Proceedings of the ACM Symposium on Document Engineering (DocEng'04), Milwaukee, USA, October 28-30, 2004, ACM Press, pp. 147-154.

The Lifecycle of a Digital Historical Document: Structure and Content

A. Antonacopoulos, D. Karatzas
Department of Computer Science
University of Liverpool
Liverpool, United Kingdom
http://www.csc.liv.ac.uk/~prima

H. Krawczyk, B. Wiszniewski
Faculty of Electronics, Telecomm/s and Informatics
Technical University of Gdańsk
Gdańsk, Poland
http://www.eti.pg.gda.pl

ABSTRACT
This paper describes the lifecycle of a digital historical document, from template-based structure definition through to content extraction from the scanned pages and its final reconstitution as an electronic document (combining content and semantic information) along with the tools that have been created to realise each stage in the lifecycle. The whole approach is described in the context of different types of typewritten documents relating to prisoners in World-War II concentration camps and is the result of a multinational collaboration under the MEMORIAL project funded (€1.5M) by the European Union (www.memorial-project.info). Extensive tests with historians/archivists and evaluation of the content extraction results indicate the superior performance of the whole semantics-driven approach both over manual transcription and over the semi-automated application of off-the-shelf OCR and the use of a conventional (text and layout) document format.

Categories and Subject Descriptors
H.3.7 [INFORMATION STORAGE AND RETRIEVAL]: Digital Libraries, I.7.1 [DOCUMENT AND TEXT PROCESSING]: Document and Text Editing --- Document management, I.7.5 [DOCUMENT AND TEXT PROCESSING]: Document Capture, I.5.4 [PATTERN RECOGNITION] Applications --- Text processing, Computer vision.

General Terms
Algorithms, Experimentation, Human Factors.

Keywords
Digital Libraries, Historical Documents, Document Engineering, Document Architecture, Text Enhancement, Document Analysis.

1. INTRODUCTION
Historical archives contain a multitude of paper-based documents documenting human decisions, actions and events. Diverse types of paper documents have been created and subsequently used by different types of readers who "enhanced" the document by marking, highlighting and annotating the text in some way which at the time made the document more legible (semantically) and functional to that reader. Modern archivists endeavor to index and search for information contained in these documents and historians attempt to reconstruct past events or to discover unknown facts by examining both the documents themselves and their semantic content. There is, therefore, a significant need to devise a suitable document architecture that supports these activities and to use it to represent the original paper documents in electronic form.

The conversion of collections of historical documents into digital archives or libraries raises a significant number of issues that are not usually encountered (at least not collectively) in the conversion of other types of documents. These include physical, semantic, structural, functional and legal issues.

Physical issues of converting old/historical manuscripts revolve around difficulties arising from the effects of ageing (document degradation) and imperfect production processes. Stains, tears and irregular accumulation of dirt (due to repeated handling) in addition to artefacts resulting from earlier attempts of physical restoration are examples of the former. Non-uniform appearance of characters of the same font is also frequently observed in machine printed documents – a large body of which (most of 20th century documents) is contained in archives.

Semantic issues in recovering and storing the information arise from the need of historians/archivists to have every foreground entity labelled. For instance, some text may actually be a person's name and, depending on which section of a given document type it appears, that name may belong to a man or a woman, a prisoner or military officer etc. Similarly, various annotations and marks (made by the document creator or subsequent readers) must be suitably identified, differentiated and labelled.

A significant issue in terms of structure is the requirement that a full representation of the document must be available at different levels. For instance, a researcher does need to examine (visually) a facsimile of the original paper document to study physical characteristics of the document and evidence of its use. On the other hand, another historian only requires that they have the information on the document displayed in the original layout (reconstructed from the recognised entities) while an archivist may only need to be able to search for certain terms in the document without any requirement for visual information from the original document.

Functional issues relate to the ability of the final (converted) document architecture to be used to reason about and recover missing/incomplete information from the original document and to be forward-compatible with future recognition tools that may yield better results than current methods. Towards the latter respect, for instance, pixel-based information on unrecognised or ambiguous characters is stored in anticipation of improved image analysis and OCR methods.

Finally, legal issues dictate the need for selective access to the document information depending on the type of user. For instance an authorised user may be able to perform a search on peoples' surnames to ascertain the presence or absence of an individual in a documented event but more privileged access (e.g. an actual list of people) may be given only to highly vetted individuals. Similarly, the full document information (including a visual representation) may only be available to the resident historian/curator of a given organisation.

It is evident, therefore, that a specialised document architecture is necessary to represent the rich structure and content of historical documents, as opposed to other types of documents.

This paper describes the lifecycle of the digital historical document, from template-based definition through to content extraction from the scanned pages and its final reconstitution as an electronic document (combining content and semantic information) along with the tools that have been created to realise each stage in the lifecycle. This paper builds on the earlier (preliminary) proposal of a document lifecycle model [2] and significantly augments it with the presentation of new developments and results of the complete template-based document processing system. The whole approach is described in the context of different types of typewritten documents relating to prisoners in World-War II concentration camps and is the result of a multinational collaboration under the MEMORIAL project funded (€1.5M) by the European Union (www.memorial-project.info).

The processes and input/output of each phase of the lifecycle of the digital historical document are described in the next section. The document architecture is presented in Section 3. An overview of the quality-driven processes that complete the document structure by recovering and validating the content from the original paper document is given in Section 4. A summary of the advantages of the semantics-driven approach is given in Section 5 and general concluding remarks are made in Section 6.

2. STEPWISE DOCUMENT ENGINEERING

The conversion of a historical paper document into an interactive electronic document is a complex multi-phased engineering process requiring a variety of specialised tools for image processing and recognition, interactive graphical and text editing, and quality monitoring and evaluation. The MEMORIAL project has introduced a *Digital Document Life-Cycle Development (DDLC)* model supported by a specially developed *Digital Document Workbench (DDW)* toolset. Resemblance of the DDLC model of document engineering to the well-known *V* model of software engineering (see Figure 1) is intended, and the rationale behind that will be evident throughout the rest of this paper.

The left arm of the DDLC model, like in the V model, represents analysis of information aided by the user, whose domain knowledge is gradually being transformed into a control structure of processes for engineering the final product, represented by the right arm. Similarly, in either model, verification of partial products of respective phases of the cycle plays a key role in assuring the quality of the final product. Quality management is an essential part of each individual phase, but the criteria for progressing between phases are also important; phases correspond to functionality of corresponding DDW tools, while criteria to quality evaluation of intermediate products.

Another important feature of the DDLC model is a supporting DDW toolset, enabling the incorporation of human intelligence in the recovery of machine-interpretable information contained in the scanned document. Naturally, it is generally not possible to totally automate this recovery process, owing to the ageing (degradation) and preservation state of the original paper document, as well as to the noise and other artefacts introduced to the images during scanning and as byproducts of the image processing methods later on. AI cannot deal yet with these problems for a realistically large class of documents, and attempts to develop algorithms dedicated to specific classes of historical documents are not practical if the class of interest is not large enough – the effort required to develop new algorithms may be unrealistically high, compared to the effort spent on direct reproduction of them in electronic form by human experts (historians) with a specialised editor.

Similarly, quality tuning is interactive and performed by an expert, as a trade-off between manual (current practice), and automatic tuning, which would be narrow in applicability and costly to develop.

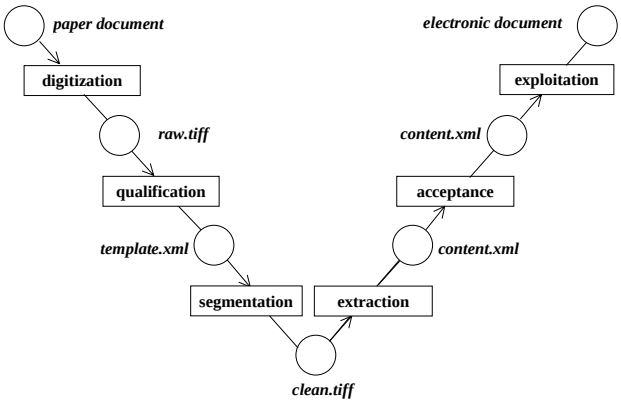

Fig. 1. Digital Document Life Cycle development model

The first phase of the DDLC model is *digitization*, which yields a raw digital image of a paper original. This process may be performed as an entirely manual activity, e.g., photographing extremely rare documents with a digital camera, as well as batch scanning of more sturdy documents (e.g. inventory cards of a museum) with an automatic document feeder scanner. The most important task of this phase is to assure proper scanning parameters, for otherwise if the quality of scanned images turns out to be unsatisfactory later on during the cycle, taking paper documents out from the archive to scan them again may be impossible or costly. Another problem is related to the naming of raw image TIFF files generated by the scanner – each generated file must have a unique name to avoid processing duplicates or overwriting (loosing) files during their processing later on. A

document Repository Management Tool (RMT) of the DDW toolset has been developed to aid the archivist in the raw image file namespace management.

The next phase of DDLC is *qualification*, when documents similar in structure, purpose and meaning are grouped into semantic classes. Examples of such classes include various types of transport lists (like the ones shown in Figures 4 and 6) and personal cards (historical documents of memorial archives), as well as index cards (post-war documents used by museums). Documents within the same semantic class can be processed throughout the rest of the cycle in a specific way, "tuned" individually for each class defined. This has been made possible by introducing two concepts: a *document template*, and *phase tuning*, explained later on

A template is an XML file, specifying formally the document layout and content in a form that is both machine readable and can be directly manipulated by an expert user knowing document semantics. It shall be noted that the distinction between classes is at the level of a single page, since a class template combines in a specific way both the page layout and the content of a page, as explained later. For example, the transport list pages in figures 4 and 6 constitute different classes, as the former is a complete one-page document, while the latter is a front page of a multi-page document. This distinction, however, is introduced at the archivist's discretion, who interprets a document and defines a template. A range of DDW component tools have been developed to operate on document templates, including a template Electronic Document eDitor (EDD) tool for generating template XML files. An intuitive graphical interface provides a fair separation of a document expert from XML intricacy.

The document template is used to control the *segmentation* phase, where the raw document image is cleaned and improved by the Image Processing Tool (IPT), transforming original TIFF files into binarised clean images.

A key phase of the DDLC is *extraction*, where the clean document image is processed by OCR, and a *document content* XML file is produced.

The following *acceptance* phase introduces again expert user interaction for two reasons. First, the content XML file generated by OCR may contain incorrectly recognized characters; therefore loading such an electronic document into the target database (digital archive) will reduce the quality of information available during the subsequent *exploitation* phase. Second, the original document may also contain errors, which automatic correction by OCR cannot be accepted by a historian, although it might be required to improve the quality of the results of queries to the target database. Typical examples for transport lists or personal cards mentioned before include incorrect dates (like a year of birth of a camp prisoner 1995 instead of 1895), as well as some geographical names misspelled by international prisoners working in camp administration. The correction of the document content in the first case requires modification of the content, and is supported by the content editor Generator of Electronic Documents (GED). In the second case only annotation is allowed, and is supported by the multivalent browser Viewer of Electronic Documents (VED).

The DDW component tools mentioned above are shown schematically in Figure 2. Editors EDD and GED, and browser VED are interactive tools, while document image handling tools RMT, IPT and OCR operate in an automatic mode. It is worth mentioning that the DDW can incorporate any OCR tool, provided that it can be controlled via an XML interface; in the current release DDW uses DOKuStar by OCE. On top of the DDW tools there is a tool for Quality Evaluation of electronic Documents (QED), not shown in Figure 2; it allows a document quality expert to tune the complex processes of DDLC phases for best performance, as described later in this paper.

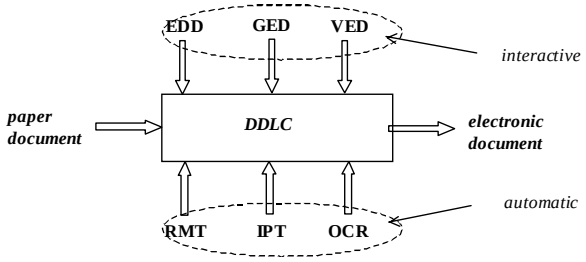

Fig. 2. DDW component tools supporting DDLC model.

3. DOCUMENT ARCHITECTURE

The document architecture is defined based on XML and plays a central role, as mentioned earlier, throughout the DDLC. The components of the document model constitute a tree, shown in Figure 3.

The document model describes the document in a top-down manner, providing separate descriptions for its background and its content. The content of the document is described as a set of rectangular regions, and line segments (since the latter contribute to the layout of the document). Regions are further specialised depending on their content into either text regions or images.

Text regions are used to represent the textual information in the document, whereas images are used to represent content entities such as signatures, stamps or handwritten notes, as will be explained later. There are two types of text regions defined: *composed text* and *tabular text*. Composed text defines textual regions in terms of text lines, which can be further broken down into their composing parts (combinations of typewritten text and/or inline handwritten notes). The tabular text on the other hand, provides the flexibility to define table layouts, which contain composed text in the nested rows and cells.

One of the important contributions of the proposed approach, however, is the semantic tagging of layout entities. To enable the semantic labelling of text regions, the document model provides a type attribute at the level of composed text in the XML tree, so that each piece of text can be assigned one of a list of pre-defined types of text (e.g. date, prisoner number, family name, first name, place of birth etc). It should be noted that these types of text can be associated either with a predefined format type (for example a date or prisoner number format) or with a higher level meaning (for example a family or geographical name that can be associated to specialist dictionaries). In the case that a specific piece of text is expected in the document (i.e. the header can be only one of a list of alternatives), lists of pre-defined strings can be compiled using tools of the DDW, and text lines can be linked with one of these lists, providing even more specific information about their

textual content. Image regions also convey semantic information, based on their arbitrary interpretation of a human expert using EDD.

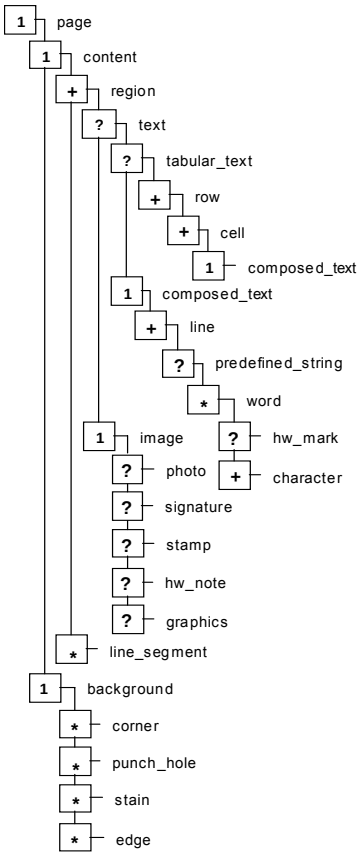

Fig. 3. Document model layout tree.

Two different examples of composed text are shown in Figure 4. The header is described in the document model as a composed text region, comprising of three lines. Since the contents of the header lines are specific to this type of documents, each text line is further specialized as a pre-defined string entity which links it with a list of pre-defined alternatives, created using the DDW tools. The table structure presents a second example of semantic label. Each cell is associated with a composed text entity (comprising of a single text line), the type attribute of which is used to label specific cells according to their content (date, name, etc).

The document architecture defined takes into account special characteristics of historical paper documents. First, it provides certain attributes at the level of individual text regions, to store information such as character spacing, font style and size (for machine-typed documents). Moreover, it also accommodates handwritten annotations as well as other semantically important non-textual entities, which often exist in such a document. To this end, a number of different *image* types are defined (e.g. signature, stamp, photo, handwritten notes etc) for content regions. An example of a signature is shown in Figure 4. Furthermore, the architecture has the capability to represent degradation artefacts, using various background entity tags, (such as stains, punch holes, edges etc) which may be of further interest. An example of a punch hole is shown in Figure 4.

Finally, the document architecture is flexible in representing unknown document types. The existing XML structure is sufficient to describe other classes of machine typed documents, (which comprise a significant proportion of the paper documents produced in the 20th century). The extendibility of the document model is ensured by providing tools (through the DDW) to define more (semantic) types for text regions, as well as additional lists of pre-defined strings. The flexibility achieved by the document model presented in Figure 3 is particularly important for the forward compatibility of documents, which may be analyzed with future generations of information extraction tools, capable of recognizing, for instance, text overlapping on handwritten notes, or signatures.

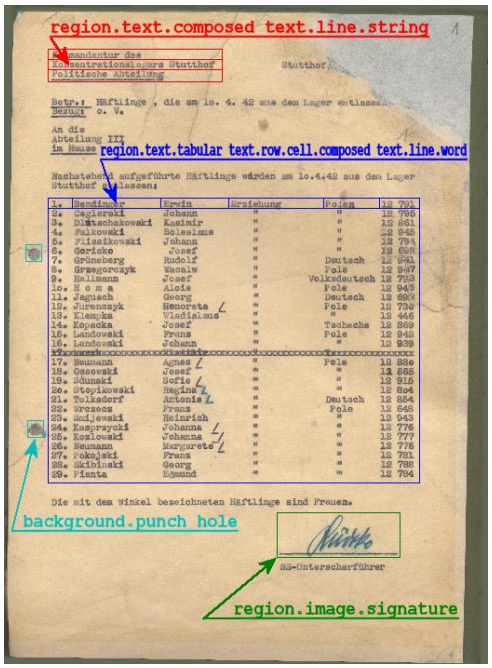

Fig. 4. Examples of different entities.

4. FROM TEMPLATE TO CONTENT

4.1 Re-typing vs. Engineering

An important distinction must be made at this point between manually entering the content (typing) and using DDW tools that interact with the template at various levels to recover the information from the scanned document.

Re-typing the content of a document, apart from being costly and time-consuming, simply makes available a stream of searchable text, possibly with some layout information for presentation purposes. More specifically, current re-typing practices across memorial places and museums involve electronic forms, designed (usually by a bureau) individually for each class of documents. Manual form filling is a time consuming process, as it relies entirely on human interpretation of a document image. For example, with typing in a single personal record taking about three minutes, and another two for its verification, the processing

of a single class of documents in the Stutthof Museum (with 32000 records) took about five years of work involving two archivists. In addition, it must be noted that document content quality control in this case concerns only textual content, as no information on the original document layout can be stored in the electronic form. Moreover, the evaluation of quality is only visual and subjective, as no formal quality metrics can be applied.

On the other hand, using the DDW tools to create and progressively fill the document template with content provides for the required higher level semantic, structural, functional and other important considerations with regard to using historical documents as outlined earlier. Higher level semantics refers to annotations and internal relationships between selected regions in the same page, as well as external relationships between regions in different pages of a multi-page document. Annotations and relationships (defined as links) can be introduced to the *content XML* document (see below) as separate layers with the multivalent VED browser by an expert historian.

The DDW tools are applied in sequence and at the end of each stage, the output of the relevant tool has added a further level of sophistication to the representation of the historical document. The performance of each tool is actively tuned to produce the best overall quality result (see Quality Tuning below).

The first tool (template editor – EDD) after document qualification is applied by the historian/archivist to define the document template for the selected class of documents. Expert domain-specific knowledge is encoded at this stage in the form of the document structure. The information recovery process then starts, operating on the scanned document, simultaneously analyzing the image content and verifying/updating the document template information. OCR is the last part of this process before the extracted information is presented to the user in a document-specific interactive editor to be corrected and, finally, accepted. These stages, as well as the overall quality assessment and parameter tuning are described next in more detail.

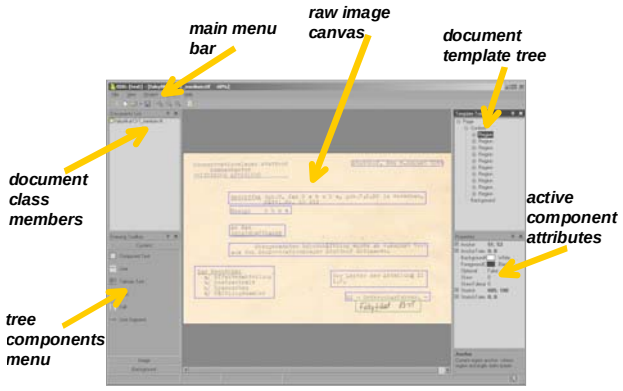

Fig. 5. A screenshot of the interactive template editor.

4.2 Structure and Semantics

The template creation process initialises a base (generic) XML document layout model. EDD, the interactive template editor (Figure 5) as well as two helper applications are provided in the DDW system to help the historian/archivist define the document structure and semantics for the selected document class. The template editor, allows the users to work directly on the document image canvas and define the layout components using easy drag and drop procedures.

4.3 Content Extraction

The content extraction phase aims at locating and understanding the information existing in the scanned document, in order to appropriately fill in the document model (XML file) for the document. The content extraction phase comprises a number of individual processes; mainly document image analysis and character recognition. All content extraction processes are tightly integrated with the document model, which is used as much to provide information about the document, as well as a placeholder to store the extracted information.

4.3.1 Document Image Processing and Analysis

The first step towards extracting information from the scanned document aims to improve the document image in a way that optimal results are achieved during the subsequent character recognition process. The necessity of this step is dictated by the intrinsic characteristics of historical documents. Due to ageing and the way historical documents have been preserved, certain artefacts are present in most of the cases, which significantly hinder the character recognition process. Specifically, historical documents have highly textured background, typewritten characters that have been transferred onto the paper with different strengths, diffused ink (especially in the cases of carbon copies) etc. An artefact resulting from earlier document restoration attempts is the presence of areas of reconstructed paper, where missing paper is “grown” back using liquid paper, introducing areas of different colour in the scanned document.

During scanning, the not-so-careful placement of the paper on the scanner is likely to introduce skew, as well as include some non-document regions (e.g., the scanner lid) in the image, if the paper does not cover the whole scanning area. Finally, certain artefacts such as stains, staples and punch holes as well as non-textual document entities such as handwritten marks, stamps or signatures, must also be segmented before the document image is subjected to OCR.

4.3.1.1 Segmentation of background entities

The background entities described above are segmented in this step and excluded from further processing. Due to the requirement (specified by the Authors) to scan each document against a dark background, a dark outer region surrounding the document exists in every image. This surrounding area is identified and marked as such.. A first attempt is also made to identify and correct skew.

The second type of background entity that is segmented at this point is that of areas of reconstructed paper. The segmentation of reconstructed-paper areas is performed in two steps. First, potential areas of reconstructed paper are identified in the image, based on their colour characteristics. Subsequently, the identified regions are filtered based on their location in the image. An example of the results of background entities segmentation is shown in Figure 6.

4.3.1.2 Character location

In order to improve the text regions to the effect that merged characters are separated, and faint ones are “lifted” from the background, the approach described here performs an individual

character location and enhancement process. This approach is novel in this type of application and is afforded by the regularity of the typewriter font.

In order to locate individual characters in the image, a top-down approach is followed [1]. First, the regions of interest are looked up in the XML document template. This minimizes the overall processing effort required, since character location only takes place within given areas instead of the whole image. For each text region of the template, a two-step process takes place to locate the characters: first, the identification of textlines in the region is performed, and then for each textline extracted, the characters within it are segmented.

The information extracted for text lines is used to update the information stored in the document template, so that it accurately matches the contents of the specific image.

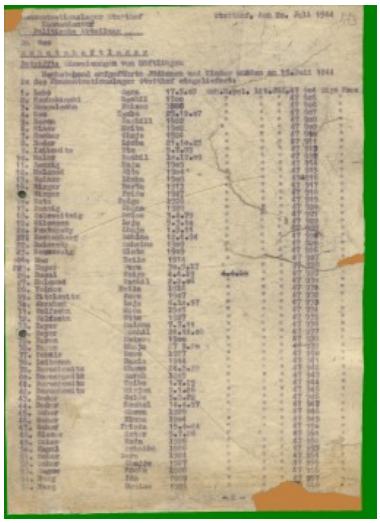

Fig. 6. The identified surrounding area (green) and the reconstructed paper areas (orange).

By locating individual characters within textlines, an important problem, which hinders the OCR stage, is readily addressed: merged characters (characters that are touching in the original image) can now be separated. An example of individual characters precisely located within the document image can be seen in Figure 7. It can be seen that, apart from the enhancement of the characters, the separators correctly split characters that were merged in the original image (e.g. the last three characters "GER" in the word "KONZENTRATIONSLAGER").

4.3.1.3 Image-based character enhancement

Having identified the position of all characters in the image, local enhancement takes place for each character. This processing, aims at improving the characters and producing a black and white image of the character, which will be used by OCR in the next stage. A local (individual character) approach can potentially produce to a great extent better results in the case of typewritten documents, since each character is formed individually and the strength of the transfer (force on the typewriter key) can vary from character to character.

A number of contrast enhancement and adaptive thresholding approaches have been implemented and tested (including variants of histogram equalisation techniques [4], Niblack's [5] method and Weszka and Rosenfeld's [6] approach).

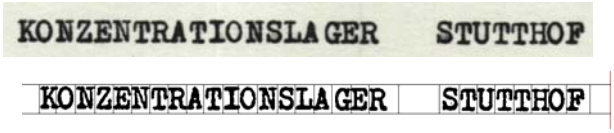

Fig. 7. Text in the original image and (below) the result of individual character location and enhancement.

Very encouraging results have been obtained so far, with merged characters correctly separated and faint characters (previously classified as background) recovered (see Figure 8). The ability to locate individual characters constitutes a very significant benefit for any enhancement process and this is one of the characteristic advantages of this project.

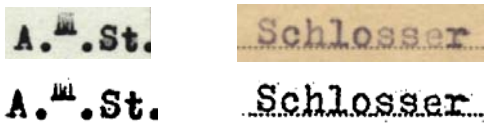

Fig. 8. Detail of text showing faint and strongly pressed characters properly recovered.

4.3.2 Character recognition

The OCR process (commercial product) is provided with the enhanced image file and the location of each logical entity (from the intermediate content XML structure). At the end of this step, the recognised characters are inserted in the content XML structure.

The OCR package used cannot be trained directly on the document class in hand. Instead individual dictionaries have been created with valid text for different semantic entities. For instance, a dictionary of first (proper) names and a dictionary of place-names is being used to improve the recognition rate for the corresponding semantic entities in personal index cards. It must be noted that this ability to apply different recognition parameters to different semantic entities, and the corresponding improvement of results, is only possible due to the semantics-rich document architecture devised.

4.4 Quality Tuning and Acceptance

Quality tuning is another important feature of the DDLC model and forms an integral part of (and in fact being enabled by) the quality management approach that underlines the whole system. Tuning is an important feature from the point of view of the performance of image processing algorithms used by DDW tools, as well as of the incorporation of human intelligence in document engineering processes. On top of that there are legal regulations, protecting copyrights as well as the content of specific documents, e.g., with personal information – which may prevent access of quality experts to certain classes of documents. Hiring such experts on-site to process each class of documents may be economically not viable, while sending documents out of the archive may be legally impossible. A solution developed in the project can circumvent such difficulties.

Consider again Figures 1 and 2 and denote by *Q(PD)* and *Q(ED)* respectively, the quality of the (input) paper document, and the quality of the (output) electronic document; it may happen that:

1. *Q(PD) > Q(ED)*, the final product quality has deteriorated during processing along DDLC phases;
2. *Q(PD) ≅ Q(ED)*, the final document quality has not significantly changed compared to the original;
3. *Q(PD) < Q(ED)*, the final document quality has been improved during processing.

The first two relations are not very interesting; the first one will occur when process parameters of DDLC phases are not set correctly (in which case a document shall not be accepted), while the second one will occur when the parameters are set sufficiently well (but not optimally). More interesting is the third relation, indicating actual increase of document quality during DDLC. It is possible only when expert user has been able to successfully contribute to the document engineering processes. One extreme of that is the manual creation of a document with GED, based on a previously defined template with EDD – if only the raw document image is legible enough for the historian. A typical situation observed during the project is the quality improvement combined with a significant reduction of time and effort in editing a document during acceptance phase, compared to manual reproduction of a document from scratch. Document quality assessment in any DDLC phase uses a specially developed Visual GQM (VGQM) method, supported by QED [2].

An overall scheme of DDLC tuning has been outlined in Figure 9.

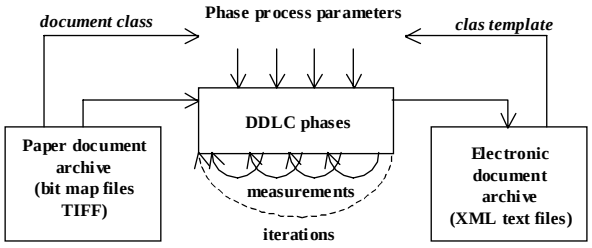

Fig. 9. Quality tuning of DDLC.

The VGQM method distinguishes between parameters and metrics. Parameters characterize processes of each phase, and their values may be used to control component DDW tools (see Figure 2). The values of metrics, specific to each phase, are measured to characterize the respective input and output data. The value of Q is calculated based on a quality tree and normalized to a five grade scale, from "very low" (VL), through "low" (L) and "medium" (M), up to "high" (H), and "very high" (VH) quality. Process tuning by a quality expert requires first defining quality tree for each respective phase, defining metrics and setting up weights. Next a representative set of three up to five documents for a semantic class of interest has to be selected and put through the cycle. In each phase input and output quality values are measured, and process parameters for the best observed relation between them is stored as the optimal phase setting. Once all optimal settings for each phase are established by the quality expert, the processing of the remaining documents of the class can be performed automatically in a batch by an archivist. Any document that cannot pass the quality threshold set up by an expert may now be rejected. Thorough selection of acceptance criteria for each respective class implies that either document processing progresses to the next phase, or is of such a poor quality that it must be processed manually (retyped).

In Figure 10, a snapshot of a QED screen shows quality trends when tuning parameters of the extraction phase for personal cards. The upper part indicates the progress achieved for preceding phases in the form of a bar diagram, while below a sample document with the regions defined in its template is displayed and process parameters for each region can be fine-tuned separately to achieve the optimal performance for the entire page.

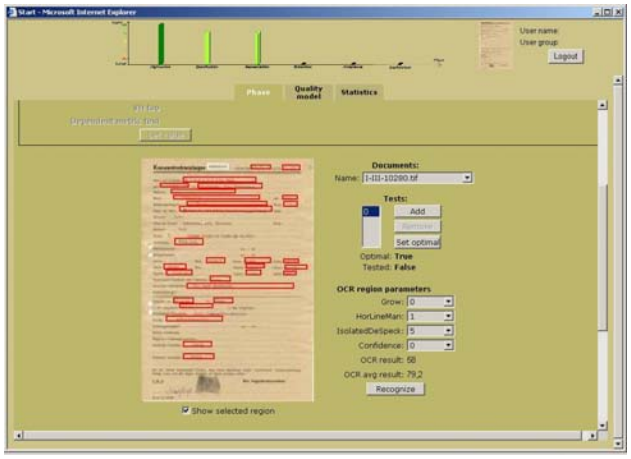

Fig. 10. Quality trends across DDLC phases.

4.5 Validation of the DDLC Model

Experimental results obtained so far indicate that by considering human effort overhead on document processing is reasonable and justified. Single "difficult" (denoting "low"—L or "very low"—VL quality) documents should be processed manually (retyped), while numerous "easier" documents processed automatically. Determining which documents are sufficiently "easy" and "numerous" in not straightforward to an expert historian (archivist) not familiar with GQM. It is worth mentioning that machine typed documents processed in MEMORIAL, commonly considered "difficult" to process with commercial OCR, have been able to give quality results comparable to printed documents. It has been made possible by the QED tool, implemented as a web-tool, to reduce external expert costs.

The effectiveness of the whole approach is assessed by evaluating acceptance-testing scenarios. The scenario related to the uniform model only is discussed here, for reasons of brevity. The model takes into account four metrics: the average OCR confidence level (as output by the package), the percentage of correctly recognised characters, the percentage of correctly recognised words and, finally, the document preparation time ratio (indicating time/cost savings as opposed to human transcription). Quality (effectiveness of the system) is expressed in the range of 0–1. Three different cases are compared in terms of quality value: the direct application of the off-the-shelf package to the document, the application of the OCR package following thresholding by Otsu's method [3], and finally, the comprehensive approach of the MEMORIAL project.

The uniform model (graph shown in Figure 11) applies equal weights to each of the metrics. It is evident that the whole approach constitutes an overall improvement to both the manual transcription and to the semi-automated application of off-the-shelf packages. Moreover, the richness of information

(semantically tagged) obtained by the approach described here is far superior to the output of generic OCR.

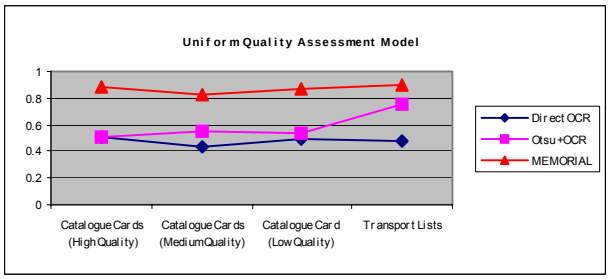

Fig. 11. Uniform quality assessment model graph.

5. ADVANTAGES OF THE SEMANTICS-DRIVEN APPROACH

The semantic-driven approach to describing and recovering the structure and content of historical documents has a number of advantages over more conventional document representations.

First, it provides for rich querying and more semantically accurate results. For instance, the document architecture enables one to perform a query such as "show all Camp Commanders' names". Second, as far as recovery of information from the document image is concerned, semantic information allows for individual tailoring of methods to improve recognition. For instance, if an entity is a date of birth, one can specify for the OCR that only numbers are present and perhaps indicate valid ranges. Furthermore, due to the nature of the historical documents, controlled access to different parts of the document and different application scenarios are required. The document architecture specifies accurately the types of entities and appropriate access can be given denied at a per-entity level.

Finally, the semantically rich architecture can be used advantageously even without the subsequent recognition processes. For documents that are very difficult to recognize, for instance, one could re-type the text (using the GED tool) and still benefit from the advantages outlined above. In addition, the functionality of the editing tools (enabling drag-and-drop operations between parts of architectures of different classes of documents) minimises user-interaction in the creation of similar class templates (effectively providing for reusability of architectural components).

6. CONCLUDING REMARKS

This paper has presented the lifecycle model and overall architecture for converting, representing and using historical documents. The document structure devised and the tools to recover the content from scanned images to fill the document template were described.

The document architecture and the content extraction methods in the form of DDW tools have been proven to be effective in trials with historians and considerably more preferable to re-typing the document content in a conventional document format (text with simple layout).

7. ACKNOWLEDGMENTS

The authors wish to acknowledge the support of the European Union for this work under grant IST-2001-33441.

8. REFERENCES

[1] Antonacopoulos, A. and Karatzas, D. A Complete Approach to the Conversion of Typewritten Historical Documents for Digital Archives. In *Proceedings of the6th International Association For Pattern Recognition Workshop on Document Analysis System* (DAS 2004) (Florence, Italy, September 8-10, 2004), Springer LNCS, 90–101.

[2] Krawczyk, H. and Wiszniewski, B. Visual GQM Approach to Quality-driven Development of Electronic Documents. In *Proceedings of the Second International Workshop on Web Document Analysis* (WDA2003) (Edinburgh, UK, August 3, 2003). PRImA, Liverpool, UK, 2003, 43–46.

[3] Otsu, N.: A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man and Cybernetics, Vol. SMC-9 (1979) 62–66.

[4] Sonka, M., Hlavac, V. and Boyle, R. *Image Processing, Analysis and Machine Vision*, 2nd edn. PWS Publishing, 1999.

[5] Niblack, W. *An Introduction to Digital Image Processing.* Prentice-Hall, London, 1986.

[6] Weszka, J.S. and Rosenfeld, A. Threshold Evaluation Techniques. *IEEE Trans. on Sys., Man and Cyber.*, Vol. SMC-8 (1978), 622-629.