

**INFORMATION SOCIETY TECHNOLOGIES
(IST)
PROGRAMME**



Project acronym: **MEMORIAL**

Project full title:
A Digital Document Workbench for Preservation
of Personal Records in Virtual Memorials

Contract no.: IST-2001-33441

***Digital Document Workbench (DDW) v1.0
Demonstration and Test Summary Report
Deliverable D9***

Date of preparation of report : 2004-01-31
Security : public
Edited by : Bogdan Wiszniewski (TU Gdansk)
Contributed by : Marcin Smółka (TU Gdansk)
Michał Melzer (TU Gdansk)
Krzysztof Drozdowski (Museum Stutthof)
Henryk Krawczyk (TU Gdansk)
Apostolos Antonacopoulos (UNIVLIV)
Wolfgang Schade (GFaI)
Cornelia Rataj (GFaI)

EXECUTIVE SUMMARY

This document reflects the state of the art of the *Digital Document Workbench, DDW v. 1.1.3*, toolset, released on Jan. 8, 2004 and demonstrated at the 2nd Annual Memorial Project Review Meeting in Luxembourg, on Jan. 12-13, 2004. Owing to the ongoing integration efforts, successive versions of DDW have been made available for testing and evaluation in the real context of work provided by two partner archives, the Stutthof Museum in Poland and the Moreshet Holocaust Study and Research Center in Israel since September 2003. All through that time DDW has been expanded in terms of its *accuracy*, implying lower error rate of the OCR output, as well as *efficiency*, allowing to process documents faster, and to reduce the amount of work spent on correcting the final document content before accepting it to the digital archive (database).

The experiments described in this document have been carefully planned, executed and their results evaluated to provide a common base for objective comparison with future versions of DDW and measuring the overall project progress. This document has been written according to the general standard IEEE guideline for test documentation, and its respective sections address all relevant issues concerning the tested product, its environment and test data used.

The current status of each DDW component has been detailed in *Section 1*, with special emphasis on features of that had to be tested during the reported experiments. Steps of the assumed test procedure have been described in *Section 2*, and an argument has been provided that the procedure can measure objectively advances of DDW beyond current practices of digitizing historical documents in memorial places. In *Section 3* a representative set of test cases has been described and explained from the point of view of objectives of the testing procedure. Organizational and legal aspects of test items needed for the test, i.e., executables of DDW components, other tools, and document files are described in *Section 4*. Incidents, failures and bugs noticed during the experiments are described in *Section 5*, including their impact on future versions of the DDW toolset and solutions to the problems they may have caused. Comprehensive assessment and evaluation of test results is given in *Section 6*. Finally, *Section 7* wraps-up the last 12 months of DDW toolset development and integration activities in Workpackage 5 with general recommendations for the future, including user site training and commercialization.

Gdansk, January 2004

1. Outline	4
1.1 Project status.....	4
1.2 Tested features.....	5
1.3 Features not tested	7
1.4 Approach.....	8
2. Test procedure.....	8
2.1 Purpose.....	9
2.2 Scenarios	9
2.2.1 Retyping.....	9
2.2.2 Raw image OCR	10
2.2.3 Context-free image improvement.....	11
2.2.4 Semantics-driven image improvement	11
2.3 Measurements.....	12
3. Test cases	13
3.1 Features	14
3.1.1 Transport lists.....	15
3.1.2 Personal cards	15
3.1.3 Index cards.....	15
3.1.4 Fakes.....	16
3.2 Required techniques.....	16
3.2.1 Manual OCR tuning	16
3.2.2 Background cleaning with Paint Shop Pro	17
3.2.3 Available functionality of DDW	17
3.3 Environmental needs.....	17
3.4 Intercase dependencies.....	18
4. Test items	19
4.1 Location.....	19
4.2 Status.....	19
4.3 Approvals	19
5. Test incident report.....	20
5.1 Description	20
5.2 Impact.....	21
5.3 Solutions.....	21
6. Test summary	21
6.1 Variances.....	21
6.2 Comprehensiveness assessment	22
6.3 Summary of results	22
6.4 Evaluation.....	23
6.5 Summary of activities	25
7. Recommendations	25

1. Outline

The *Digital Document Life-Cycle (DDLC)* development model, supported by DDW involves six phases, specified in detail in Deliverable D6 [5]:

1. *Digitization*; conversion of a paper document page into its digital (facsimile) image with a digital device like a scanner or camera;
2. *Qualification*; selection and assessment of scanned pages of paper documents constituting a semantically coherent class of archival items;
3. *Segmentation*; identification of major components (regions) of a document informational content in a document page image;
4. *Extraction*; automatic conversion of an image (portion of a document page image) representing a text into a string of ASCII characters;
5. *Acceptance*; spell checking, correction and edition of a text extracted from a document image;
6. *Exploitation*; analysis, interpretation and synthesis of information contained in the extracted document text.

DDLC phases can be performed in two modes:

1. *Manual (tuning) mode*, when a single document is processed repeatedly during each phase with various manual settings of control parameters of each relevant phase process until the best quality of each phase output can be achieved.
2. *Batch mode*, when documents are processed one by one along the entire cycle with process parameters set to fixed (default or tuned) values.

Distinguishing these two modes is necessary, because there is no any fixed set of values of process parameters that could suit all classes of documents equally well. Therefore a set of representative (sample) documents representing a class (project) is used first to tune the cycle and to find values of parameters that give the best results, and next all documents represented by the sample are processed with the same parameters.

1.1 Project status

DDW components and their features, described in detail in Deliverable D6 [5] are briefly summarized in Table 1.

Table 1: Summary of DDW components and their functionality

<i>Tool</i>	<i>Functionality</i>
<i>IDT</i>	<i>InDexing Tool</i> for automatic naming and management of raw image files generated by scanning devices
<i>RLT</i>	<i>Repository Loading Tool</i> for creating and storing links to raw image TIFF files in a working repository, along with automatically generated raw image JPEG and thumbnail files.
<i>EDD</i>	<i>Electronic Document eDitor</i> for creating and editing electronic document template files
<i>IPT</i>	<i>Image Processing Tool</i> for preparing raw document page images for page content retrieval with OCR
<i>OCR</i>	<i>Optical Character Recognition</i> tool for extracting text (strings of characters) from document page images.
<i>GED</i>	<i>Generator of Electronic Documents</i> for editing content files interactively



VED	<i>Viewer of Electronic Documents</i> , a multivalent browser for browsing, annotating and linking content of selected documents
QED	The tool for <i>Quality Evaluation of Documents</i> across all DDLC phases
WR	Working Repository, an internal DDW database for storing project data (document template, content and JPEG image files, document quality data and process parameters)

DDW toolset can be used in two possible configurations:

- *stand alone*, without any connection to the working repository, with all relevant files stored directly in a common filesystem, useful when processing just a few documents in manual mode.
- *connected* to the working repository (DDW database), providing a better control on intermediary document forms in between DDLC phases, in particular when operating DDW in a batch mode.

Components of DDW do not bear any specific version numbers, although their features have been evolving in time. A brief history of each component with regard to the project Gantt is summarized in Tab. 1.

Tab. 1: History of DDW components

<i>Tool</i>	<i>Period (project months)</i>	<i>Expansion up-to-date</i>
IDT	15-18	Naming raw image files and directories, metadata description
RLT	15-18	Uploading WR with initial data (raw image and descriptive information)
EDD	15-23	Interactive design of resizable contextual regions
IPT	15-21	Background cleaning and character improvement of machine typed text.
OCR	15-21, 21-22	Embedding DOKuStar OCR tool with pluggable external interface adaptor (not object-oriented and object oriented versions)
GED	15-21	Interactive (manual) edition of region content
VED	21-23	Visual browsing of document layers with lenses
QED	15-23	Manual (tuning) mode operations for all DDLC phases except exploitation

Since the first successful integration of DDW, demonstrated at the meeting in Month 15 (Technical Week I, Gdansk, Apr. 28 - May 4, 2003) they have been expanding in parallel, as autonomous units developed and tested individually by each respective consortium partner. However, the external interface based on XML and agreed in Month 15 guaranteed compatibility (with one exception explained later), so replacing one component by its newer version has been possible without any need to redesign other DDW components.

1.2 Tested features

In order to measure the real advances of DDW beyond the current state of the art in the area extracting content of machine typed documents two quality aspects have been considered, namely *accuracy* and *efficiency*.

Accuracy indicates how good in the content extracted with the OCR tool from the improved image. Based on the algorithmic aspects of DDLC of the respective phase processes (see Deliverable D6 [5]) the following features shall to be considered when planning and executing tests of DDW:



1. Defining region semantics; as indicated in Deliverable D2 [1] any document consists of regions, and each one may contain a *composed* or *tabular text*, splitting further down into more detailed structures. During the experiment all possible types of regions defined formally in the document model (see Deliverable D6 [5]) shall be exercised.
2. Image improvement; since the project assumes embedding a ready-to-use, off-the-shelf OCR, special processing of the color raw image is required before OCR, expecting a binary black and white image. The related processes mentioned below have been explained in detail in Deliverables D4 [3] and D5 [4], but considering them here is important from the point of view of the final OCR output quality:
 - i. *background cleaning* requires testing DDW with documents of various quality, with the originals exhibiting different types of noise and fatigue;
 - ii. *character improvement* also requires documents of various physical condition, but it is also important to make sure that they contain as many characters as possible – first to represent a complete alphabet of symbols, and second, to provide various combinations of character subsets.
3. Content quality; extracted document content will consist of characters, which must be compared by a knowledgeable user to the original (document image) content. In general, evaluation of content quality shall involve two forms of information contained in the document:
 - i. *text*, i.e., machine readable strings to be evaluated with regards to the correctness of their characters
 - ii. *text image and layout*, i.e., information representing text but in a pixel form to be evaluated with regard to the appearance of characters.

Another quality aspect considered for the tests reported in this document is *efficiency*, which indicates what is the relative effort spent by user on tuning and using DDW when generating electronic documents from their paper originals, compared to the cost of creating the same document with alternative tools and methods. From the point of view of efficiency the following features of DDW shall be explored:

1. User's effort; once a document is scanned its processing towards a final content file takes a certain amount of time and number of people. An alternative to using tools in that process is to type each document, what as explained later is the common practice across memorial places known to the project consortium. Having this in mind, measurements shall take into account larger quantities of documents prepared in a longer time period, rather than single 'ad hoc' processed documents in isolated tests. Effort characteristics should indicate average figures, since individual archivists may work with different speed, depending on personal preferences, time of day, other activities performed at the same time, level of training, etc.
2. Time; since DDW provides automation of most of the processes of the DDLC model defined in Deliverable D6 [5], exact measurements of execution time of various component DDW tools is not required, if only the computer used for running DDW during the experiments satisfies some minimum performance criteria. However, three processes of DDLC require interaction with the user, and although supported by the respective tools, users may affect the overall processing time with their actions:



- i. *scanning* requires user interaction when feeding the scanner with paper originals, and next when storing the raw image data in a computer filesystem. Scanner feeding cannot be automated very much, as mass scanning of historical documents has been rejected by the consortium due to the risk of document destruction (see Deliverable D3 [2]). Storing raw image data after scanning requires interaction with the system for file management, i.e., creating and naming files and directories, copying, moving, deleting files etc. Owing to this, file management features of the IDT and RLT tools shall be assessed with regard to the time spent on loading DDW with input data.
- ii. *layout definition* requires user interaction with the EDD tool to define regions, their content, internal structure and attributes, written in a document template file (see deliverable D6 [5]). Owing to the concept of templates, creation of a template file can be performed once for a larger set of documents, so the overall speed-up of processing a batch of documents with one template, compared to processing each document with its individual template, should be observed. Measurements are expected to indicate that such a speed-up is noticeable and realistic.
- iii. *acceptance* of a retrieved document content involves its final edition with the GED tool. This process cannot be automated by using a spell checker for example, since each phrase, word and even a single character requires approval of an expert archivist, before uploading the retrieved document content to the database. Original paper documents may contain errors, i.e., originally misspelled or wrong characters, so correcting them would alter the content of a historical document. On the other hand, the OCR processing may introduce errors if the image improvement is poor, or the original scanned document is illegible. The ideal situation will be when the final document content file will contain exactly the characters that appear in the original scan. Checking that (by visual comparison) and typing in necessary corrections will take user's time. Measurements are expected to indicate whether the effort spent on preparing (tuning) DDLC processes is acceptable compared to the time of correcting the final document content, what in the worst case would mean retyping the document from scratch.

1.3 Features not tested

It shall be noted that during the experiments only the features mentioned explicitly in the previous subsection will be tested. Other features that may affect the overall quality assessment of DDW, but will not be tested are listed below. The features affecting accuracy *aspects* that have not been tested include the following:

1. OCR tool accuracy; the DDW toolset can incorporate any commercial OCR tool, if it only can be controlled via some well defined XML interface. For the current project the DOKuStar OCR tool by OCE has been selected, and no "best OCR tool" alternative is sought by the consortium. However, if future versions of DDW are to incorporate other OCR that DOKuStar some measurements may be relevant with this regard.
2. Consistency and completeness of document data; as indicated before original documents may contain errors. These errors may affect the final document acceptance, as legible but complicated document content may require more effort from an archivist correcting it with GED. This issue has not been taken into account when evaluating tests reported later in this document

3. *Quality of scanning devices*; it has been assumed in all experiments reported in this document that raw image data are of sufficient quality. Experiments aimed specifically at the quality of raw image data (document scans) with regard to various scanning devices were performed in Workpackage WP2 and reported in Deliverable D3 [2]. It has been discovered that scanning in infrared, with optical filtering, as well as with alternative devices (like cameras) reduces in general the quality of OCR output.

Features concerning *efficiency* aspects that have not been tested include the following:

1. *Manual scanning devices*; manual feed scanners may exhibit different speed and access time, determined by particular details of their design. These features have been assumed irrelevant and the origin of measuring document processing time has been chosen as the moment of storing the raw image data file on disk.
2. *Mass scanning*; as indicated before (see Deliverable D3 [2]) mass scanning devices have been rejected for the sake of document security, and no mass scanning has ever been used in DDLC.

1.4 Approach

In order to provide a common basis for an objective comparison of the successive versions of DDW, (current version 1.1.3, and a few more that are planned towards the end of the project) the "benchmark" set of documents has to be selected. The set shall represent various classes of documents from the point of view of their *physical state* and *originality*, content *semantics*, and *quality* levels of paper originals, all of which are assumed to be machine typed documents.

Physical state and originality will be represented by *historical documents* (created in the period between 1939-45), *contemporary documents* (created in the period between 1965-75), and "ideal" documents, specially prepared *fakes* of real historical documents, using the original paper and typing machine from the time of WWII possessed by the Stutthof Museum.

Semantic classes of documents shall satisfy the above mentioned criteria and constitute the richest possible classes of documents in the archive. Based on the analytic work of Workpackage WP1 reported in Deliverable D2 [1] three semantical classes of documents have been selected: *transport lists* documenting shipments of groups of prisoners between camps, to the camp, and very rarely out of the camp, *personal cards* (records) of individual prisoners held in the camp, and *index cards* documenting archival items.

Quality classes of documents include five standard levels: *very low* (VL), *low* (L), *medium* (M), *high* (H) and *very high* (VH).

Measurements will involve global DDLC quality, relating the input material quality, i.e., visual assessment of a raw image of each document selected for processing, to the output quality, i.e., extracted document content accuracy combined with the total development time up to the final document acceptance.

2. Test procedure

The objective of a testing procedure governing the experiments reported in this document is to compare and contrast current practices used in memorial places for generating electronic

documents from their paper originals to the potential brought to the scene by functionality of DDW and other tools available on the market.

2.1 Purpose

Test procedure specified in this document is aimed at measuring objectively advances of DDW beyond the current practice and state of the art. This has been made possible by processing a benchmark set of documents in several possible ways, including current practices, a new DDW based approach, and several other approaches based on the tools currently available, but not necessarily in a wide use.

2.2 Scenarios

The test procedure includes four scenarios, each one involving the same set of benchmark documents and performed independently. In each scenario the same set of metrics is collected and the same quality evaluation model used. Scenarios may be performed in an arbitrary order, however the sequence described below specifies the range of methods from intuitively the most to the least costly approach (and efficient) approach.

2.2.1 Retyping

Current practices at memorial places across the world involve mostly manual retyping of a document and relies entirely on human expert interpretation of the content (see Figure 1)

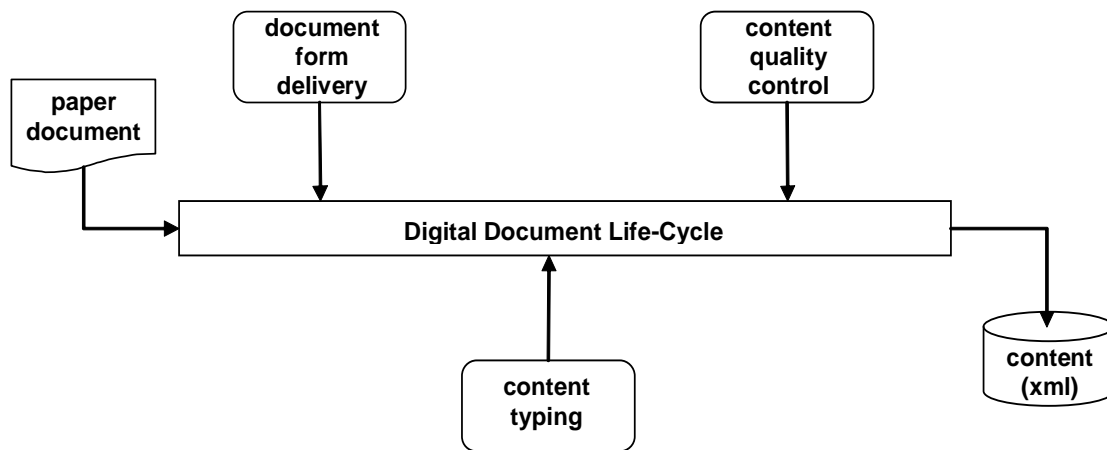


Figure 1: Manual document form filling

Paper document is used as a basis for an electronic form (HTML for example) designed and created by a third party on request of the archive staff. The form is a collection of uniquely named fields of specific format. After some time the form is delivered and a digital document life-cycle starts. Each respective field of the form is filled (typed in) by an archivist. Owing to the predefined structure and layout of the electronic form, upon completing all its fields a document content can be moved to the respective tables in a database automatically.

This practice can be summarized in the following way:

- document form delivery is a time consuming operation, requiring interaction with an external company developing it, and selling as any other software product. For example, a typical form requested by the Stutthof Museum from Neurosoft cost was about 1000 Euro and its delivery time was 3 months;
- manual form filling is a time consuming process, as it relies entirely on human interpretation of the document image. For example, with typing in a single personal record taking about three minutes, and another two for its verification, processing of one class of documents in the Stutthof Museum with 32000 records took about five years of work of two archivists;
- document content quality control concerns only textual content, as no information on the original document layout can be stored in the form. Moreover the inspection is only visual and subjective, as no formal quality metrics may be applied.

2.2.2 Raw image OCR

Direct application of OCR tools available on the market to machine typed historical documents is possible, although no popular across archive sites, as such documents are of much lower quality than “fresh-printed” office documents, for which most of the existing OCR tools have been developed. The respective digital document life cycle development may look like the one shown in Figure 2.

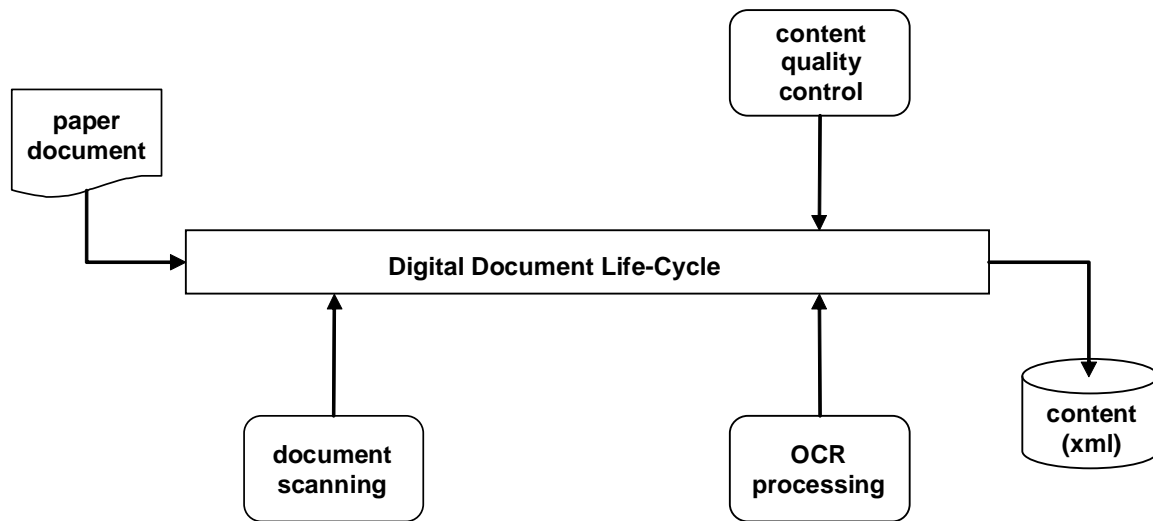


Figure 2: Scanning directly into OCR

Experiments with several leading OCR tools performed by the consortium indicate that this approach is not suitable for machine typed and often fatigued historical documents. In summary the following remarks are in order:

- document scanning is time consuming and prone to error because of the problems with file namespace management. Specifically lack of mechanisms preventing users from accidental file duplication or overwriting reduces overall process efficiency, which is hard to trace.
- advanced OCR tools require interactive input manipulation and manual tuning for best performance. This kind of expertise is hard to acquire for the average historical archive staff.

- document content quality control involves only textual content as no layout information is preserved by OCR. Moreover, errors injected by the OCR during the extraction process can be corrected only manually with a text editor.

2.2.3 Context-free image improvement

Since direct application of OCR tools seems to be impractical due to poor quality of the original document image an approach involving image improvement may be attempted. Some off-the-shelf image processing tools, like Paint Shop Pro or GIMP may be used to filter out noise and some artifacts interfering with the OCR process. Consider Figure 3 outlining the approach.

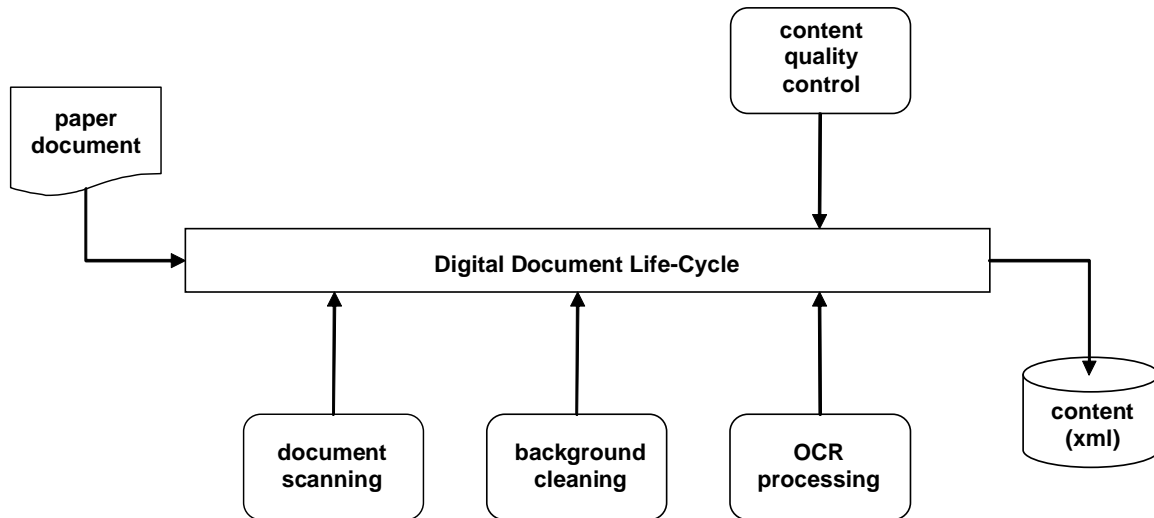


Figure 3: Content extraction with context-free image improvement

Although image improvement for selected documents may be quite impressive in some cases, cleaning noise from the document background with stand-alone tools mentioned before can be applied only to the entire document, without any respect to the local contexts of various regions of the document. In summary the following remarks apply:

- stand-alone tools that may be used in the improvement process have to be purchased separately, as they usually are not associated with any particular OCR;
- selection of image processing methods is manual, so is the subsequent tuning of OCR for best performance. This requires user archivists to have an advanced knowledge on image processing;
- content quality control relies on implicit evaluation, as the cleaned document quality can be assessed after inspecting OCR output;
- textual content is structured owing to the use of XML, so the original document layout can be partially preserved, but corrections of OCR errors require user ability to manually edit XML code.

2.2.4 Semantics-driven image improvement

Document image improvement that takes into account context of each region requires integration of several methods in one document. This requires introduction of a document template, defining semantics of each region of interest to enable the image improvement tool to treat each

region separately for the best result. This is the approach developed in the project and experiments described in this report were supposed to provide evidence that this approach is optimal and cost effective compared to the current practices (manual form typing) as well as approaches that may use advanced stand-alone tools available on the market today on in the near future. Consider Figure 4, outlining the approach introduced by MEMORIAL.

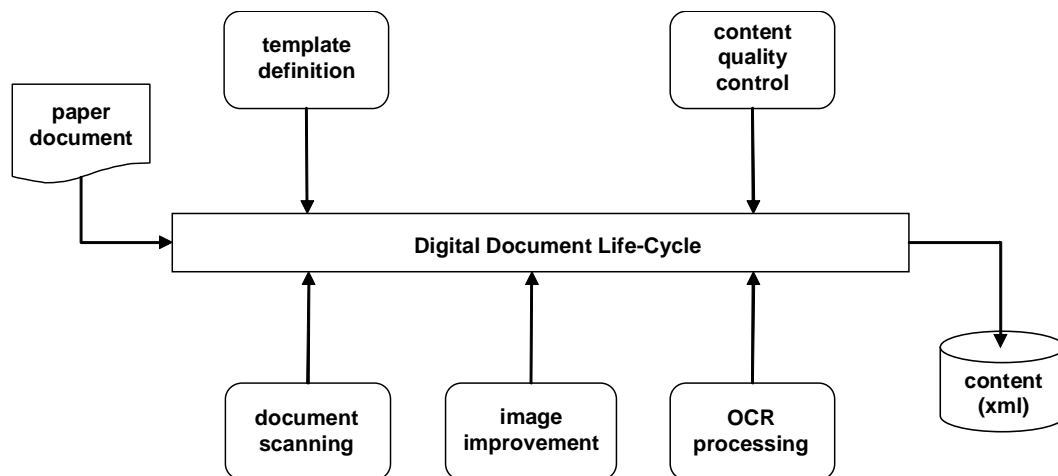


Figure 4: Semantics driven image improvement

Note that owing to the template defining the original document layout and structure the background cleaning process concerning *removal* of pixels representing noise can be integrated with the process of character improvement, concerning actual *adding* of pixels (for details see Deliverable D4). The following remarks apply before presenting concrete results of the experiments later on:

- file namespace management is provided by the indexing tool (IDT), making the scanning process robust and precise;
- template creation and edition with the EDD tool enables direct manipulation of document regions, their content, structure and semantics;
- image improvement integrates background cleaning and character improvement in one tool IPT;
- any XML controlled OCR (off-the shelf tool) can be used;
- content quality control is supported by several tools: a template driven document content editor GED, textual content and layout viewing tool VED, and multi-phased quality monitoring and tuning tool QED.

2.3 Measurements

A comprehensive set of metrics has been defined in order to capture all major features of DDW including both accuracy and efficiency, as described in Section 1.2 before. The number of metrics had to be reasonably small, to avoid excessive amount of data to be processed when testing DDW repeatedly with different classes of documents, but on the other hand should be meaningful and indicative. Given the major features of DDW the following metrics have been selected:

1. An average OCR confidence level (so called *OCR quality*), returned by the OCR tool and indicating a number of alternatives taken when recognizing a single character;

2. Percentage of characters correctly recognized, indicating extracted text quality from the lexical/syntax point of view;
3. Percentage of words correctly recognized, indicating contextual aspect of recognized characters and more semantics oriented;
4. Document preparation time ratio, calculated as $(1-t/t_{max})$, where t represents a total time spent by an archivist on correcting the document content after extraction with DDW and before accepting it to the archive, and t_{max} represents a total time of manual creation of the document content using the form filling method described in Section 2.2.1.

Assessment of quality of each document was based on a quality tree, shown in Figure 5.

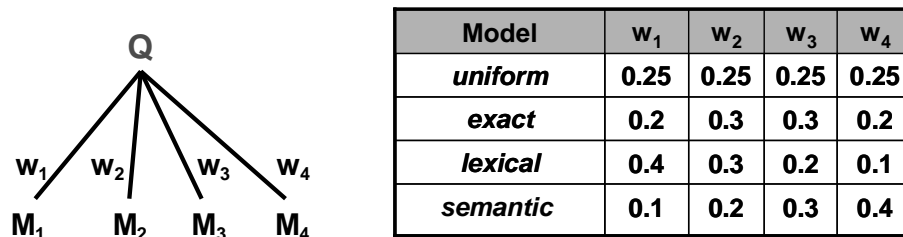


Figure 5: Quality tree and models used for testing DDW

Leafs M_1, \dots, M_4 represent values of the respective metrics defined before, while weights w_1, \dots, w_4 determine contribution of each metric to the overall quality of the final document content, according to the formula $Q=w_1*M_1+w_2*M_2+w_3*M_3+w_4*M_4$. Each vector $\langle w_1, w_2, w_3, w_4 \rangle$ of specific values of weights represents a quality model, emphasizing various aspects of the assessed DDW output. During the experiments the following models were used:

- *uniform*, where each metric had the same significance for the final document content quality;
- *exact*, which focuses more on characters and words, thus emphasizes quality of text rather than OCR process difficulty and human effort in correcting errors;
- *lexical*, which emphasizes character related aspects of the text recognition processes;
- *semantic*, where the meaning of the text counts most, i.e., correct representation of each word of the text, even at the cost of manual edition counts the most.

If DDW and the supported technology constitute a real advance beyond the current state of the art in converting paper documents into electronic ones, each quality evaluation model shall indicate that. Results of experiments in that regard have been described in this report later on.

3. Test cases

Along with defining an appropriate quality evaluation model a representative set of test cases (document samples) had to be selected. These documents should be representative in the following ways:

- *historical significance*; most interesting documents from the point of view of successful extraction are those that contain important and useful information.
- *physical state*; although paper documents in the archive may be of the same type and structure, their individual samples may exhibit various level of legibility to automatic content extraction tools, strongly determined by their physical state.



- *size*; document scanning may generate image files ranging in size, depending on physical dimensions of individual sheets, total number of characters typed in them, as well as the resolution used.
- *content semantics*; since DDW relies on semantic information, document templates required for various classes of documents selected for the experiments should exercise possibly all tags and their attributes of the document model (see Section 3.2 of Deliverable D6).

3.1 Features

The following classes of documents have been selected to provide test cases for the experiments:

- transport lists, including single and multiple page documents,
- personal cards,
- index cards,
- specially prepared “real fakes”.

It will be argued throughout the rest of this section that they are representative enough to serve as a basis for DDW validation, since they contain all elements of the generic document model introduced informally in *Deliverable D2* and next defined formally in *Deliverable D6* (see Table 2).

Table 2: Checklist for DDW validation test cases

<i>Elements of the document model (XML tags)</i>	<i>Test cases</i>			
	<i>Transport lists</i>	<i>Personal cards</i>	<i>Index cards</i>	<i>“Real-fakes”</i>
page	√	√	√	√
content	√	√	√	√
region	√	√	√	√
text	√	√	√	√
composed_text	√	√	√	√
line	√		√	√
word	√	√	√	√
characters	√	√	√	√
hw_mark		√		
predefined_string	√	√	√	√
tabular_text	√			√
row	√			√
cell	√			√
image	√	√		√
signature	√	√		√
stamp		√		
photo	n/a	n/a	n/a	n/a
hw_note		√		
graphics	n/a	n/a	n/a	n/a
line_segment		√	√	√
background	√	√		
corner	√			
edge	√			

Two elements of the document model, namely **photo** and **graphics** did not appear in any class of documents that might be considered significant (neither MST nor Moreshet archives), therefore they have not been tested.

3.1.1 Transport lists

Documents constituting transport lists are the most difficult ones, as their individual pages can differ significantly. Therefore they have been split in four classes, each one corresponding to a different document (page) template. Sample pages of each class have been shown in Figure 6.

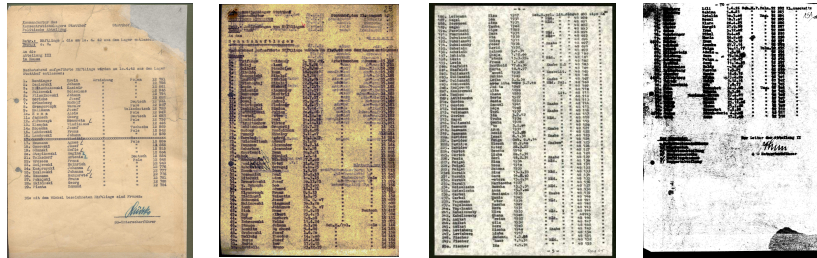


Figure 6: Four classes of transport lists used in tests

They are respectively (from left to right): a single page document, and next front, middle and last pages of a multi-page transport list. It can be seen that their major component is a region representing a tabular text. These are historical documents with a lot of important information, however many of them are fatigued and hard to read.

3.1.2 Personal cards

Personal cards differ significantly from transport lists in their layout and content, as they contain only single lines of composed text. Most of the personal cards in the archive of MST are printed forms filled with a machine typed text, but a small subset containing documents machine typed from scratch has also been found (see Figure 7).

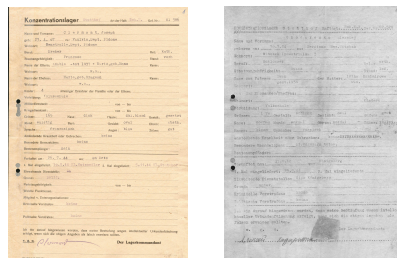


Figure 7: Two types of personal cards used in tests

These documents constitute also an important class of historical documents with useful (and most complete) personal information, and except the small subset of typed (not filled) personal cards, they are well preserved and legible.

3.1.3 Index cards

DDW validation tests included also index cards, contemporary machine typed documents, which have no historical value but interesting to archive staff due to archive related information, collected once in the past when cataloguing resources of MST archives. A sample document of this class is shown in Figure 8.

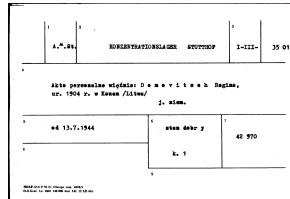


Figure 8: Index card used in tests

Documents of this class are of excellent condition and their successful processing with DDW tools might aid many museums and archives, as this is a standard form used in libraries across Poland.

3.1.4 Fakes

Along with real documents, a class of “ideal” (benchmark) documents have been prepared by MST archivists, taking an advantage of the fact that an original (historic) typing machine used by the WWII administration of the Stutthof concentration camp has been preserved, as well as the original paper used by camp administration. Although the fake documents contained no useful information, they were free of artifacts and damages, except the results of the natural aging process of the paper itself. They have been treated then in the validation tests as a set of “new” documents, of the highest possible quality, a perfect basis for comparison with their original (damaged) counterparts. It was particularly important to have benchmark documents for transport lists, as the originals were the most difficult test cases selected for DDW validation. A sample fake transport list is shown Figure 9.

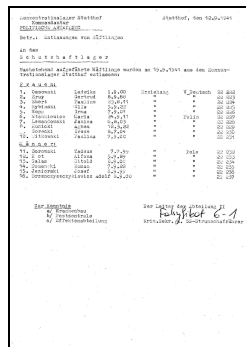


Figure 9: A "real fake" transport list used in tests

Another reason for preparing fake documents, has been a need to circumvent legal barriers in using personal data for integration testing of DDW components by physically distributed partners (for details on legal aspects see *Deliverable D2*). Since fake documents contained no information on any real person they could have been sent over Internet and used for demonstration purposes at the project Web site.

3.2 Required techniques

Proper execution of test scenarios intended for DDW validation required test staff with the following skills and knowledge:

- manual OCR tuning,
- background cleaning with Paint Shop Pro,
- using all available functionality of DDW.

They are described in more detail below.

3.2.1 Manual OCR tuning

Validation experiments reported in this document involved only one OCR product, DOKuStar by OCE. To reduce the effort and avoid excessive number of parameters to be tuned the following parameters have been selected:

- *grow/erode*, determining maximum depth (in pixels) of smoothing outer edges of each character during the extraction. The range assumed for the experiments varied from -2 (erosion) to $+2$ (growth);
- *horizontal line removal*, an off/on flag for detecting and removing horizontal lines and assuming values 0 (off) or 1 (on);
- *isolated despeckle*, defining the maximum diameter of groups of pixels considered a noise that should be filtered out. The range assumed for the experiments varied from 0 (no noise) up to 20 (stains smaller than a single character);
- *confidence*, determining the size of a set of alternative characters analyzed by OCR when looking for the best guess. The range assumed for the experiments varied from 0 (first best fit selected) to 50 (maximum number of best fits that might be considered).

3.2.2 *Background cleaning with Paint Shop Pro*

PaintShop Pro ver. 8.0, a commercial product by Jasc Software used in the experiments is an advanced tool for image processing. Exercising of all its features during the experiments was not in the scope of the validation exercise, since as mentioned before background cleaning with stand alone tools always concerned the entire document page. Nevertheless, two classic binarization methods of document images have been used: noise reduction with thresholding, and advanced binarization with the Otsu method.

3.2.3 *Available functionality of DDW*

Functionality of the project product required for its validation testing has been achieved a few months ahead of the time of the scheduled date of the validation event (conclusion of workpackage WP5) close to the project end, as due to the very strict project timetable there was practically no room for delaying the necessary level of DDW maturity. Therefore a preliminary validation (public demo for non project users at Yad Vashem archives in Month 23) preceded a principal validation (concluded with presentation of test results at the second annual project review in Month 24). In summary, for successful completion of validation tests, the required functionality of DDW involved:

- management of document image file namespace;
- interactive document template definition;
- direct manipulation of document regions, their content, structure and semantics;
- image improvement including background cleaning and character improvement tool;
- integration with the embedded off-the-shelf OCR tool;
- interactive edition of document content;
- multi-layered document viewing;
- multi-phased quality monitoring and tuning;

Respective DDW components supporting the required functionality have been listed in Table 1 before.

3.3 Environmental needs



THE VIRTUAL MEMORIAL PROJECT

Validation of DDW concerned several versions of component tools, configured to support various scenarios of use and speed up the related activities by splitting the work between various types of computers. Additionally, the experiments were intended to demonstrate versatility of DDW and its stability of performance on various machines. The following configurations have been tested:

1. Stand-alone version, without a working repository, document files stored in a standard personal directory of the user, with only EDD, GED and VED tools, for dynamic creation of document forms and manual generation of documents;
2. Stand-alone version as in p.1, but with the IPT tool, intended for preparation of images for off-line (independent of DDW) processing with OCRs;
3. Stand alone version as in p.2 but integrated with OCR;
4. Integrated client-server configuration with the working repository for batch processing with client and server located at the same computer
5. Client-server configuration as on p.4, but with clients and server located at different computers.

Minimum requirements for the testing configurations listed above are specified in Table 3.

Table 3: Minimum requirements for validation testing configurations

#	CPU	RAM	Disk space	System
1	700 MHz	128MB	5 MB working, 40 MB images	<ul style="list-style-type: none"> - MS XP with SP1, or MS2000 with SP4 - MS .NET Framework 1.1
2	1.2 GHz	256 MB	5 MB working, 40 MB images	<ul style="list-style-type: none"> - MS XP with SP1, or MS2000 with SP4 - MS .NET Framework 1.1
3	1.2 GHz	256 MB	5 MB working, 40 MB images	<ul style="list-style-type: none"> - MS XP with SP1, or MS2000 with SP4 - MS .NET Framework 1.1 - DOKuStar 3.x Edition version with Hardlock key
4	1.5 GHz	256 MB	5 MB working, 40 MB images	<ul style="list-style-type: none"> - MS XP with SP1, or MS2000 with SP4 - MS .NET Framework 1.1 - DOKuStar 3.x Edition version with Hardlock key - MS SQL Server 2000 Developer Edition with SP3
5	1.5 GHz (server) 800 MHz (client)	256 MB	5 MB working, 40 MB images	<p><u>server:</u></p> <ul style="list-style-type: none"> - MS XP with SP1, MS2000 with SP4, MS 2000 Advanced Server with SP4, or MS 2003 Server - MS .NET Framework 1.1 - DOKuStar 3.x Edition version with Hardlock key - MS SQL Server 2000 Developer Edition with SP3 <p><u>client:</u></p> <ul style="list-style-type: none"> - MS XP with SP1, or MS2000 with SP4 - MS .NET Framework 1.1 - DOKuStar 3.x Edition version with Hardlock key - MS SQL Server 2000 Developer Edition with SP3

3.4 Intercase dependencies

Although all selected document classes were significantly different from the point of view semantics, as described in Section 3.1, they exhibited a number of dependencies. This property of test cases is strongly required in acceptance and validation testing, as it reinforced diversity of testing contexts of the product features under test. In particular:



THE VIRTUAL MEMORIAL PROJECT

- all font characteristics important from the point of view of character improvement and recognition were shared by transport lists, personal cards and fake documents, as they have been typed on the same (or similar) machine.
- document background characteristics, related to paper aging were shared by transport lists and fake documents, as they have been typed on paper of the same origin and state of preservation.
- two main types of document layout: tabular-based and form-based were present. Features of tabular-based documents were shared by transport lists and fake documents (fake transport lists in fact), while features of form-based documents were shared by personal cards and index cards.

4. Test items

Test items, including DDW components integrated and configured in five different versions listed in Table 3 have been installed on several computers, ranging from the small laptop up to a server, along with image files of documents split in several groups constituting a set of test cases described in Section 3.

4.1 Location

Respective configurations (see Section 3.3) of DDW have been installed at the following machines:

- #1 and #2 on the least powerful laptop Compaq Evo N110 (996 MHz, 250 MB RAM),
- #3 and #4 on two more powerful laptops, namely Acer TravelMate (1.6 GHz, 512 MB RAM), and Dell Latitude D400 (2.0 GHz, 512 MB RAM) and
- #5 on two main project servers, Dell PowerEdge 4000 (2x1.5 GHz, 1 GB RAM), one located at TUG and another at MST.

Compaq, Acer and Dell laptops were used for demonstration and exploitation of DDW at different geographical locations (Berlin, Tel Aviv, Luxembourg and Sztutowo), while the main project server (docmaster.eti.pg.gda.pl) was used by MST and TUG testers (including TUG students of the course on Software Quality participating in DDW testing exercises).

Groups of documents constituting a set of validation test cases included about 20 image files in TIFF format per each class (single and multi-page transport lists, personal cards, index cards), and a few fake transport lists, over one hundred image files in total. They have been selected from about 300 GB of image data provided by MST.

4.2 Status

DDW configurations listed before and considered for validation tests have been unit- and integration-tested by the responsible consortium partners and accepted for public demonstration prior to the final validation exercise in Month 24.

Document images selected as test cases for DDW validation constituted final scans, intended by MST for complete processing upon concluding workpackage WP5 and developing a virtual memorial database.

4.3 Approvals

Owing to the sensitive copyright and personal data issues transfer of document scans to the main project server at TUG required approval of their owner, MST. Such an approval was granted by MST and having two “twin” project servers (‘archmaster’ at MST and ‘docmaster’ at TUG) was an



important advantage: four disk units (each one with 75 MB with image data) were simply brought from MST to TUG and plugged for the duration of tests into the 'docmaster' frame. In this way no need existed to copy large volume of image files, nor to transfer them via Internet. This operation was supervised by MST server administrator.

5. Test incident report

A principal objective of validation tests has been measurements and evaluation of quality of electronic documents and quality of the document engineering process with and without DDW tools. During the tests, however, some unexpected situations occurred, revealing programming errors in some units, as well as a few architectural and conceptual flaws, all of which have been readily corrected. It shall be emphasized that DDW components and their units were rigorously and carefully tested during prior DDW integration by responsible partners, and all detected errors were duly eliminated.

5.1 Description

Unit errors revealed during the validation exercise concerned template editor EDD and image improvement tool IPT. EDD related incidents involved several exceptions not appearing before, and related to various hardware configurations used when testing DDW. They were:

1. Various unhandled mouse events (different units with different user timing),
2. Exception raised by region reopening operation (new user with different habits)
3. Negative row shift, by one pixel outside of a tabular text region (spotted only at the slowest laptop used during the test).

IPT related incidents were observed when adding new test cases (documents) to the test suite used before during unit testing:

4. Lines of white pixels stretching over characters in some lines of cleaned images of fatigued transport lists,
5. Inter-character noise introduced to small regions in cleaned document images of fatigued transport lists,
6. Unnecessary dotted line reinforcement in cleaned images of personal cards.

Architectural flaws spotted when exercising validation test cases concerned only one incident, namely:

7. Parsing of OCR output (XML files) produced by two consecutive versions of DOKuStar (used respectively by TUG and GFaI).

Conceptual flaws identified with the help of validation test cases concerned the following incidents:

8. Region matching in documents belonging to the same class, i.e., described with the same template, as actual class documents had some corresponding regions shifted from the origin defined in the class template);
9. Parsing of a **predefined_string** element specified in the respective document class template resulted later on in a content file with **word** and **character** elements substituting the former, causing certain problems in locating some words by the content editor GED.

5.2 Impact

Incidents 1-3 listed before caused only minor inconvenience during testing, as a simple way around it was available, and the exceptions did not cause any failure.

Incidents 4-6 affected slightly quality of the OCR process only for a few documents, and were not able to delay or stop the validation exercise.

Incident 7 affected the validation exercise in that it could have been continued only by TUG. This was not a big problem either, as all validation test case documents belonging formally to MST were located at TUG server, as explained in Section 4.3.

Incident 8 caused some delay in completing the validation exercise, as it indicated a need for introducing in template editor EDD an additional algorithm specifically for automatic region matching, and a slight modification of the basic EDD data flow, as described in the next section.

Incident 9 also indicated a need for some modifications (in the main GED algorithm), but it practically had no impact on completing the validation exercise, as a way around that problem was possible (avoid **predefined_string** elements in templates of documents scheduled for automatic processing)

5.3 Solutions

All issues brought to the attention of consortium partners during incidents described before have been solved. In particular,

- reasons for the reported exceptions have been eliminated from the DDW code,
- some internal parameters of IPT have been tuned for better performance,
- new parser for XML output produced by DOKuStar has been developed
- an additional pixel analysis algorithm has been developed and added to EDD, which now offers extra functionality for automatic shifting and resizing of regions,
- usage of **predefined_string** elements in document templates have been substituted by word elements, which are automatically associated with dictionaries, used next by GED during document content edition.

6. Test summary

Validation of the DDW toolset is completed, which has now achieved a level of maturity suitable for its exploitation as a beta-release product. Nevertheless its testing (now maintenance testing during its regular use by MST archive staff, and TUG students in selected courses) is continued.

6.1 Variances

The test procedure adopted for DDW validation exercise assumed in principle a selection of five representative documents for each class of documents (see test cases specified in Section 3.1), with regard to five quality levels: very high (VH), high (H), medium (M), low (L) and very low (VL). The sample set was next used to tune all respective phases of the DDLC document engineering model (see Deliverable D6) for the best possible quality. Next the remaining documents of the class were supposed to be processed in a batch mode, with parameters set up for the representative set.

In some cases, however, minor deviations from the principal testing procedure have been accepted. The involved:

- selection of just VH, M, VL samples for personal and index cards, what helped to shorten time of the validation exercise, without compromising on test comprehensiveness, as the document classes mentioned here were of a relatively good quality and did not require finer tuning.
- selection of only VH samples for fake documents, as they all were of ideal quality anyway.

6.2 Comprehensiveness assessment

Features that could have been tested more during the reported exercise concerned **tabular_text** in transport lists. These particular structures could be arbitrary complex, owing to practically unlimited number of possible combinations of **tabular_text** components. The test completion criterion assumed for the testing exercise reported here required complete coverage of each element in the generic document tree (see Deliverable D6), but not each possible combination of such elements. Given the nature of testing, some combinations of elements not appearing in the set of test cases used for the reported validation exercise may potentially reveal new errors.

6.3 Summary of results

Sample measurements of metrics M1-4 introduced in Section 2.3, which have been collected for selected documents from the set of test cases described in Section 3.1 are listed in Table 4. Test IDs represent the following scenarios for document content extraction:

1. Scanning directly into OCR (see Figure 2);
2. OCR after binarization and tresholding with Paint Shop Pro (see Figure 3)
3. OCR after Otzu binarization with Paint Shop Pro (see Figure 3)
4. Stepwise extraction with DDW (see Figure 4)

Moreover, values of metric M4 reported in Table 4 required values of parameter t_{max} , defined in Section 2.3. These have been measured during manual form filling (see Figure 1), where forms have been generated with the use of template editor EDD and filled with the use of content editor GED.

Table 4 : Measurements for selected document classes

<i>Document class</i>	<i>Document ID</i>	<i>Test ID</i>	<i>M1</i>	<i>M2</i>	<i>M3</i>	<i>M4</i>
Personal cards	1. I-III-10280.tif	1	0,79	0,83	0,50	0,52
		2	0,78	0,84	0,47	0,36
		3	0,79	0,84	0,57	0,56
		4	0,75	0,71	0,37	0,88
	2. I-III-11698.tif	1	0,68	0,47	0,33	0,48
		2	0,86	0,78	0,67	0,28
		3	0,80	0,69	0,47	0,56
		4	0,86	0,81	0,70	0,88
	3. I-III-17417.tif	1	0,85	0,82	0,70	0,60
		2	0,85	0,63	0,67	0,44
		3	0,85	0,70	0,70	0,56
		4	0,80	0,56	0,40	0,88
Index cards	4. 0000017F.tif	1	0,67	0,84	0,60	0,00
		2	0,81	0,88	0,60	0,00
		3	0,70	0,86	0,60	0,00
		4	0,89	0,97	0,90	0,80
	5. 0000018F.tif	1	0,71	0,77	0,47	0,00

		2	0,81	0,89	0,58	0,00
		3	0,74	0,90	0,68	0,00
		4	0,81	0,94	0,74	0,80
	6. 0000019F.tif	1	0,70	0,80	0,58	0,00
		2	0,85	0,85	0,63	0,00
		3	0,73	0,86	0,63	0,00
		4	0,89	0,96	0,84	0,80
Fakes	7. falsyfikat6-1_medium.tif	1	0,56	0,44	0,44	0,50
		2	0,76	0,87	0,76	0,40
		3	0,77	0,92	0,82	0,48
		4	0,87	0,94	0,89	0,90
Transport list (last page)	8. I-IIb12_151.tif	1	0,27	0,40	0,14	0,46
		2	0,51	0,52	0,45	0,37
		3	0,34	0,44	0,20	0,46
		4	0,21	0,26	0,06	0,89

6.4 Evaluation

Results of DDW validation with test cases described in Section 3.1 are given below in numerical and graphical form. Consider again Figure 5 and four models for assessing document quality, explained in Section 2.3. Quality values of each respective document listed in Table 4, calculated with weights corresponding to each model, have been presented in the following tables. Figures illustrating quality levels based on data from the tables indicate clearly an advance of DDW (Test 4 scenario) over the current state of the art.

Table 5: Quality values according to the uniform model

Document	Personal cards			Index cards			Fakes	Transport lists (LP)
	1	2	3	4	5	6	7	8
Test 1	0,659	0,493	0,742	0,502	0,438	0,496	0,486	0,317
Test 2	0,611	0,647	0,647	0,496	0,496	0,483	0,696	0,462
Test 3	0,689	0,629	0,702	0,514	0,555	0,529	0,747	0,359
Test 4	0,677	0,811	0,659	0,890	0,822	0,872	0,901	0,353

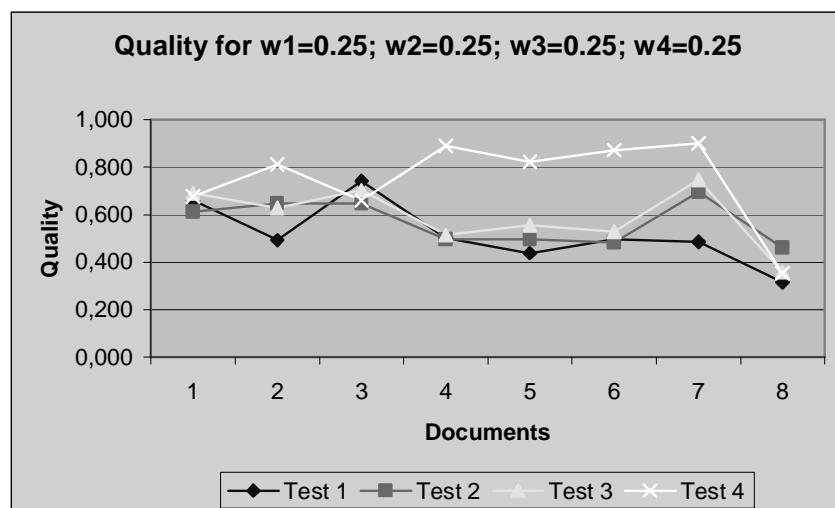


Figure 10 : Comparison of quality levels according to the uniform model

Table 6: Quality values according to the exact model

Document	Personal cards			Index cards			Fakes	Transport lists (LP)
	1	2	3	4	5	6	7	8
Test 1	0,660	0,475	0,745	0,546	0,475	0,535	0,476	0,307
Test 2	0,620	0,663	0,647	0,544	0,544	0,535	0,720	0,467
Test 3	0,692	0,619	0,702	0,557	0,603	0,573	0,772	0,351
Test 4	0,649	0,800	0,623	0,899	0,826	0,878	0,904	0,314

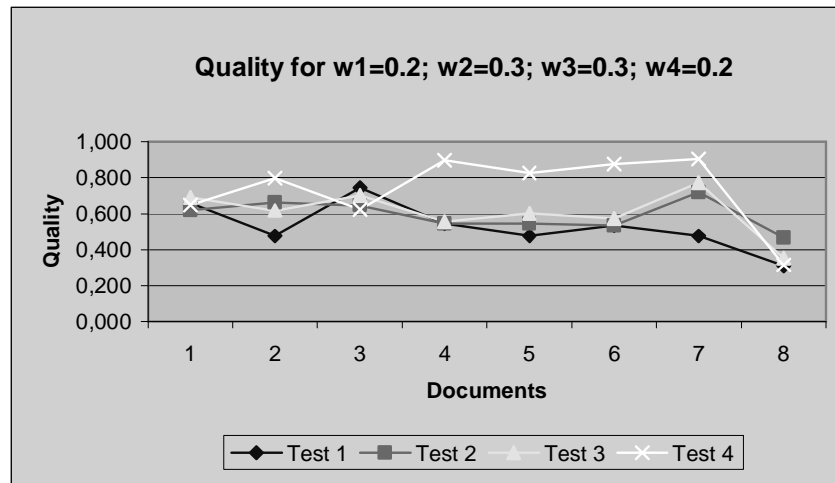


Figure 11 : Comparison of quality levels according to the exact model

Table 7: Quality values according to the lexical model

Document	Personal cards			Index cards			Fakes	Transport lists (LP)
	1	2	3	4	5	6	7	8
Test 1	0,716	0,531	0,785	0,629	0,590	0,627	0,495	0,302
Test 2	0,692	0,740	0,706	0,675	0,679	0,682	0,755	0,486
Test 3	0,738	0,677	0,746	0,646	0,692	0,665	0,796	0,354
Test 4	0,674	0,813	0,655	0,907	0,834	0,891	0,900	0,261

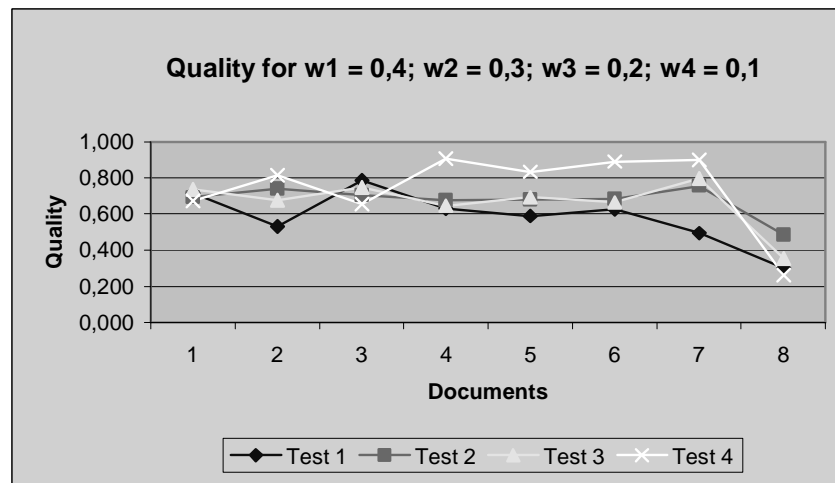


Figure 12: Comparison of quality levels according to the lexical model

Table 8: Quality values according to the semantic model

Document	Personal cards			Index cards			Fakes	Transport lists (LP)
	1	2	3	4	5	6	7	8
Test 1	0,602	0,455	0,699	0,375	0,287	0,365	0,476	0,332
Test 2	0,530	0,554	0,587	0,316	0,313	0,285	0,638	0,438
Test 3	0,641	0,582	0,659	0,381	0,419	0,394	0,698	0,365
Test 4	0,679	0,809	0,664	0,873	0,810	0,853	0,903	0,445

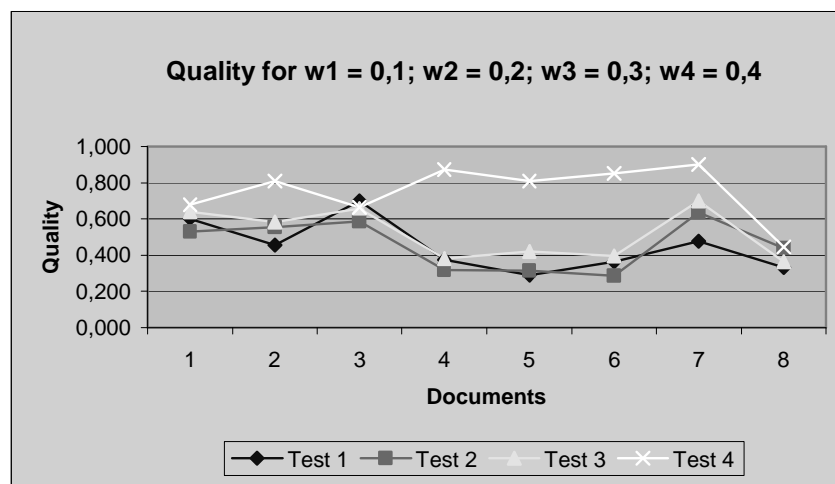


Figure 13: Comparison of quality levels according to the semantic model

6.5 Summary of activities

In summary, it shall be stressed that the validation exercise has been completed successfully and on time: selected test cases constitute a meaningful set of documents, with a great potential for batch processing, and of significant interest of archive staff in processing them to the interactive, electronic form.

7. Recommendations

Although incidents reported before indicate some room for improvement of the DDW toolset, a major product of the MEMORIAL project, it is mature enough to enter the beta-phase of its exploitation by interested memorial places, museums and archives outside of the project consortium.

Bibliography

- [1] Deliverable D2: Specification of a personal record paper document layout, structure and content
- [2] Deliverable D3: Specification of the integration concept for document input
- [3] Deliverable D4: Knowledge based preprocessing
- [4] Deliverable D5: Document input and knowledge based preprocessing
- [5] Deliverable D6: DDW and information exchange infrastructure