Visual GQM approach to quality-driven development of electronic documents

Henryk Krawczyk and Bogdan Wiszniewski
Technical University of Gdańsk
Faculty of Electronics, Telecommunications and Informatics
ul. Narutowicza 11/12, 80-952 Gdańsk
{hkrawk,bowisz}@eti.pg.gda.pl

Abstract

Project MEMORIAL [3] is aimed at developing a new technology for creating Web based information systems using interactive electronic documents extracted from their paper originals with advanced optical character recognition tools. A multi-phased digital document life-cycle development model is proposed to assure high quality of the process and its final product.

1. Introduction

Image analysis and pattern recognition techniques are capable today of converting practically any kind of well structured and legibly printed documents into highly interactive Web documents. For example, the META-E project [4] has been concentrating on printed books and journals, which consist of such elements as chapters, subchapters, page numbers, running heads, graphs, captions, etc. The MEMORIAL project [3] reported in this paper have been focusing on the analysis of less structured machine typed documents. Such documents may constitute just typed sheets of blank paper, or be more complex forms mixing printed and typed text and graphical borders with hand-written annotations, signatures, rubber stamps and photographs. Moreover, typed original and carbon copy pages may include characters which are of different color, overstricken, shifted, only partially visible, etc. Due to a physical condition of a document some portions of the text may also be blurred with stains, punch holes, torn out edges, corners and other noise-like effects.

In order to tackle these problems in a systematic way the MEMORIAL project has introduced 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 any standard Web browser. In between we have to deal with various forms of "digital" (binary image) document page representation. 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.

2. Digital Document Life Cycle development

Phases of the DDLC model, outlined in Figure 1, include: *qualification* of a paper original for digitization, *digitization* of a document page to get a page image, *segmentation* of a document page image into regions of content, *extraction* of each region content into an XML file, *acceptance*, and finally *exploitation* of an electronic document page.

Qualification; The quality of a selected paper document depends on its context of use throughout the entire cycle. There are four basic aspects of data quality to be considered in this regard: *complexity*, *originality*, *background*, and *content* quality. *Complexity* reflects its semantical quality by characterizing retrieveability of the information, its completeness, conformance to other document standards, stability and fidelity. *Originality* of a document refers to its principal point of origin, its relationship to other documents in the archive, traceability, etc. Finally, *background* and *content* quality refer to its physical suitability for scanning and OCR processing, as documents may exhibit different degrees of fatigue.

A set of selected documents may be satisfactory for some of these aspects, and at the same time inadequate for others, making the digitization process hard to access, so it is necessary to find a way for balancing conflicting requirements for input data quality by deriving objective evaluation criteria. These criteria involve specially defined quality metrics to measure the indicated document aspects.

A methodology for measuring quality of documents across DDLC phases is introduced in Section 3 and 4; before that let us first complete quality analysis of the remaining phases of the cycle presented in Figure 1.

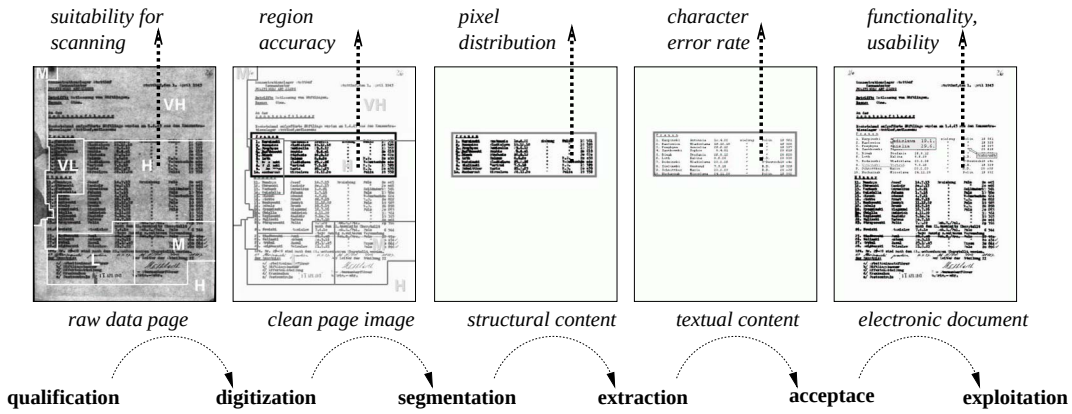

Figure 1. Quality driven Digital Document Life Cycle development

Digitization; A first step to get better recognition results is the adaptation of the scanning technology to a group of documents similar to the process. Optical filtering with glass filters and different illumination have been tried in the project in order to improve the recognition rate of poor print quality documents. Wavelength discrimination was used with relative wide band pass.

Other issues of image quality involves capabilities of existing OCR tools. Most of ready to use off-the-shelf OCRs require binary (black and white) input images. However, it seems impossible to get good results of recognizing blurred characters when using only binary images [3]. The specialized *Digital Document Workbench (DDW)* toolset being developed in the MEMORIAL project is aimed at handling color images before transforming them into binary image OCR inputs.

Segmentation; Known algorithms for text image separation from graphics are rather restrictive in the type of accepted documents and cannot be directly applied to machine typed texts. One problem for example is that graphics in the analyzed page image is the portion of (handwritten) text illegible for optical recognition. On the other hand, such information shall be saved for the future, when eventually more powerful algorithms for recognizing handwritten text may be discovered. Therefore page content segmentation must involve contextual analysis of its layout and structure to isolate illegible portions of a document page from regions with a readable content. Moreover, the latter must provide sufficient information for each region to ease as much a possible the textual content extraction.

Extraction; Respective regions defined in a template document are gradually filled with information extracted by the OCR tool under the control of a document analysis tool of the DDW toolset. In order for this scheme to work a document page image must be cleaned off background noise, in a process called *background cleaning*, and the characters must be rebuilt in a process called *character improvement*.

Acceptance; Carefully selected parameters of processes, and metrics measured for each respective DDLC phase product enable objective quality evaluation of a digital document built-up from individual region contents. The Goal-Question-Metric method [1] used in DDLC has been implemented in a form of a quality evaluation tool QED described in Section 4. An important feature of QED is that archivists monitoring individual phases of DDLC do not have to be experts in scanning, background cleaning, character improvement, etc. Actual ”tuning” of process parameters for each respective phase can be performed by an expert. When the tuning is complete, e.g. upon evaluating a set of benchmark documents, the respective parameters can be stored in a database; further evaluation may be performed then by an archivist using QED that can mimic the expert.

Exploitation; Extracted page content constitutes a sort of an enlivened electronic document page image. Pieces of scanned graphics and textual objects have the appearance of a scanned image, but could also be manipulated as if they were a word processing application. For example, even though users are presented with a document image, they can select text from the document, search for and highlight target terms and perform annotations. Annotations may provide such functionality as for example active manipulation of documents to which they are attached, re-attachment to any other document the one can be linked to, reaction to changes in the underlying document, etc.

With such a functionality, a machine typed paper document may be converted into an interactive Web page, or become a component of some larger dynamically created virtual document.

3. Visual GQM

In order to evaluate quality of a digital (yet non-interactive) as well as an electronic document (already accepted for exploitation) a sequence of standard steps is re-

Table 1. GQM metrics for paper documents

	Metrics	Values				
		VL	L	M	H	VH
Category 1	*Document originality*	unknown source	suggestions about the source	opinion about the source	incomplete knowledge	detailed knowledge
	Document simplicity	regions differ in many pages	regions the same in many pages	regions differ in one page	one region in one page	one region with few characters
Category 2	*Background quality*	diversified texture and defects	texture with many defects	texture with minor defects	no texture with minor defects	clean without defects
	Renovation status	laminated and glossy	japaneese paper	some fragments missing	some fragments restored	well preserved
Category 3	*Page content simplicity*	regions hard to determine	difficulty in region location	redundand regions defined	empty regions defined	optimal region distribution
	Layout visibility	regions must overlap	regions may overlap due to high tolerance	alternative distribution of regions	all regions distiguishable	unique separation of regions
	Region font readability	unreadable characters	displaced or over−stricken	mixed character faces	single characters unreadable	all characters readable
	Region content clarity	content illegible	content mostly illegible	problems with understanding the content	small fragments illegible	legible content
	User defined exceptions	not defined	postulated	intuitive description	informal definition	formal definition

quired, namely identification of (business) *goals*, definition of quality *metrics*, selection of automatic *measurement procedures*, collection of *quality data*, and *decision making* based on metrics values.

The main goal is to improve quality of documents created during the life cycle, in a way similar to the software life-cycle.It means that for each phase of the cycle, quality of it's output document should equal at least the quality of it's input document.

For example, in the digitization phase (see Figure 1) the input document is a paper page ready for scanning. Before that, however, we may want to have a closer look at the original page paper sheet. Based on the visual examination of the sheet we may distinguish and mark on its scanned image adjacent areas of various quality, which in our view may influence further processing of this page scan. Let the allowable range of categories be VL (very low), L (low), M (medium), H (high) and VH (very high), represented respectively by numbers 1 through 5.

According to the GQM standard a set of questions and possible answers should be determined first. For electronic documents there are three respective categories of questions, concerning respectively document complexity and originality, background quality, and content quality; see Table 1 which combines these categories of questions with possible answers and values of metrics. Based on that a questionnaire pointing out the most important quality parameters of the scanned page may be prepared. With such a questionnaire a meaningful and systematic analysis of each marked area of the selected paper page can be performed, and the overall page quality assessment derived as an average of all relevant areas. Next, when preparing for the segmentation phase a document template layout can replace the questionnaire. Now, instead of answering questions for each identified region a resultant estimate can be calculated *automatically* as an average value of all respective portions of the page areas (previously marked during the qualification phase) that intersect with the selected region. Owing to easy to use and intuitive manipulation of region location and size, a quality of the template itself can be tuned to the best possible level. We call this refinement a *Visual GQM (VGQM)*. Further on, during subsequent phases, the quality of such a "tuned" region can be traced, by using metrics relevant to the actual phase. For example, pixel distribution metrics may be used to assess the content of a region prior to character extraction, and character error rate prior its acceptance. Finally upon acceptance, the same region of the original page, now converted into an electronic document may be evaluated from the point of view of its functionality and usability.

4. Document process quality improvement

From the quality point of view, processes describing each phase of a document life cycle can be characterized with a relation $\{< Q^i_{in}, \bar{P}^i, Q^i_{out} > | i = 1, \ldots, n\}$, where n denotes the number of experiments, $\bar{P}^i$ denotes a vector of process quality parameters set independently for each i-th experiment, and Q^i_{in} and Q^i_{out} denote measured values of quality metrics for respectively, input and output documents. During these experiments vector $\bar{P}^i$ of process quality parameters is found, for which the best quality improvement can be achieved. More formally, for each experiment $i = 1, \ldots, n$ we have to calculate $QI_i = \frac{Q^i_{in}}{Q^i_{out}}$, where Q^i_{in} and Q^i_{out} are called *quality factors*. Each quality factor is defined as $QF = \sum_{k=1}^{5} k \cdot w_k$, where w_k represents a regression coefficient (weight) of the resultant value of document quality, calculated as a ratio of all sizes of regions of the same quality level to the total size of all regions.

Table 2. Selected DDLC processes

Phase	*Process*	*Sample parameters*
digitization	scanning	speed, resolution, illumination, contrast [5]
segmentation	layout analysis	number of regions, relative positioning, tolerances
extraction	region extraction	specified region content domains
acceptance	post OCR	% of accepted regions, % of accepted characters

Process quality parameters for each respective phase of the document life cycle shown in Figure 1 are given in Table 2. Their values and corresponding measurement proce-

dures can be described with XML files, which specify concrete type names of parameters, their range and default values. Also the name and the respective number of parameters of each measurement procedure can be specified. Based on the defined quality relation we can choose parameters that lead to quality improvement of the document being created. Based on the respective quality factors, for each phase of DDLC a relevant decision on proceeding to the next phase can be worked out with the QED tool.

Quality improvement that can be really obtained there requires adding to the DDLC development some essential "external" knowledge provided by a human expert; if only image processing and automatic character recognition techniques were used, the best what one could get for each phase would be output quality Q_{out} not less than input quality Q_{in}, i.e. $Q_{out} \leq Q_{in}$. If we want to achieve more, i.e., $Q_{out} > Q_{in}$ a provision shall be made for adding knowledge on document semantics, in a form and content relevant to each phase. For example, *qualification* is aimed at selecting similar paper documents, constituting a class of semantically similar documents. Based on that a template structure of a document class can be defined with a specially developed template editor tool. Regions of a document can be specified down to each individual line (composed text) or cell (tabular text) and further on to words belonging to some semantical domains defined by the expert. This can effectively narrow down the search space for the OCR and drive selection of dictionaries in post-OCR processing. Moreover, parameters P^i for each phase 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.

5. QED tool architecture

Figure 2 outlines the general architecture of the *Quality Evaluation of Document (QED)* tool, which supports the VGQM method and DPQI procedures; it consists of four basic components: *user interface* for defining quality model and interpretation of commands, library of *document quality evaluation* procedures, *document process quality improvement* procedures enabling comparative quality analysis of each DDLC phase, and *document repository* for storing analyzed documents.

Upon conclusion of a given phase of a document lifecycle, its current output can be evaluated with VGQM, and next with DPQI procedures of a subsequent phase, a new output is produced. Quality of the next phase output is compared to the quality of the previous one, and if the result is not satisfactory (no improvement observed) phase tuning is performed, i.e., either DPQI procedures are replaced or their parameters are modified [2]. The QED tool supports

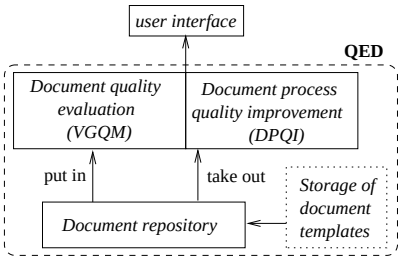

Figure 2. Conceptual architecture of QED

each respective DDLC phase and makes this comparison objective and effective. With the repository for storing intermediate document representations it is always possible to backtrack and re-tune preceding phases.

6. Conclusions

We have defined a quality driven document life cycle development, originating from the paper document qualification and ending up with exploitation of a corresponding interactive electronic (Web) document. With a special tool QED developed to support quality assessment of that cycle it is possible to interactively tune the entire process to obtain high quality level. Once tuned properly, the development process can be repeated automatically for a large volume of documents. Experiments carried out so far with machine-typed documents indicate that without a predefined document template the OCR error rate is about 50% and 70%. With the layout and region content information introduced by a human user this error rate is expected to be significantly reduced.

References

[1] IEEE. IEEE Standard for a Software Quality Metrics Methodology. *IEEE Computer Society*, December 1998.

[2] H. Krawczyk, M. Sikorski, S. Szejko, and B. Wiszniewski. A tool for quality evaluation of parallel and distributed software applications. In *Proc. Parallel and Applied Mathematics, PPAM'99*, pages 413–426, Kazimierz Dolny, Poland, September 14-17, 1999.

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

[4] MetaE. The Metadata Engine Project, IST-1999-20021. `http://meta-e.uibk.ac.at`.

[5] NIST. Draft standard Z39.7-2002. *Metrics & statistics for libraries and information providers – data dictionary*, Version 2002a.