Definition gleichartiger Dokumententypen zur Verbesserung der Erkennbarkeit und ihre XML-Beschreibung

Definition of similar document types for better recognition and the corresponding XML-description

Janusz Jarzemski, Henryk Krawczyk, Michal Melzer, Marcin Smolka, Bogdan Wiszniewski
Faculty of Electronics, Telecommunications and Informatics, Gdansk University of Technology
ul. Narutowicza 11/12, Gdansk, Poland
Tel.: +48 58 347-1018, -1089, Fax: +48 58 347-2222
E-mail: {avatar,hkrawk,mentos,zefir,bowisz}@eti.pg.gda.pl, Internet: http://docmaster.eti.pg.gda.pl

Zusammenfassung:
Im Vortrag wird ein schrittweises Vorgehen zur Überführung eines vorliegenden analogen (d.h. mit Schreibmaschine geschriebenen) Originalschriftstücks in ein interaktives elektronisches Dokument vorgestellt. Mit "elektronisch" ist dabei ein vollständig interaktives Dokument gemeint, das in XML beschrieben und für die Verarbeitung durch jeden Standard-Web-Browser geeignet ist. In den Zwischenschritten werden verschiedene digitale Darstellungen der Seiten verwendet. Das vorgestellte Modell eines "Lebenszyklus von digitalen Dokumenten" besteht aus Phasen, von denen jede wohldefinierte Prozesse und Resultate von kontrollierbarer und voraussagbarer Qualität einschließt.

Abstract:
The paper introduces a stepwise approach to the development of an interactive electronic document from its analog paper origin. By "analog" we mean a typed piece of paper, by "electronic" we mean fully interactive document, described in XML and suitable for processing by any standard Web browser. In between analog and electronic page extremes we have to deal with various digital (binary image) intermediary page representations. The proposed Digital Document Life-Cycle (DDLC) model consists of phases, each one involving well defined processes and products of controllable and predictable quality levels.

1. Introduction

In the paper we report on the ongoing project aimed at the development of interactive electronic documents and creation of a Web portal based on information extracted from machine-typed documents [5]. Although *Optical Character Recognition (OCR)* tools today can reach a ratio of correctly recognized characters of printed text well over 95% [3], machine typed text still constitutes a barrier; success rate in recognizing typed characters, especially in the presence of even minor noise, is well below 50%. This is because the OCR tools normally expect binary image inputs. The MEMORIAL project consortium has developed a framework enabling expansion of OCR capability by taking advantage of color information and document content semantics. The proposed *Digital Document Life Cycle (DDLC)* [4] introduces a series of inter-related and quality driven processes for systematic extraction and engineering of knowledge contained in paper documents.

2. Digital Document Workbench tools

Digital document technologies have been rapidly advancing in many areas, from more powerful scanning devices, more advanced image processing methods for skew elimination, image improvement and color based segmentation, through intelligent character recognition and text extraction algorithms, up to electronic document information management in distributed databases

and Web information systems. Unfortunately, their effective deployment in a real context provided by archives holding large volumes of historical documents of various origin and form and containing often very specific information is not straightforward.

First of all, historical documents may be preserved in a various physical condition. Automatic detection of their defects during background segmentation, as well as character recognition and text extraction are context sensitive, thus algorithmic solutions cannot produce good quality output. Even narrowing the class of analyzed documents to some specific medium, form and purpose is not sufficient to achieve the level of quality that may be achieved, as for Web documents for example. Research and development work of the MEMORIAL project focuses on machine typed documents. Although they may be considered a class, owing to the fixed character pitch and line skip, or the same font type and size used consistently throughout the entire document, direct analysis of their content with OCR tools currently available on the market gives poor results. While for documents, printed on white paper sheets with a stable texture, the ratio of correctly recognized words measured in the experiments carried out by the consortium with several leading OCR tools was about 95%-98%, experiments carried out with the same OCR tools on archival machine typed documents showed that ratio to be around 30%-35%. Based on the latter figure, the effort estimate for correcting extracted texts of such a poor quality could be expected to be equal (if not higher) to the effort estimate for just typing documents right into a database by the archivist.

The MEMORIAL project has adopted the following means to overcome the barrier created to the OCR process by machine typed documents:

1. In-depth analysis of archival resources in order to identify and define classes of documents with regard to their content semantics, layout and scenarios of use.
2. Assessment of scanning techniques and devices from the point of view of the scanned image quality and performance of the mass scanning process.
3. Contextual analysis of document content in a way enabling to build into the segmentation and extraction processes historian expertise on the analyzed document semantics.
4. Integration of document image processing and recognition tools in a coherent framework, enabling effective management of information concerning document content.
5. Document process quality improvement by introducing multi-phased data quality assessment methodology.

3. Document lifecycle development

In place of a traditional OCR process model the MEMORIAL project has introduced a *Digital Document Life Cycle (DDLC)* development model, which distinguishes manageable sequence of phases and assumes that an accepted output of one phase becomes input to the following phase. The DDLC model has been shown schematically in Fig. 1. DDLC phases include:

- *digitization;* a scanned raw image file is produced based on its paper document page origin;
- *qualification*; a set of similar page images is selected and the class is defined by a special template;
- *segmentation*; elements of the page background are filtered out from the page image;
- *extraction*; textual content of page regions according to their definition provided by the template is extracted;
- *acceptance*; correction and final assembly of regions' content is made to obtain the final page representation;
- *exploitation*; the page content is used for interaction in the Web environment.

By using the notion of classes, i.e., a set of documents of a similar layout, structure and semantics, it has been made possible to introduce the concept of document engineering, where information is converted from its pixel form (image) to textual (editable) form using a common and predefined document template throughout the entire cycle.

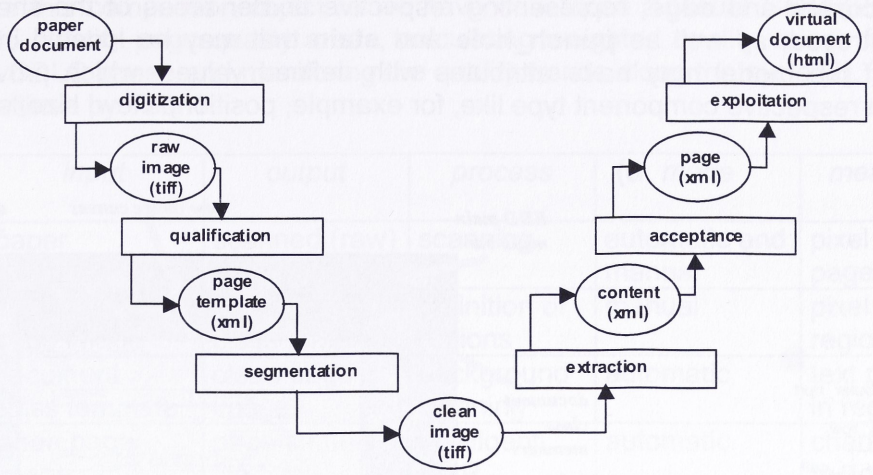

Fig. 1: Digital Document Life Cycle model

3.1. Template driven document engineering

A template can be defined by an archivist, for a class of similar documents, as an initially "empty" XML document. Such a document specifies regions of a precisely predefined type of content. Region specifications are interpreted by DDW tools to extract characters from the scanned document pages into respective regions of the template to produce *content* files - one for each respective page of the class. Classic contextual support for the OCR process is based on dictionaries, which may help to resolve problems with recognizing individual words and groups of characters. In the case of mass scanning of similar documents their common template can provide additional information on a specific layout of various components in each analyzed page, especially relationships between elements of tables, for which OCR dictionaries provide no support. In order to enable archivists to introduce a valuable semantical information on document regions two problems had to be resolved:

1. Definition of a uniform document layout and content model that is suitable for the universe of machine typed documents;
2. Development of a tool for generating templates that is intuitive and easy to use by a non-technical user (historian, archivist).

Based on the specific features of machine typed document a formal document layout definition has been introduced. It uses a hierarchy of *components* shown in Fig. 2a. Each component (a **<tag>** in the corresponding XML file) can appear in a document tree *at most once* (marked as '?'), *exactly once* (marked as '1'), *at least once* (marked as '+'), *arbitrary many times* including 0 (marked as '*'). Each machine typed **page** includes one **content** component, consisting of pixels of **text**, and one **background** component, consisting of pixels of dirt, stains, punch holes, etc., considered a noise. Page content consists of at least one **region**, or arbitrary many printed **line_segment** components, possibly dividing the page into disjoint parts. Component region may include exactly one component **text** or **image**. Component **text** can either contain a **tabular_text** component or a simpler, **composed_text** component. The main feature of tabular contents is that they contain composed texts in the nested **row** and **cell** components, as shown in Fig. 2. The only content of a composed_text can be **line**, **word** and **character** components. Certain classes of documents may also include certain individual hand written characters **hw_marks**, appearing in a line as single words. Moreover, characters may sometimes appear in groups representing constant **predefined_string** components. Besides textual regions a document page may include non-decomposable **image** regions. The latter consist of pixels for further (non-textual) analysis, such as **photograph**, **signature**, **stamp**, hand written annotation **hw_note**, or drawn or printed **graphics**. Another basic component of a document page is **background**. It contains all "physical" objects in the (paper) sheet on which the analyzed document has been typed. The model described in Fig. 2

distinguishes: **corner** and **edge**, representing respective border areas of the sheet that may be missing or destroyed, as well as **punch_hole** and **stain** that may be located inside the sheet. Components of the model may have attributes with defined values, which provide information specific to each respective component type like, for example, position, skew, size, skip, etc.

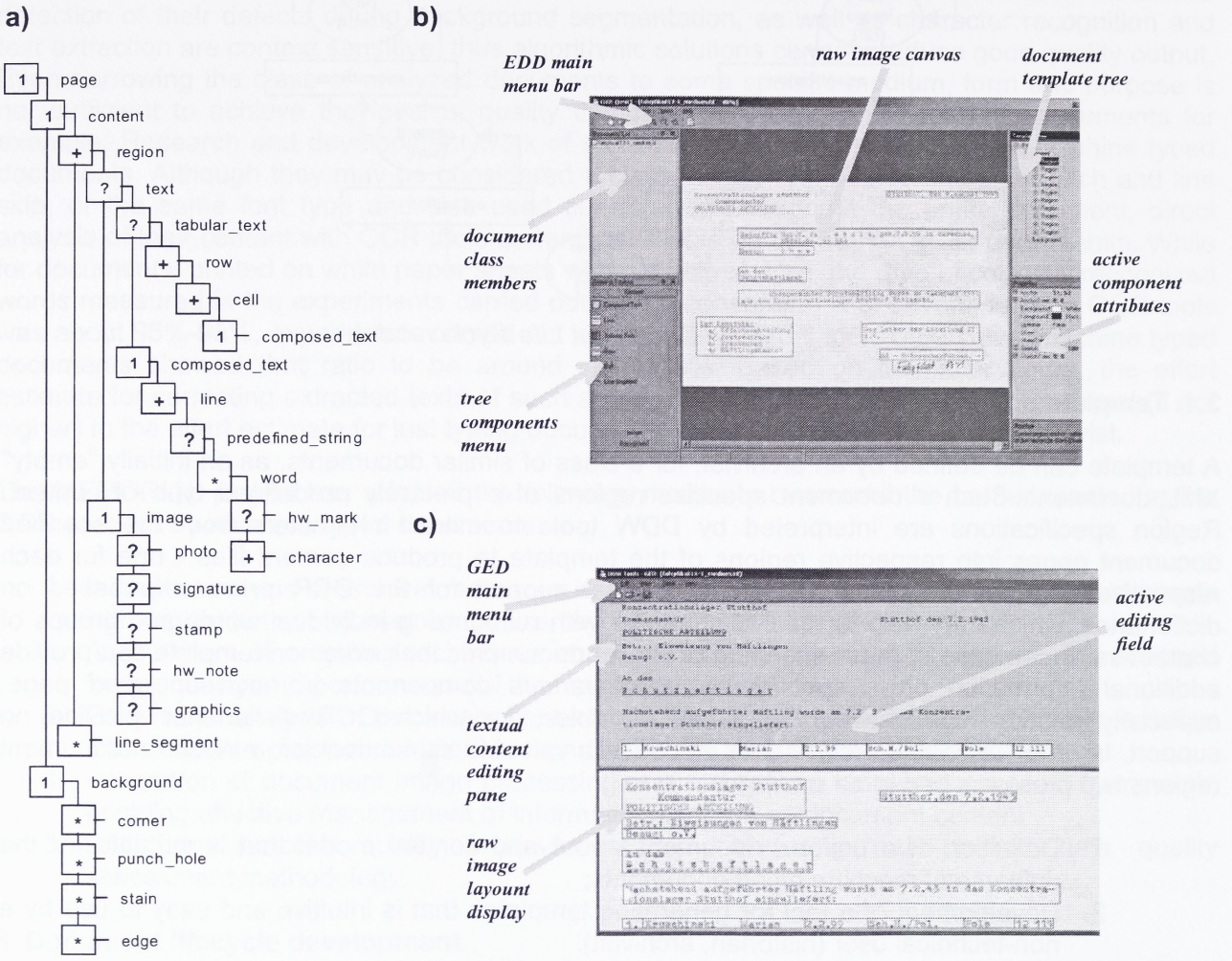

Fig. 2: Document layout definition: (a) tree, (b) template editor EDD, (c) content editor GED

3.2 Document process quality management

A key concept of the process wide quality management is that each document engineering life-cycle phase is monitored in terms of the input and output data quality, and the parameters characterizing the respective phase. Consider for example the digitization phase, which besides physical browsing of the paper archive by an expert historian involves the main scanning process. Evaluating quality of input data, i.e, selected pages of documents, implies visual examination of the physical state of each single page by an expert, in order to identify areas of various quality, which may influence further processing of that page scan. Output of the digitization phase in the raw image file, which quality can also be measured. The simplest means for that is again a visual inspection by the archivist (but this time of a raw image), and comparison of the respective areas in the image and its paper origin. Parameters of the scanning process, like scanner device make and model, serial number, name and version of the scanning software, resolution and image size, exposition time and brightness [1] shall be recorded in the database. Acceptance of the scanning process parameters is possible then if the observed output data (raw image) quality is not less than the observed input data (paper page view) quality.

In order to make this comparison objective we have developed a *Visual GQM (VGQM)* method, allowing users to identify page quality areas, calculating weighted quality values for a document page after each DDLC phase and monitoring the quality trends along the cycle [2]. This process is schematically outlined by Tab. 1.

phase	*input*	*output*	*process*	*mode*	*measurement base*
Digitization	paper document page	scanned (raw) page image	scanning	automatic and manual	pixel areas in page
Qualification	scanned (raw) page image	document class template	definition of regions	manual	pixel areas in regions
Segmentation	document class template	clean page image	background cleaning	automatic	text pixel areas in regions
Extraction	clean page image	page content file	intelligent OCR	automatic	characters in regions
Acceptance	page content file	electronic document page	edition of region content	automatic and manual	words in regions
Exploitation	electronic document page	virtual document (Web page)	browsing, linking, annotating	automatic and manual	user interaction with content

Tab. 1: Multi-phased document quality control

Refer to Fig. 1 and note how important pieces of information can be added step-by-step into the document life-cycle. In the *digitization* phase all pixels in the document raw image are taken into account, combined with the expert user information on physical condition of specific areas of the page. In the next, *qualification* phase, similar documents are distinguished by the user as a class, who can next define their generic layout, structure and content with a template. The information provided by the template in next used by the background cleaning process of the *segmentation* phase to filter out pixels belonging to background artifacts and constituting a noise. Assessment of this phase quality uses both: regions to be analyzed by OCR and their relationship to previously identified areas of page quality. A resultant quality value for each processed document page image should then take into account quality values for each region defined in the template, weighted accordingly to their overlapping with the previously indicated page quality areas. The same kind of evaluation is made next in *the extraction* phase, where characters in the respective regions are recognized. This time quality assessment takes into account quality of the extracted characters rather then pixels of the respective regions. Finally, during the *acceptance* phase, words of text in each respective region form the basis for an overall extracted document page content quality. This can be done automatically, by measuring extracted words quality with regard to dictionaries relevant to each type of region (e.g. surnames or birthdates), or manually, by counting the number of corrections made by the user with a human editor support tool. Upon acceptance, document content becomes a resource that may be used in Web information systems. Quality assessment there (*exploitation* phase) should be based now on measurements of end-user interactions with the document content, e.g. number of annotations to a given electronic document may indicate problems in interpreting its content, number of hyperlinks to other documents may be used to measure importance of its content, and so on.

The document process can be fine-tuned by the quality-control expert, so that mass processing of large volumes of documents belonging to the same class (associated with the same template) can be done automatically, without loosing a grip on the total quality management in DDLC. This requires selecting of a few sample (representative) documents of the class to be processed and using them to determine experimentally the optimal values of parameters of all respective processes in each phase with the VGQM method. Once determined, the same parameters are set for the respective processes for processing all documents of the class. If some documents do not meet the quality standards set up for the sample documents in some phase they are disqualified

and left in the database for further inspection by the expert. Their processing may be continued manually, e.g., seriously damaged or rare documents, or automatically, after being qualified to another class with a different template. Fine-tuning of DDLC processes with samples and monitoring of document flow in mass processing is supported by the document quality evaluation (QED) tool.

4. Document engineering tools

Tools implementing DDLC phase specific processes are shown in Tab. 1. The complete set of tools constituting the *Digital Document Workbench (DDW)* and developed in the MEMORIAL project includes also the tool for *Quality Evaluation of electronic Documents (QED)* mentioned before and the *Working Repostitory (WR)*, a specialized database integrating QED and phase-specific tools listed in Tab. 2.

Phase	*Tool*	*Description*
Digitization	IDT	**I**n**D**exing **T**ool for automatic naming and management of raw image files generated by scanning devices
	RLT	**R**epository **L**oading **T**ool for creating and storing links to raw image TIFF files in a working repository of DDW, along with automatically generated raw image JPEG and thumbnail files.
Segmentation	EDD	**E**lectronic **D**ocument e**D**itor for creating and editing electronic document template files
Extraction	IPT	**I**mage **P**rocessing **T**ool for preparing raw document page images for page content retrieval with OCR
	OCR	**O**ptical **C**haracter **R**ecognition tool for extracting text (strings of characters) from document page images
Acceptance	GED	**G**enerator of **E**lectronic **D**ocumenst for editing content files interactively
Exploitation	VED	**V**iewer of **E**lectronic **D**ocuments - multivalent browser for browsing, annotating and linking content of selected documents

Tab. 2: Digital Document Workbench tools

Due to the limited space of this paper we describe briefly just two DDW tools, namely EDD for defining document class templates, and GED for editing document content files.

4.1. Document template editor EDD

As indicated before a tree defining a document layout and structure is represented in a machine readable form with XML files. While a document *content* file is generated and updated automatically by various DDW tools, a *template* file must be generated from scratch by the expert historian. Of course it can be done in a very primitive way practically with any text editor, but in such a case a great deal of knowledge and expertise on XML file syntax and structure would be required. In order to ease the creation process an interactive template editor tool *EDD* (*Electronic Document eDitor*) has been developed. It has a simple and intuitive interface that allows archivists and other non-technical users to operate directly on the document image canvas using drag and drop facility with all document tree components represented as drawable objects, as shown in Fig. 2b. *Document class members* that are supposed to fit the same class of documents (a project) can be selected by browsing the list of available raw image files. A selected document image can be displayed in the central part of EDD window as a canvas for drawing regions selected from the *tree components menu*. Upon selection and drawing, a component is added to the document tree displayed in the *document template tree* area. The tree may be further expanded either by selecting and clicking components directly in the *raw image canvas* viewing area, or in the *tree viewing area*. Attributes of the currently selected (active) component are displayed for initialization, inspection and/or modification in the *active component attributes* viewing area. At any time a

document template tree (a project) can be saved for further development or modification, using the *main menu bar* entries. The saved project includes binary representation of the project status and automatically generated template XML file, which can be viewed with any XML enabled browser.

4.2. Document content editor GED

The extracted document content described with the XML content file can be browsed with any XML enabled browser and edited with any text editor tool, like its template counterpart. However, such an exercise would not be cost effective, and as in the case of template files would require XML expertise from archivists using DDW tools. Therefore a document content editor *GED* (*Generator of Electronic Documents*) for interactive edition of content files has been developed.

Since extraction of characters by OCR into various areas of the XML content file may be not satisfactory and may require edition by the human user, the GED editor tool has been introduced to the process. GED window is shown in Fig. 2c. Upon opening a raw image file of interest, by using the respective buttons of the *main menu bar*, the *raw image layout* is displayed with the graphical representation of regions defined in the corresponding template file. By clicking on the region of interest in the lower display area users may select (activate) any region defined in the template. Each activated region component becomes an *editing field*, where the existing content may be modified, altered, and/or typed from scratch. Saving of the changed XML content file overwrites permanently the previous content file in the DDW database. In the current version GED supports only interactive edition by the human user, but by incorporating a spell-checking facility in the near future it will also enable automatic edition.

5. Summary

Quality improvement that can be really obtained in document engineering based on the DDLC model relies on the essential "external"' knowledge provided by a human expert. The template based approach to document image improvement, and the subsequent content segmentation, recognition and extraction, makes it possible. It can effectively narrow down the search space for the OCR and drive selection of dictionaries in post-OCR processing. Moreover, parameters of all related processes can be fine-tuned by a human expert working with a representative sample of a larger batch of documents. Upon setting up the relevant process parameters the remaining portion of documents can be processed and assessed automatically. The associated set of DDW tools, supporting one another and integrated on top of a common specialized database (working repository), the XML technology, and a total quality management paradigm is flexible and easy to use by non-technical users.

References

[1] NIST Draft standard Z39.7-2002: Metrics and statistics for libraries and information providers - data dictionary, Version 2002a, < http://www.archivists.org/catalog>

[2] Krawczyk, H., Wiszniewski, B.: Visual GQM approach to quality-driven development of electronic documents. Proc. 2nd Worskhop on Web Document Analysis, August 2003, Edinburgh, UK, pp.43-46

[3] DoKuStar: Océ Document Technologies, http://www.oce.de

[4] Krawczyk, H., Wiszniewski, B.: Digital Document Life Cycle Development, Proc. Int. Symposium on Information and Communication Technologies ISICT 03, September 2003, Dublin, Ireland, pp.262-267

[5] MEMORIAL-IST-2001-33441: A Digital Document Workbench for Preservation of Personal Records in Virtual Memorials (2001-2004), http://docmaster.eti.pg.gda.pl