive annotation processes (names, cities, dates, events, etc.). Expert practitioners are often the final users that make specific queries after the metadata on image contents has been captured both by semi-automatic image analysis processes and manual annotations.

(a) (b) (c)

Figure 4: Sample images of the target collection: (a) Cover page of a record (b, c) Some contained documents

From the digital archive in question, the "Expedients de Frontera" collection, what is of interest to extract information regarding people (names, nationality, age, civil status, dates, etc), events they participated in (arrests, interrogations, expatriations, etc) and place names related to the above. People, events and places will therefore constitute the *concepts* we are interested to define for this particular document collection, which can subsequently used for indexing purposes.

Relationships in this example refer to semantic links between the *concepts* defined above. For example, someone arrested in a city would indicate *relationships* between a person, an event and a place.

Since virtually all the information that lies in the document collection is expressed as text, we have chosen in this particular scenario to define regions that correspond to individual words as the image-based *terms* based on which we aim to describe each document image.

Based on these definitions, the specific implementation of each of the four modules of the ArCAnOID is described in detail below.

4.1 Image Analysis Module

Due to the special organisation of the collection, we are presented with specific opportunities regarding the implementation of the Image Analysis module. Specifically, since each bundle of documents contains a cover page listing the names of the persons appearing in the documents of the bundle, it is possible to process these cover pages separately, and make use of the results to facilitate the analysis of the rest of the documents in the bundle.

The main workflow is illustrated in Figure 5. First, cover pages are segmented at the level of words, and these words are used as keywords in a word spotting process aimed to segment similar words in the rest of the documents in the bundle. Word spotting is performed both the image level (using the segmented words' regions as image queries) and the OCR level (searching for the recognised words within the OCR results obtained for the documents). Specifically for the OCR based word spotting method, a dictionary of pre-defined words which carry special semantic value is also employed. This dictionary comprises 170 words deemed important, and contains place names, words indicative of personal relations (e.g. "son", "father"), nationalities, words indicative of specific events (e.g. "arrested", "imprisoned") etc.

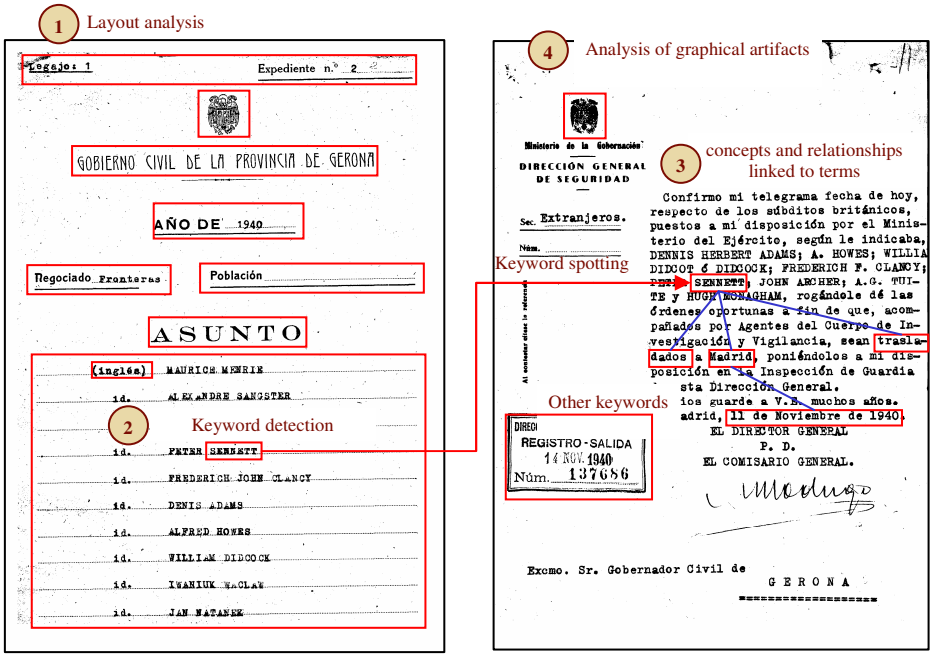


Figure 5: Process workflow

More than just convenient, such an approach is necessary, as most of the subsequent bundle pages have complex layouts with overlapping text and graphics and generally suffer from noise and other artefacts. As a result classic DIA methods alone are not adequate to extract accurate information.

In addition to word spotting, graphical symbols (stamps, logos, etc) are also segmented for image categorization and indexation purposes. Let us briefly describe the word spotting process. For further details the reader is referred to [Lladós, 07].

4.1.1 Layout analysis and key word detection

This process is applied to each cover page of the collection. Figure 4 (a) shows an example. Cover pages have a regular layout. A list of names and nationalities appears in a form-like region in the bottom part of the document. The aim of this process is to segment these names automatically to spot them in the rest of the pages of the record, or even in the rest of the collection. Names have been filled in the form using a typewriter; thus they appear underlined by dotted horizontal lines. To detect long horizontal lines, the classic Hough Transform (HT) is applied. Once the position of words is detected, underlines are removed and characters are reconstructed using a specific line removal algorithm based on run length analysis. Since the collection was scanned in binary raw images using a global binarization process, some words are partially deleted or present stains. A median filter is hence applied to remove noise. The final cropped word images are used as shape models in the spotting process.

4.1.2 Word spotting

The process of *word spotting* aims to locate query keywords in document images. Image regions where the searched keywords appear are the *terms* defined according to our proposed implementation. Two strategies exist. First, transcribing the document using OCR and a subsequent inexact string search process to locate the keyword. Nevertheless, in some documents commercial OCR engines are not able to produce good results. This is due to heterogeneous layouts (i.e. text areas overlapping with graphics areas), or due to document aging that results in degraded and noisy images.

To address such cases, some authors have developed word spotting techniques to search keywords at the image level. Keywords are first modelled with shape signatures in terms of image features. The detection of a keyword in a document image is then done by a cross-correlation approach between the prototype signature and signatures extracted from the target document image. A number of contributions exist in the literature on word spotting methods both for typewritten documents [Kuo, 94, Lu, 08], and handwritten documents [Rath, 03, Tomai, 02].

We apply here a combined strategy, namely word spotting at both the text level and the image level. During the digitization process, all the documents of the corpus were OCRed using an off-the-shelf engine. The recognition rate depends on the kind of document: with handwritten letters the OCR does not detect any word, while in typewritten documents the rate on misrecognized words is 20%. Thus, keywords are first searched in an alphanumerical mode by means of an inexact substring matching approach. To model the character edit errors the Levehnstein distance is used.

In addition to it, and to overcome the text-level errors, a second word spotting approach has been developed at the image-level. A compact shape signature inspired by the well-known shape context descriptor [Belongie, 02] is used to represent keyword images. Skeleton points are taken as feature points. Any given shape context h_i of a feature point p_i, can be converted to a bit vector representation b_i by binarizing h_i using a threshold T experimentally set. This bit vector of a shape context is called

the *codeword* of the point p_i. Since our shape contexts are organized in 24 zones, distributed in three concentric rings, each codeword is a 24-bit number (see Figure 6). For the sake of compactness, the codeword is split in three codes of 8 bits each $b_i = (b_i^1, b_i^2, b_i^3)$. These codes are used as hashing keys to index in a look up table that generates votes in a set of bins defined over the image. Those bins having a high number of votes are zones likely to contain the queried keyword.

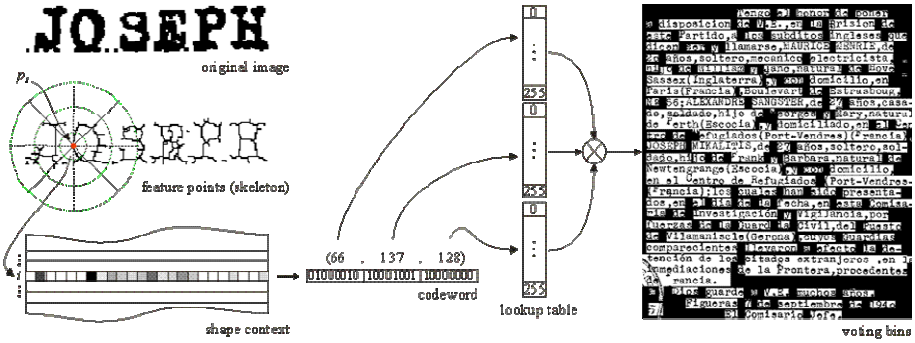
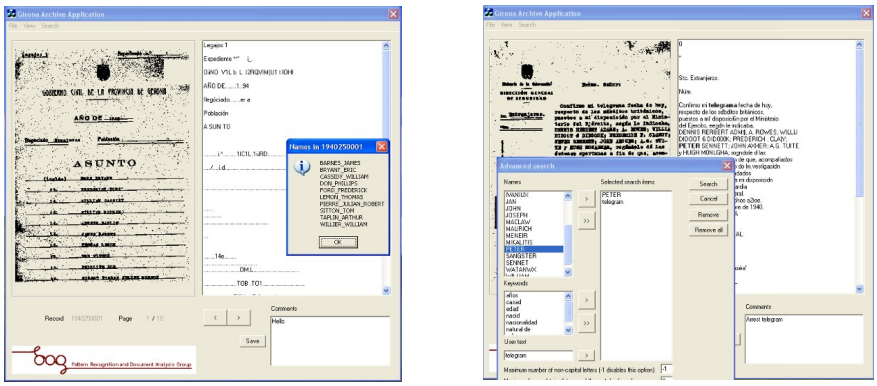


Figure 6: Codeword indexation scheme

4.2 Correction and Validation Module

In our case-study, we have developed a simple front-end module that allows the execution of specific image analysis algorithms on demand such as segmenting cover pages, OCRing, word spotting, etc. The user is therefore able to define new processing workflows using this interface if necessary. More importantly, the interface allows the user to correct and validate the image analysis results.



(a) (b)

Figure 7: Two snapshots of the correction and validation module

This module is illustrated in the screenshots showed in Figure 7. Figure 7 (a) shows the process of extracting words from cover pages. The list of extracted *terms*

after being validated by the user, are saved in a database described next in section 4.4. The user is also able to type in other keywords and execute a word spotting task at OCR level as illustrated the Figure 7 (b). The text locations where the query words are found are highlighted. Once the user validates these results, the corresponding *terms* are stored into the knowledge database.

It should be noted that in terms of validation the sketching interface described in the following section, in addition to being used for knowledge capture, is also used to perform some validation actions.

4.3 Knowledge Capture Module

This module is the part of the system dedicated to the experts. They will provide the necessary knowledge introduce *concepts* and *relationships* and link them to existing *terms* where possible. In our case of study the experts are mainly historians. They typically start from *terms* that represent segmented words in the documents and create *concepts* describing persons (names, age, nationality…), places (countries, prisons, etc) and events (interrogation, detention, etc) and the possible relationships among these concepts. We may find relationships between persons (married, child of, etc), between persons and events (being arrested, being interrogated, etc) and between places and events (arrested in, interrogated in,).

In order to introduce this knowledge in the collection of documents we have developed a sketch based interface. This kind of interfaces use a pen input device to interact with the system. Here, the user draws a set of gestures to navigate, annotate, etc in a collection instead of using specific commands or menu items. A gesture may be seen as a set of strokes with an associated meaning. A stroke in this case is defined as the set of points captured by the device between a pen-down and pen-up user action. In our application we have defined three categories of gestures depending on the associated meaning: Browsing/Navigation, Knowledge Capture and Relationship. The first category describes the set of gestures that allow the user to navigate in the collection. In this category we defined gestures to manage the zoom level of a page and go forwards or backwards in the collection. The Knowledge Capture category contains the set of gestures that allow the user to deposit his own knowledge in the collection, built knowledge based on the document terms or search for relevant information in the rest of the collection as a means to help him perform his task. Finally, the Relationship category contains gestures that allow the user to relate different *terms* to *concepts* or instances of *concepts* to each other. A graphical description of the gestures may be seen in Figure 8.

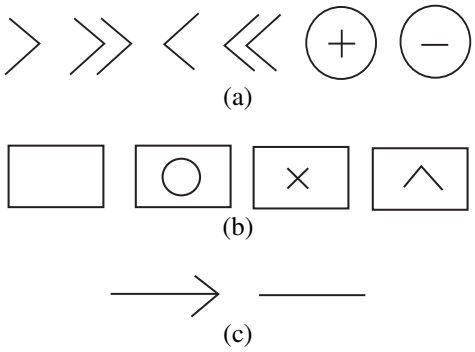


Figure 8: Application Gestures Sets: a) Browsing set, b) Knowledge Capturing set and c) Relationship set.

A schema of the application is shown in Figure 9, we may distinguish three different components. The core component is the one dedicated to visualize the documents in the collection and to capture the gestures drawn by the user. The component sends the gesture to the recognition component and obtains the class of the gesture, then sends this information to the action interpreter component to raise the adequate action. The pre-processing and gesture classification component contains the different recognition techniques. Depending on the technique the gestures will be pre-processed in a specific way. This component obtains the corresponding class of gesture. Finally, the application raises the corresponding action to the gesture drawn by the user. The application is connected both to the database that contains the knowledge extracted from the documents and to the repository of the document images. For more details on the gesture recognition system the interested reader is referred to [Rodriguez, 07, Mas, 06].

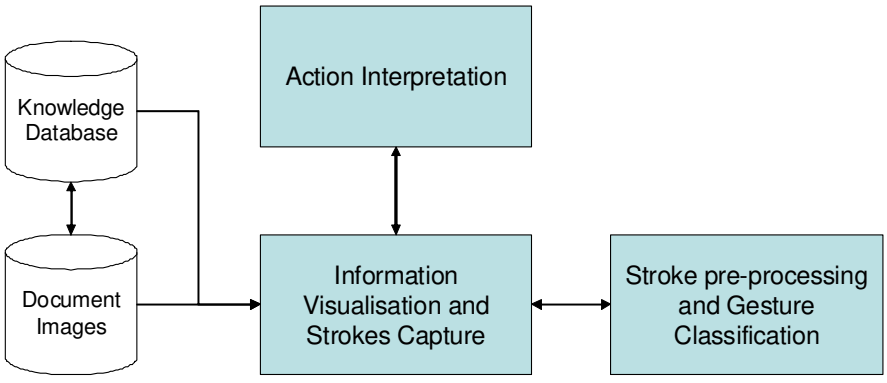
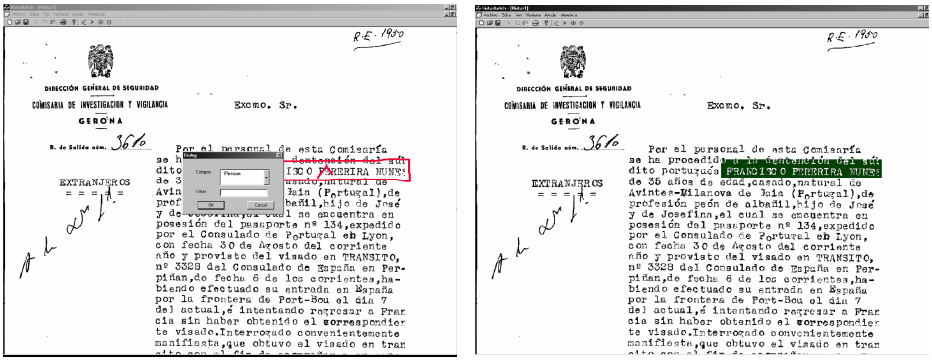


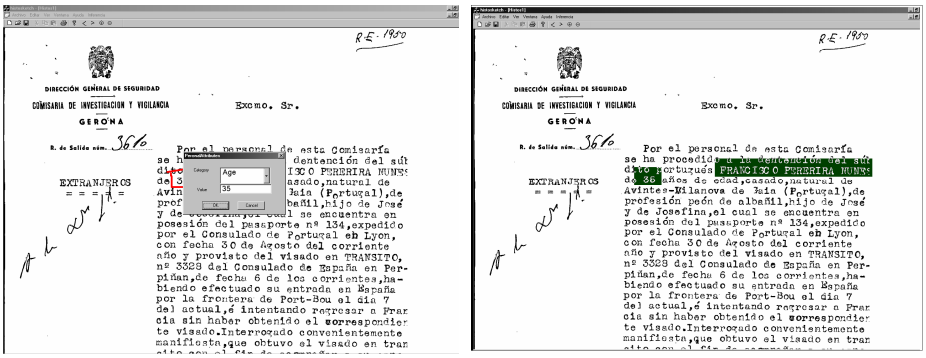
Figure 9: Application Schema.

A series of figures below present the steps that have been followed by an expert when extracting the knowledge from a document. First, the expert has selected the set of *terms* (segmented words in this case) corresponding to the name of a person ("Francisco Pereira Nunes"), see Figure 10, and created a *concept* (a new person in this case). Then he selected the *term* "35" and added this information to the previously created person as an attribute – see Figure 11. Then, a new *term* is selected and used to create a second person and the expert creates a relationship between the two persons by linking with a gesture the two annotated zones – see Figure 12.



(a) (b)

Figure 10: The process of combining terms (segmented words) into higher level concepts (persons)



(a) (b)

Figure 11: Adding information to existing concepts by linking them to terms in the document

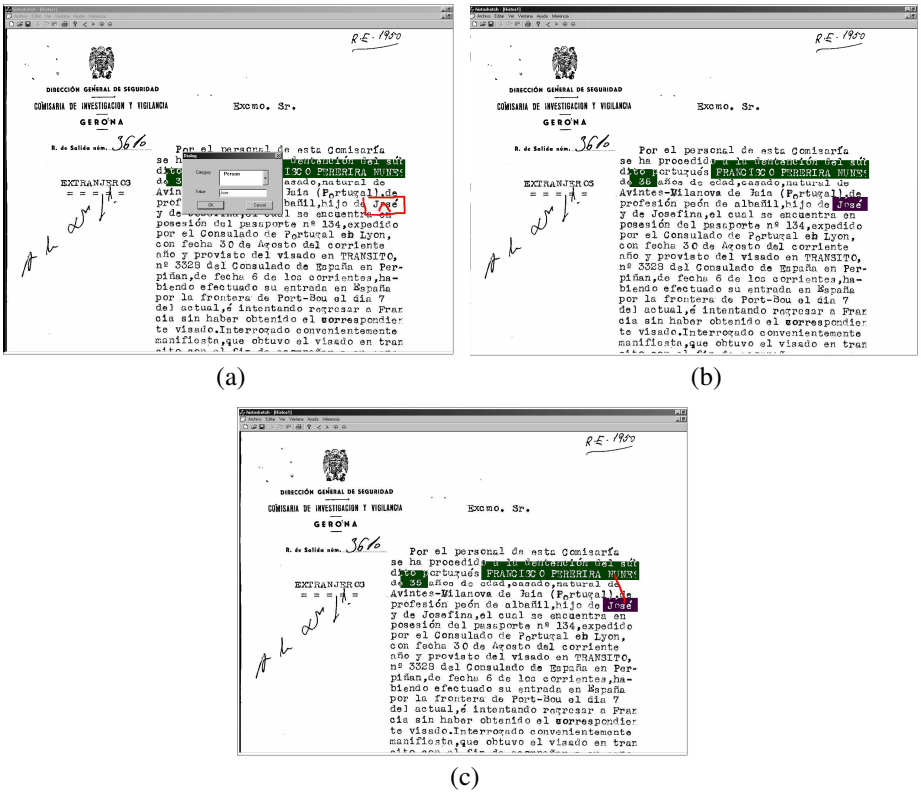


Figure 12: Defining relationships between concepts using the gesture-based interface

4.4 Knowledge Management Module

The KM module for this particular application was implemented as a relational database. An overview of the database can be seen in Figure 13 below. Some of the fields and intermediary tables have been omitted from the figure for the sake of clarity.

The three information repositories are implemented here as self-contained parts of the database. The Collection Organisation repository reflects the physical organisation of the archive into records, each of which contains a number of documents consisting of individual pages. Each page corresponds to a scanned document image, and supplies the link between Collection Organisation and Image-based Description.

Image-based Description is here achieved by expressing each page as a set of words. Words are chosen as *terms* in this case, as they can describe completely and accurately each typewritten document page of the collection. Each word in our case is specified by its bounding box and its OCR transcription if available.

Words give rise to higher-level semantic information. As explained before, in the case of the "Expedients de Frontera" archive, the information of interest relates to specific people identified in the documents, the events they participated into, the

location where they took place and any relations between different persons identified in the archive. As a result, the *concepts* selected here are Persons, Events and Places, and a number of *relationships* are built around them (Person to Person, Person to Event, Event to Place). Each of these *concepts* and *relationships* are linked to individual *terms*.

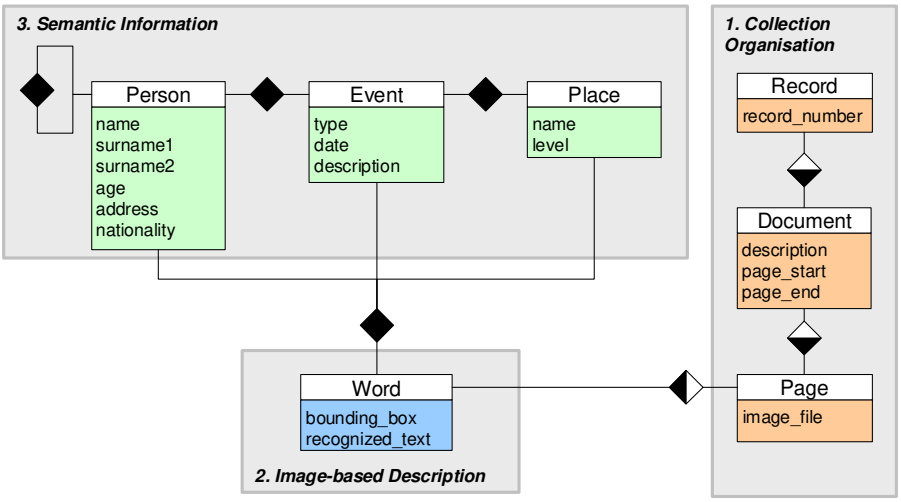


Figure 13: The relational database implementation of the KM module

The *terms* (segmented words in our case) can be linked to each of the higher-level *concepts* (Persons, Events, Places). But at the same time it is quite possible that some instances of *concepts* (for example a person) might be introduced directly from the historians examining the collection without them being explicitly mentioned in the documents. In a similar fashion some related information (for example the nationality of a person) might be introduced by the historian without an explicit link to a specific word in the documents. For that reason, the real information about each *concept* is stored in the related table, and is independently supported (or not) by existing document *terms*.

Finally, the relational database based implementation of the KM module is making use of a number of pre-defined enumerations, in order for the information to be more easily indexed and retrieved. For example specific types are pre-defined for events (e.g. arrest, transportation, imprisonment etc), nationalities are enumerated and so on.

5 Conclusions

In this paper we have addressed the problem of converting high volumes of document collections to Digital Libraries semantically enriched. We have presented the Architecture for the Conversion and Annotation Of Information in Documents

(ArCAnOID), a modular paradigm that assumes a semi-automatic conversion and knowledge capturing process. We have proposed a flexible architecture adaptable to the heterogeneity of document collections and also the domains were the conversion task is performed. According to such a wide range of document types and contexts, we have given to the human intervention a key role in the process. Firstly, at a technical level, when document images are processed by document image analysis and understanding techniques, the expertise is necessary for a human operator to invoke automatic processes, visualise, check and validate their results. Secondly, the ArCAnOID addresses the necessity to integrate external knowledge to the archive and suggests specific ways to achieve it. We have proposed the need of involving certain groups of expert practitioners in the conversion process to provide semantic annotations and otherwise deposit their knowledge about the collection in the digital archive. In addition to the components able to capture knowledge from the two stated sources of information, namely document images and humans; in the ArCAnOID we have defined a structured way to represent the information. Knowledge representation is scalable but structured around specific blocks that correlate to the different physical and informational aspects of the document collection. We have considered three information artefacts. First, *terms* have been defined as basic image units once it has been segmented somehow. Semantic knowledge, sometimes domain-dependent and not explicit in the document images, is represented by *concepts* and their *relationships*. Both, *concepts* and *relationships* are linked to terms.

To illustrate well the advantages of the generic framework described, and give a comprehensive example of how it could be implemented given a particular collection, we have described in Section 4 a complete application scenario. The scenario was devoted to the conversion of a collection of typewritten documents, the "Expedients de Frontera" from the Historical Archive of Girona, Spain. In such a scenario we have proposed a number of document image analysis and recognition components to extract civil information related to people, places, events, etc. A sketching interface allows completing the knowledge by semantic annotations, as well as running processing actions using a gesture alphabet.

We believe that ArCAnOID is general enough to be extrapolated to other application cases, not only in the case of libraries and archives, as the application example, but also in different business processes where massive analysis of mails, faxes, forms, invoices, etc. has to be done in efficient and fast way.

Acknowledgements

Work partially supported by the Spanish projects TIN2006-15694-C02-02 and CONSOLIDER INGENIO 2010 (CSD2007-00018), the fellowship 2006 BP-B1 00046 and the Subdirecció General d'Arxius de la Generalitat de Catalunya.

References

[Adam, 04] Adam, S., Rigamonti, M., Clavier, E., Trupin, E., Ogier, J.-M., Tombre, K., Gardes, J. : DocMining : A Document Analysis System Builder, Document Analysis Systems VI, S. Marinai and A. Dengel (Eds.), Springer Lecture Notes in Computer Science, LNCS 3163, pp. 472 – 483, 2004

[Antonacopoulos, 04a] Antonacopoulos, A., Karatzas, D., Krawczyk, H., Wiszniewski, B.: The Lifecycle of a Digital Historical Document: Structure and Content, Proceedings of the ACM Symposium on Document Engineering (DocEng2004), Milwaukee, Wisconsin, ACM Press, pp. 147 – 154, October 2004

[Antonacopoulos, 04b] Antonacopoulos, A., Karatzas, D.: A Complete Approach to the Conversion of Typewritten Historical Documents for Digital Archives, Document Analysis Systems VI, S. Marinai and A. Dengel (Eds.), Springer Lecture Notes in Computer Science, LNCS 3163, pp. 90 – 101, 2004

[Belongie, 02] Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(24):509–522, April 2002

[BVH, 08] Les Bibliothèques Virtuelles Humanistes (BHV), 2008, http://www.bvh.univ-tours.fr/

[Clavier, 03] Clavier, E., Masini, G., Delalandre, M., Rigamonti, M., Tombre, K., Gardes, J. : DocMining: A Cooperative Platform for Heterogeneous Document Interpretation According to User-Defined Scenarios, Graphics Recognition : Recent Advances and Perspectives, J. Llados and Y.-B. Kwon (Eds.), Springer Lecture Notes in Computer Science, LNCS 3088, pp. 13 – 24, 2003

[Dengel, 02] Dengel, A., Klein, B.: smartFIX: A Requirements-Driven System for Document Analysis and Understanding, , Document Analysis Systems V, D. Lopresti, J. Hu and R. Kashi (Eds.), Springer Lecture Notes in Computer Science, LNCS 2423, pp. 433 – 444, 2002

[Downton, 07] Downton, A., He, J., Lucas, S.: User-configurable OCR enhancement for online natural history archives, International Journal on Document Analysis and Recognition, Vol. 9, Nos. 2 – 4, pp. 263 – 279, April 2007

[Gutenberg, 08] The Project Gutenberg, 2008, http://www.gutenberg.org

[Hamza, 08] Hamza, H., Belaid, Y., Belaid, A., Chaudhuri, B.B.: An end-to-end Administrative Document Analysis System, In Proc. 8th IAPR Int. Workshop on Document Analysis Systems, pp. 175 – 182, September 2008

[Hitz, 00] Hitz, O., Robadey, L., Ingold, R.: An architecture for editing document recognition results using XML, In Proc. 4th IAPR International Workshop on Document Analysis Systems, Rio de Janeiro (Brazil), pp. 385 – 396, 2000

[IMPACT, 08] EU Project: Improving Access to Text (IMPACT), 2008, http://www.impact-project.eu

[Kuo, 94] Kuo, S., and Agazzi, O.: Keyword spotting in poorly printed documents using pseudo 2-D hidden markov models, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 16, No. 8, pp. 842 – 848, 1994

[Lladós, 07] Lladós, J., Sánchez, G.: Indexing historical documents by word shape signatures, In Proc. 9th Int. Conf. on Document Analysis and Recognition, pp. 362 – 366, 2007

[Lu, 08] Lu, S., Tan, C.: Retrieval of machine-printed latin documents through word shape coding, Pattern Recognition, Vol. 41, No. 5, pp. 1816 – 1826, 2008

[Mas, 06] Mas, J., Sanchez, G., Lladós, J.: An incremental parser to recognize diagram symbols and gestures represented by adjacency grammars, Graphics Recognition: Ten Year Review and Perspectives, Liu, J.L.W. (ed.), Lecture Notes in Computer Science, Vol. 3926, pp. 252 – 263, Springer-Verlag, 2006

[Mello, 02] Mello, C.A.B., Lins, R.D.: Generation of Images of Historical Documents by Composition, In Proc. 2002 ACM Symposium on Document Engineering, Virginia (USA), pp. 127 – 133, 2002

[OAIS, 02] CCSDS 650.0-B-1: Reference Model for an Open Archival Information System (OAIS), Blue Book (Standard), Issue 1, January 2002

[PGDP, 08] Distributed Proofreaders, 2008, http://www.pgdp.net

[Phelps, 01] Phelps, T.A. and Wilensky, R.: The multivalent browser: a platform for new ideas, In Proc. 2001 ACM Symposium on Document Engineering, pp. 58 - 67, Atlanta, Georgia, 2001

[Rath, 03] Rath, T., Manmatha, R.: Word image matching using dynamic time warping, In Proc. Conf. on Computer Vision and Pattern Recognition (CVPR), volume 2, pp. 521 – 527, Madison, WI, 2003

[Rodriguez, 07] Rodríguez, J.A., Sánchez, G., Lladós, J.: A pen-based interface for real-time document correction, In Proc. 9th Int. Conf. on Document Analysis and Recognition (ICDAR2007), Vol. 2, pp. 939 – 943, 2007

[Sánchez, 04] Sánchez, G. Valveny, E. Lladós, J. Mas, J. Lozano, N.: A platform to extract knowledge from graphic documents: Application to an architectural sketch understanding scenario, Document Analysis Systems VI, S. Marinai and A. Dengel (Eds.), Springer Lecture Notes in Computer Science, LNCS 3163, pp. 389 – 400, 2004

[Sinaiticus, 08] Codex Sinaiticus project, 2008, http://www.codex-sinaiticus.net

[Tomai, 02] Tomai, C., Zhang, B., Govindaraju, V.: Transcript mapping for historic handwritten document images, In Proc. 8th Int. Workshop on Frontiers in Handwriting Recognition, pp. 413 – 418, Ontario, Canada, 2002

[von Ahn, 08] von Ahn, L., Maurer, B., McMillen, C., Abraham, D., Blum, M.: reCAPTCHA: Human-Based Character Recognition via Web Security Measures, Science, Vol. 321, no. 5895, pp. 1465 – 1468, September 2008