

DELIVERABLE

Project Acronym: Europeana Newspapers
Grant Agreement number: 297380
Project Title: A Gateway to European Newspapers Online

D3.3 Evaluation Tools – Final Versions

Revision: 1.0
Authors: Stefan Pletschacher, USAL
 Christian Clausner, USAL
 Christos Papadopoulos, USAL
 Apostolos Antonacopoulos, USAL

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
PU	Public	x
PP	Restricted to other programme participants (including Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium and the Commission Services	

Revision History

Revision	Date	Author	Organisation	Description
0.1	10-07-2014	Stefan Pletschacher Christian Clausner Christos Papadopoulos	USAL	First draft
0.2	17-07-2014	Stefan Pletschacher	USAL	Updated version
0.3	23-07-2014	Stefan Pletschacher	USAL	Appendix added
0.4	27-07-2014	Apostolos Antonacopoulos	USAL	Minor changes
0.5	28-07-2014	Stefan Pletschacher	USAL	Final version for internal review
0.6	29-07-2014	Sandra Kobel	SBB	Internal review
1.0	30-07-2014	Clemens Neudecker	SBB	Final version

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Table of Contents

1. Introduction	4
2. Deliverable	4
3. Tools	5
3.1 File Format Support	5
3.1.1 PAGE XML	5
3.1.2 Page Layout Libraries	6
3.2 Ground Truth Production	7
3.2.1 Aletheia	7
3.2.2 PAGE Converter and Validator	9
3.3 OCR Integration	9
3.3.1 FineReader Integration	10
3.3.2 Tesseract Integration	10
3.4 Evaluation Tools	11
3.4.1 OCR Evaluation	11
3.4.2 Layout Evaluation	11
3.5 Workflow and Auxiliary Tools	12
3.5.1 Extractor/Exporter Tool	12
3.5.2 PAGE Metadata Scanner	12
3.5.3 Feature Extractor	13
3.5.4 Image Characteristics Tagging Interface	14
3.5.5 PAGE Viewer	15
4. Conclusion	16
APPENDIX: Tool Documentations	17

1. Introduction

The analysis and recognition (Optical Character Recognition - OCR) of historical documents, and in particular newspapers, is still a technically very challenging task. Depending on the quality and complexity of the input images this typically leads to processing results that are affected by errors of different degrees of severity. Such errors (and their significance) are measured by comparing the output obtained from a system with an ideal representation of the fully-recognised document (called ground truth). For different aspects of the document information (such as layout, text, and reading order) different metrics are used as a basis for calculating potential deviations from the ground truth.

In order to measure the quality actually achieved on representative examples from the Europeana Newspapers production workflow, as well as to put these results into the context of actual use scenarios (the requirements on a *phrase search* use scenario, for instance, are much higher than those on *keyword search*), it was necessary to implement and extend a number of tools. The main categories of software which were required so as to support all evaluation-related activities in the project (most importantly Tasks T3.4, T3.5, and T3.6) were:

- File Format Support
- Ground Truth Production
- OCR Integration
- Evaluation Tools
- Auxiliary Workflow Tools

In many cases it was possible to build on existing tools, modifying and adding extra features which were required for handling newspapers and the scenario-based evaluation approach taken in this project. The remainder of this document summarises the contributions that were made specifically within the scope of the Europeana Newspapers project.

2. Deliverable

This report gives an account of the tools that were developed as part of T3.3 *Evaluation Tools*. The *actual deliverable* is the software which was (and still is) used to perform evaluation tasks in the Europeana Newspapers project.

Depending on the nature of the individual tools, it was aimed at making them available to the widest possible audience. Any stand-alone software which might be relevant to other researchers in the field is freely available through the partner web-site of the Pattern Recognition and Image Analysis Research Group (PRImA) at the University of Salford (USAL) at <http://www.primaresearch.org/tools>. Tools depending on other components were integrated into the respective services and/or workflows so as to be accessible by users of those (mostly web-based) systems. Examples for this are the integration of the Ground Truth Viewer and the Image Characteristics Tagging Interface into the Image and Ground Truth Repository (<http://www.primaresearch.org/datasets/ENP>, see also D3.2 *Evaluation Dataset*).

3. Tools

In the following, an overview of all the tools and major contributions made within the scope of the Europeana Newspapers project is given.

3.1 File Format Support

There were two main aspects which had to be addressed by implementing and extending support for a variety of OCR-related file formats. Firstly, it had to be ensured that the format used for storing ground truth was covering all relevant features of newspapers. This was achieved by extending the existing PAGE (Page Analysis and Ground truth Elements)¹ format framework in a way that newspapers could be adequately represented (for instance by introducing a new advert region type, allowing nested regions etc.). Secondly, to ensure compatibility of formats required for automation of evaluation workflows and in order to allow meaningful comparison of processing results and ground truth (potentially originating from different tools and source formats) specific file import handlers had to be implemented (e.g. for ALTO, ABBYY XML etc.).

3.1.1 PAGE XML

A great achievement of the Europeana Newspapers project was the release of significantly improved PAGE XML Schemas with regard to page content and layout evaluation data.

- Page content:
<http://schema.primaresearch.org/PAGE/gts/pagecontent/2013-07-15/>
 - More compact (more efficient representation of polygons – required for handling potentially huge files stemming from broadsheet newspapers)
 - More robust (stricter schema – required for detecting potential issues in automated evaluation workflows)
 - New features
 - Nested regions
 - New region types
 - New attributes
 - Baselines
 - New relations
 - Custom fields
- Layout evaluation data (evaluation profiles and results):
<http://schema.primaresearch.org/PAGE/eval/layout/2013-07-15/>
 - Adopted new region types and attributes from the page content XML format
 - More compact
 - Easier to understand
 - Border detection results incorporated

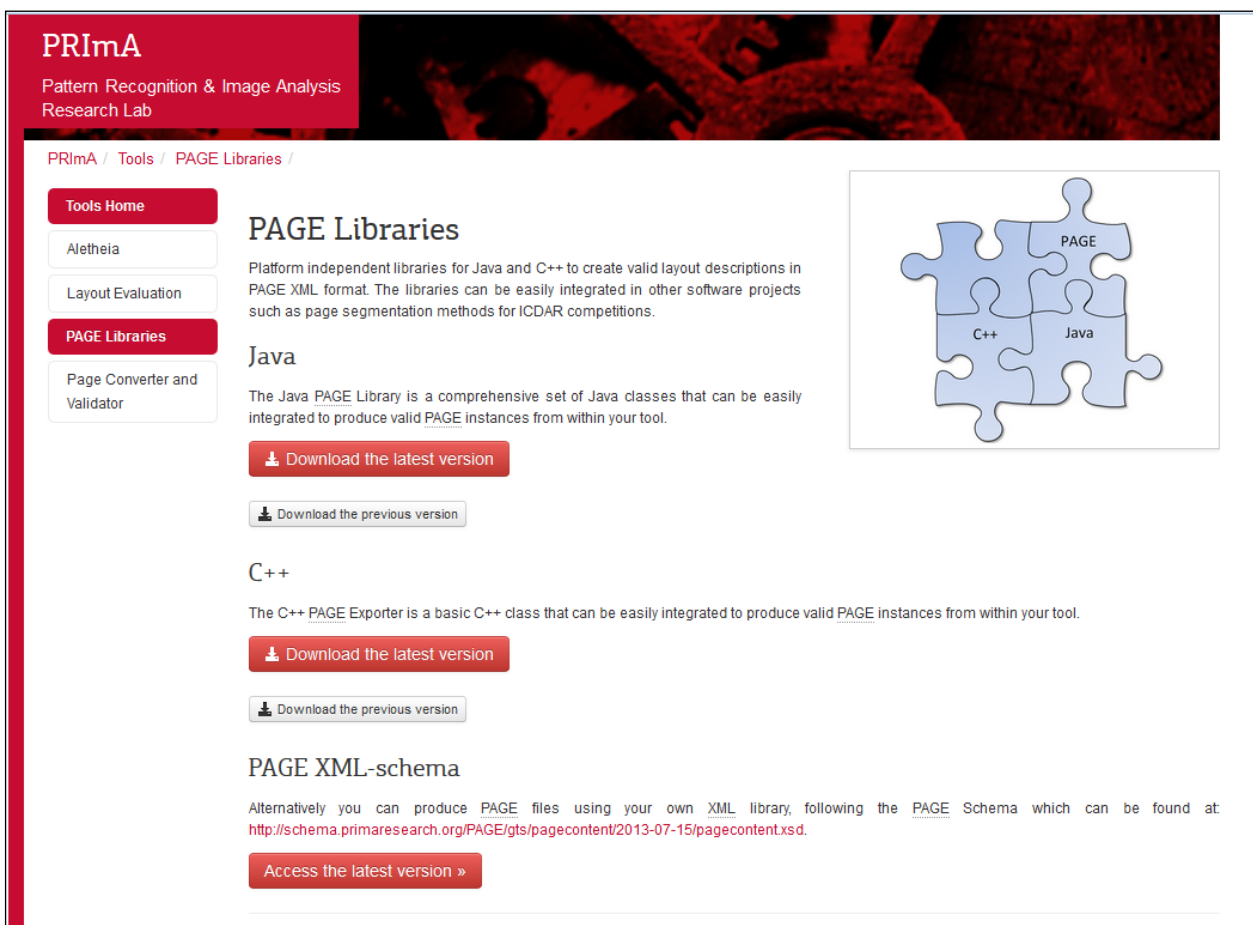
¹ http://www.primaresearch.org/publications/ICPR2010_Pletschacher_PAGE

3.1.2 Page Layout Libraries

In order to make the improvements of the new PAGE XML Schema accessible in all related tools it was necessary to implement support for the new format in the intermediate Page Layout Libraries (feeding the internal, format independent data model of every single tool). To this end, existing C++ libraries (for instance used by Aletheia, the Layout Evaluation Tool etc.) had to be updated and, as a result of new requirements to also drive web-based services, a completely new Java library had to be developed. In addition to supporting all versions of PAGE XML and following the initial requirements from above, both libraries were extended in order to also support

- ALTO 2.0
- ALTO 2.1
- ABBYY XML
- hOCR

The Java version and a simplified C++ version of the PAGE library are available to other researchers and developers to be integrated in their software.



PRImA
Pattern Recognition & Image Analysis
Research Lab

PRImA / Tools / PAGE Libraries /

Tools Home

Aletheia

Layout Evaluation

PAGE Libraries

Page Converter and Validator

PAGE Libraries

Platform independent libraries for Java and C++ to create valid layout descriptions in PAGE XML format. The libraries can be easily integrated in other software projects such as page segmentation methods for ICDAR competitions.

Java

The Java PAGE Library is a comprehensive set of Java classes that can be easily integrated to produce valid PAGE instances from within your tool.

[Download the latest version](#)

[Download the previous version](#)

C++

The C++ PAGE Exporter is a basic C++ class that can be easily integrated to produce valid PAGE instances from within your tool.

[Download the latest version](#)

[Download the previous version](#)

PAGE XML-schema

Alternatively you can produce PAGE files using your own XML library, following the PAGE Schema which can be found at <http://schema.primaresearch.org/PAGE/gts/pagecontent/2013-07-15/pagecontent.xsd>.

[Access the latest version »](#)

Diagram illustrating the components of the PAGE Libraries: PAGE, C++, and Java.

Figure 1: PAGE Libraries on the USAL website

3.2 Ground Truth Production

Ground truth as the reference for all evaluation activities played a crucial role in work package 3. In order to allow partners and service providers to create, validate, and correct ground truth files efficiently as well as support new features introduced with the aforementioned format upgrades, a general overhaul of all related tools (including Aletheia, the ground truth production software which was partly developed during the EC-funded project IMPACT²) was necessary.

3.2.1 Aletheia

Aletheia is specialist software for highly accurate and yet efficient ground-truthing of large amounts of documents. Besides a complete redesign of the User Interface, better following the natural order of tasks related to ground truth production, numerous automated and semi-automated tools were integrated in order to aid the user.

² <http://www.impact-project.eu/>

PRIMA / Tools / Aletheia /

Tools Home

Aletheia

Layout Evaluation

PAGE Libraries

Page Converter and Validator

Aletheia Ground-Truthing System

[Download the latest version](#)

Overview

Aletheia is an advanced system for accurate and yet cost-effective ground truthing of large amounts of documents. It aids the user with a number of automated and semi-automated tools which were partly developed and improved based on feedback from major libraries across Europe and from their digitisation service providers which are using the tool in a production environment.

Novel features are, among others, the support of top-down ground truthing with sophisticated split and shrink tools as well as bottom-up ground truthing supporting the aggregation of lower-level elements to more complex structures. Special features have been developed to support working with the complexities of historical documents. The integrated rules and guidelines validator, in combination with powerful correction tools, enable efficient production of highly accurate ground truth.

System Features

- Mature XML schema which is part of the PAGE format framework
- Targets production environments (large scale digitisation)

Built-in Operations

- Image Binarisation
- Noise Removal
- Automatic Layout Analysis and OCR (using [Tesseract OCR Engine v 3.02](#))

Ground Truth Production

- Border and Print Space
- Layout Regions
- Modification of Layout Regions (merge, split, edit outline)
- Region Attributes
- Text Content (Unicode with virtual keyboard for special characters)
- Reading Order
- Layers
- Text Lines, Words and Glyphs also with text content
- Validation against Ground Truthing Rules and Guidelines

Special Characters

- A list of special characters currently available through the virtual keyboard is maintained [here](#).
- For more details on how to customise the virtual keyboard refer to the Aletheia user guide.

[Download the latest version](#)




Figure 2: Download page for Aletheia – an advanced ground truth production system

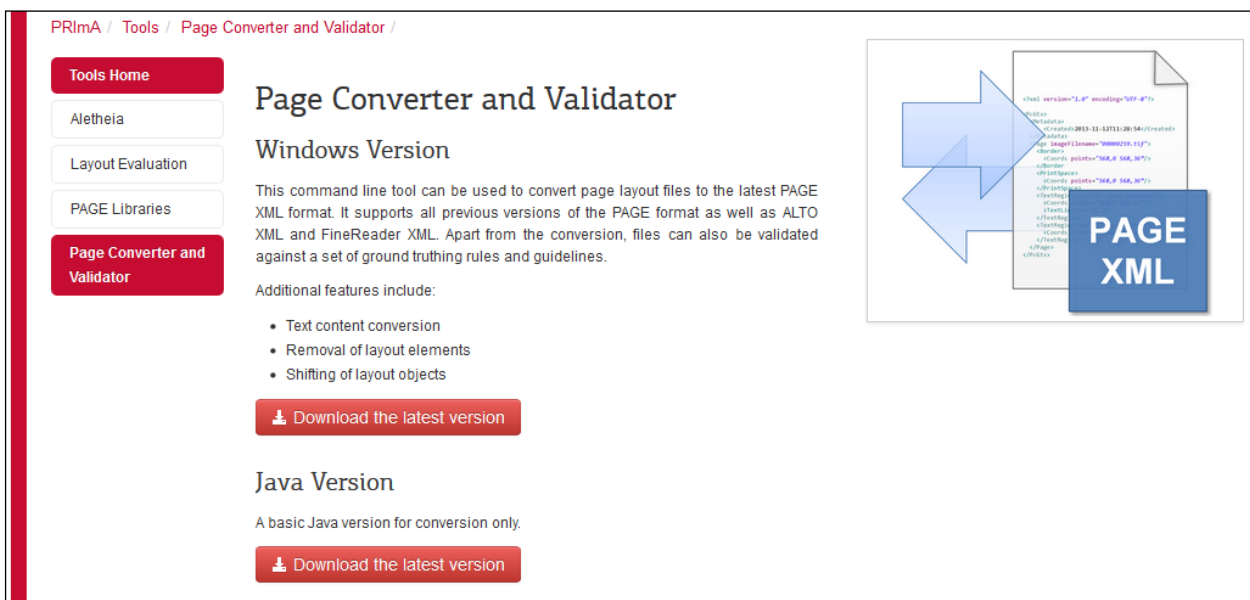
The most significant improvements to Aletheia were:

- Major improvements in user interface (Ribbon Toolbar)
- Tesseract OCR integration
- Image related tools
- Additional manual outline creation/editing tools
- Support of new features introduced with the upgraded PAGE XML format (e.g. base lines, nested regions, etc.)
- ALTO XML (2.0 and 2.1) and FineReader XML support
- Border detection
- Gamera OCR format export
- Image thumbnail (overview with position of current view)
- Support of JPG and PNG images

For more details see the attached Aletheia User Guide (an alphabetical list of all attachments is provided in the Tool Documentations APPENDIX).

3.2.2 PAGE Converter and Validator

In order to ensure consistent quality and compliance with the project-internal ground truth production guidelines, a tool which was already used in IMPACT in the context of historical books was extended and developed further in order to satisfy the more complex requirements of newspapers. Again, this meant implementing support for the new PAGE XML format but also to create more complex rules against which ground truth files were checked as part of the automated quality assurance workflow. Another feature which was integrated into this tool is the ability to convert (and save) files from any of the previous PAGE XML formats to the latest version.



PRIMA / Tools / Page Converter and Validator /

Tools Home

Aletheia

Layout Evaluation

PAGE Libraries

Page Converter and Validator

Page Converter and Validator

Windows Version

This command line tool can be used to convert page layout files to the latest PAGE XML format. It supports all previous versions of the PAGE format as well as ALTO XML and FineReader XML. Apart from the conversion, files can also be validated against a set of ground truthing rules and guidelines.

Additional features include:

- Text content conversion
- Removal of layout elements
- Shifting of layout objects

[Download the latest version](#)

Java Version

A basic Java version for conversion only.

[Download the latest version](#)

Diagram illustrating the conversion process to PAGE XML format.

Figure 3: PAGE Converter and Validator download page

3.3 OCR Integration

The need for integrating third-party OCR-engines (i.e. via their APIs) was mainly due to two reasons: Firstly, accessing recognition results directly from the internal representation of an OCR-engine provides much more detail than can be found in standard output formats. This is due to the fact that most of these formats are limited in terms of their level of detail and precision in which results can be stored. Secondly, for ground truth reproduction as well as evaluation of OCR results it was necessary to have command line versions which could be integrated in automated workflows. The two OCR-engines considered for this kind of integration were ABBYY FineReader as the state-of-the-art commercial product and Tesseract representing the most advanced and popular Open Source project.

3.3.1 FineReader Integration

Parameterised command line tools were implemented which allowed running different versions of the ABBYY FineReader OCR-engine, returning recognition results in the latest PAGE XML format:

- Abbyy FineReader Engine 10
- Abbyy FineReader Engine 11

It has to be noted that a valid license key is required for using this tool, due to the commercial nature of the underlying API.

3.3.2 Tesseract Integration

Tesseract as the state-of-the-art Open Source OCR engine was made accessible as command line executables for the two latest versions:

- Tesseract 3.02
- Tesseract 3.03

In addition, not being subject to licensing fees, Tesseract has also been directly integrated in Aletheia (as already indicated in 3.2.1).

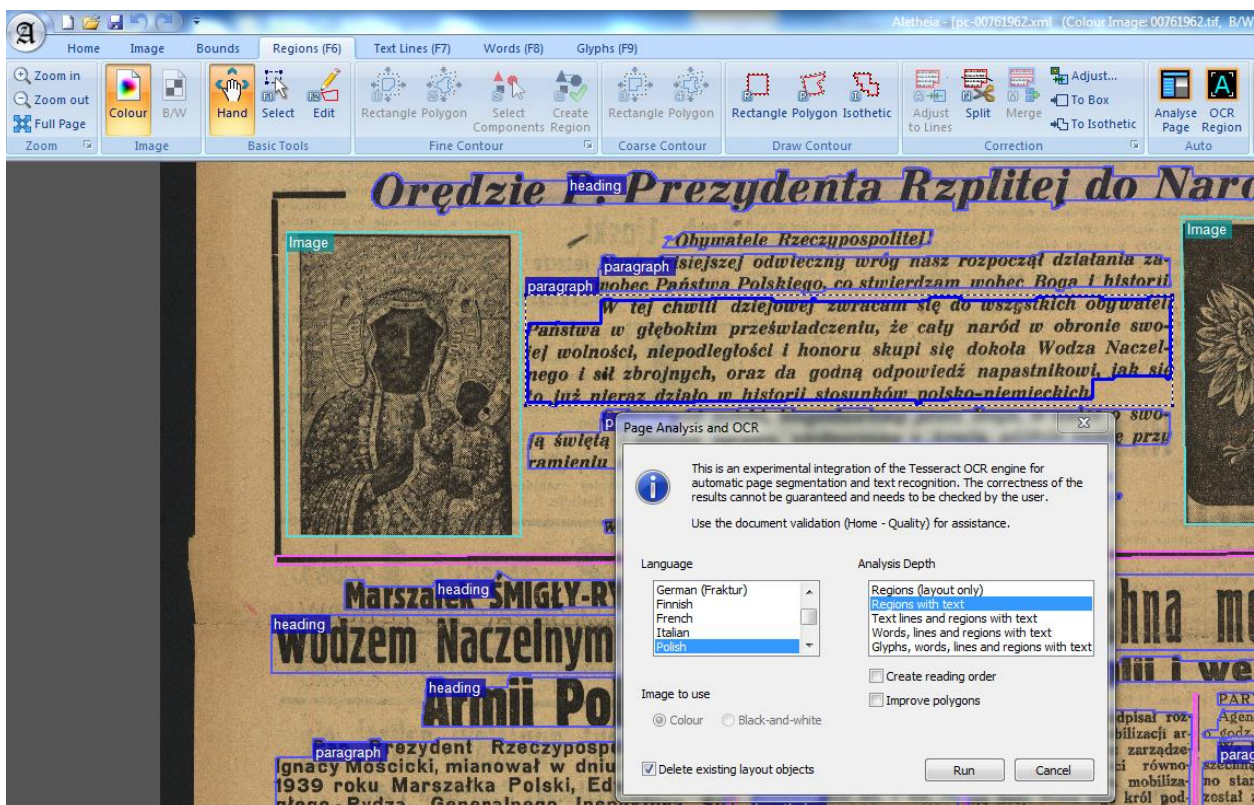


Figure 4: Integration of the Tesseract OCR-engine in Aletheia.

3.4 Evaluation Tools

Automated evaluation tools are essential not only for efficiently performing evaluation tasks but also to generate objective and reproducible results. The two main tools implemented and used for this purpose were a string comparison software analysing the text output of OCR-engines and a highly complex layout evaluation system covering image segmentation, region classification, and reading order.

3.4.1 OCR Evaluation

Due to the limitations of existing software, it was necessary to implement a completely new command line tool for evaluating the text accuracy of OCR engines. The implemented approach follows the ISRI OCR Evaluation Tools, originally developed at the University of Nevada, but adds a number of extra features:

- “Bag of words” performance measure (text accuracy disregarding the order of words)
- Word and character accuracy (with the option to ignore stop-words)
- Support of all latest file formats (PAGE XML, ALTO, ABBYY XML, hOCR)

3.4.2 Layout Evaluation

The layout evaluation tool used in Europeana Newspapers is an improved version of the software that was originally developed in the IMPACT project. Major extensions were:

- Introduction of new weights corresponding to the special needs of newspapers (for instance to define the importance of correctly recognising adverts)
- Performance improvements in order to cope with potentially very big ground truth and result files (compared to digitised books)
- Support of the latest ground truth and result file formats (PAGE XML, ALTO, ABBYY XML, hOCR)
- Support of the extended Layout Evaluation Data file format (see 3.1.2)

The layout evaluation tool comes in two versions – a command line executable for workflow integration and batch processing as well as a stand-alone GUI software for inspection of individual results and in-depth analysis of specific problems. The GUI version is also the main tool for defining evaluation profiles (sets of error weights) which are used to express specific use scenarios (compare D3.1 *Evaluation profiles for use scenarios*).

PRImA / Tools / Layout Evaluation /

Tools Home

Aletheia

Layout Evaluation

PAGE Libraries

Page Converter and Validator

Layout Evaluation Performance Analysis System

[Download the latest version](#)

Overview

This tool is part of a framework for evaluating the performance of layout analysis methods. It combines efficiency and accuracy by using a special interval based geometric representation of regions. A wide range of sophisticated evaluation measures provides the means for a deep insight into the analysed systems, which goes far beyond simple benchmarking. The support of user-defined profiles allows the tuning for practically any kind of evaluation scenario related to real world applications.

Features

- Evaluation Profile editor with Novel Metrics
 - Fully customisable metrics for different Region Types (Text, Image, Table, ...)
 - Fully customisable metrics for different Error Types (Merge, Split, Miss, Partial Miss, False Detection and Missclassification)
 - All possible combinations of Region Types and Error Types can be defined
 - Support for Reading Order (using both ordered and unordered groups)
- Evaluation result viewer, featuring:
 - Colour and pattern coded, interactive overlay on the document image
 - Highlights segmentation and classification errors

[Download the latest version](#)

[Download the previous version](#)

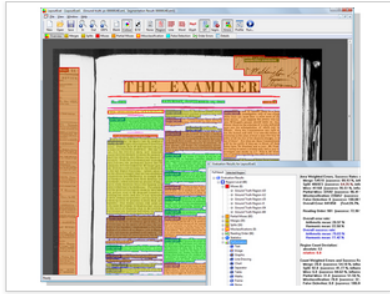


Figure 5: Download Page of the Layout Evaluation Tool

3.5 Workflow and Auxiliary Tools

In order to avoid (manual) gaps in the general evaluation infrastructure and to provide the greatest possible flexibility in terms of how the created resources can be used a number of smaller tools were implemented and further improved.

3.5.1 Extractor/Exporter Tool

The main purpose of this tool was to extract and feed raw text (without layout information) into the OCR Evaluation Tool. Stemming from a tool called “Image Extractor” (originally developed in IMPACT) it was significantly extended with a number of new features:

- Support of the latest PAGE XML format
- Gamera OCR export (for training, used for studying the potential of optimising OCR engines)
- Text extraction from text lines, words, or glyphs
- Improved command line interface

3.5.2 PAGE Metadata Scanner

Being a new development, this tool allows the automated extraction of rich metadata from a PAGE XML file. This can then be used for refined searches in the Image and Ground Truth Repository as well as for correlation analysis of certain document characteristics and evaluation results.

Major features of the tool are:

- Extraction / calculation of metadata and statistics
 - Metadata (ID, creator, creation time, modification time, width, height)
 - Border and print space presence (true/false)
 - Content objects count (per type and sub-type)
 - Text content statistics (number of characters and white spaces)
 - Language and script (semicolon separated list)
 - Reading order and layers (number of region references)
 - Character codes (e.g. to discover usage of long s in historical documents)
- CSV output

3.5.3 Feature Extractor

The Feature Extractor is again a completely new development and related to the ongoing Task T3.6 *Planning resources and quality estimation tools*. Similar to the PAGE Metadata Scanner, this tool aims at (programmatically) extracting specific characteristics from PAGE XML files but also from the corresponding document images as to be used for automated quality estimation. Its main features are:

- Prototypes for C++ and Java
 - Image based features, e.g.:
 - FOREGROUND_PIXEL_DENSITY
 - CONNECTED_COMPONENT_COUNT
 - IMAGE_NOISE
 - IMAGE_BRIGHTNESS
 - IMAGE_CONTRAST
 - EDGE_DETECTION
 - BRIGHTNESS_UNEVENNESS
 - Layout based features, e.g.:
 - TEXT_REGION_COUNT
 - REGION_OVERLAPS
 - FOREGROUND_OUTSIDE_REGIONS
 - REGIONS_WITHOUT_FOREGROUND
 - MISSING_REGION_TEXT
 - TEXT_LINE_COUNT_MISMATCHES
 - Text based features, e.g.:
 - WORDS_WITH_DIGITS
 - ALPHABETIC_CHARACTER_COUNT
 - WHITESPACE_COUNT
 - DIGIT_COUNT
 - PUNCTUATION_COUNT
 - AVERAGE_WORD_LENGTH
 - WORDS_IN_DICTIONARY

3.5.4 Image Characteristics Tagging Interface

In order to correlate evaluation results and specific types of artefacts in scanned images it was first necessary to manually tag all documents in the evaluation dataset. Not only did this involve a considerable number of instances (more than 600) but also a broad spectrum of keywords (86) organised in different categories (8). It was therefore decided to implement a tailor-made tagging interface which could be integrated into the Image and Ground Truth Repository as this would dramatically speed up the process of manually annotating hundreds of images and, moreover, would be a very useful new feature for other existing and/or future datasets.

While the software was a new development, the underlying list of keywords was based on resources from the IMPACT project which had to be consolidated, extended with additions specific to newspapers, and reorganised into the following eight categories:

- Ageing/Preservation (e.g. warped paper)
- Digitisation - Geometric Distortions/Properties (e.g. skew)
- Digitisation - Noise/Artefacts (e.g. paper clips visible)
- Document/Content (e.g. mixed languages)
- Layout/Formatting (e.g. mixed typefaces)
- Production Characteristics (e.g. textured paper)
- Production Faults (e.g. bleed-through)
- Use/Wear (e.g. folds)

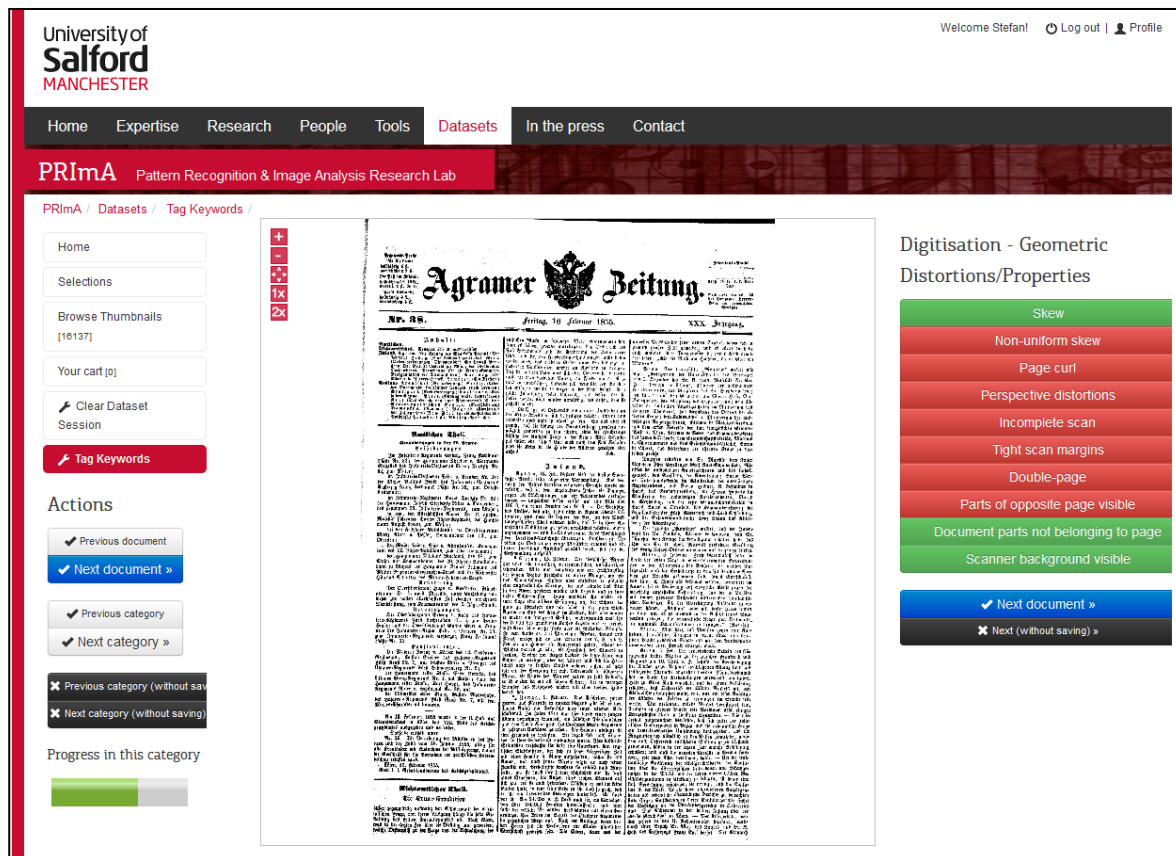


Figure 6: Tagging Interface integrated in the Image and Ground Truth Repository

3.5.5 PAGE Viewer

The PAGE Viewer was developed to provide a convenient and platform independent way (Aletheia, which can be used to view PAGE files, is only available for MS Windows operating systems) of displaying ground truth and processing results in PAGE XML format. It is based on Java and can be used on any system for which a compatible Java Runtime Environment is available.

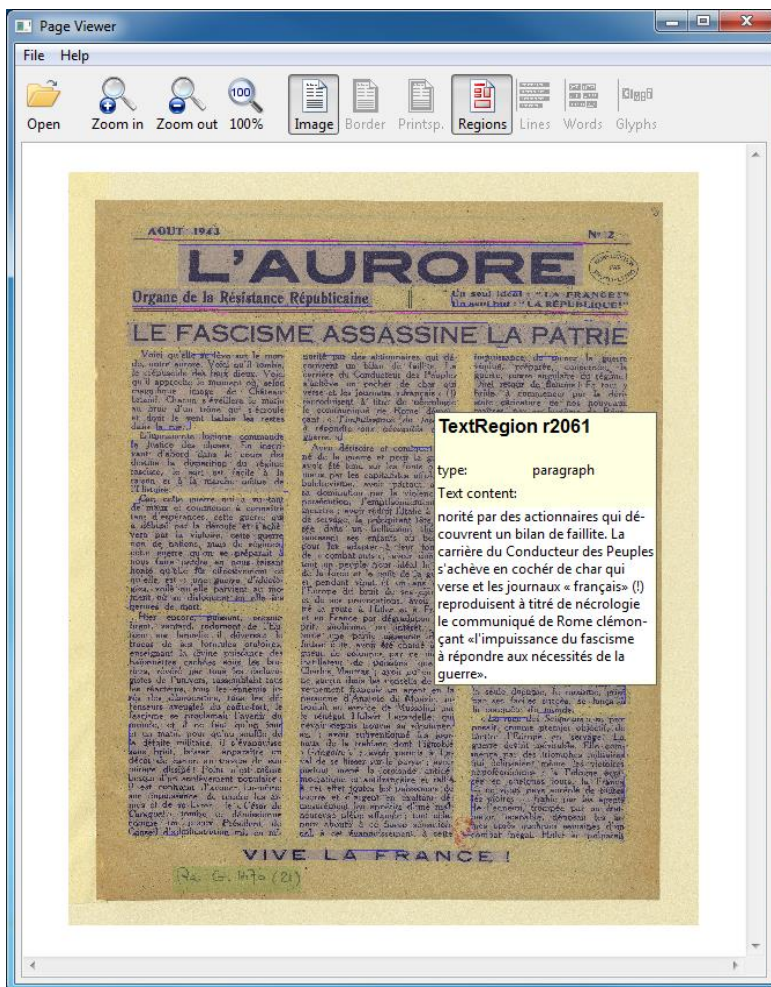


Figure 7: The cross-platform PAGE Viewer

4. Conclusion

The Task T3.3 *Evaluation Tools* has not only achieved to create the required infrastructure for all the evaluation activities as set out in the Description of Work but also produced a number of additional tools which will be very useful to future digitisation and research projects. The outcome, as documented in this report, is a comprehensive collection of software covering all aspects of performance evaluation related to OCR-workflows. The availability of tools for every step allows for a high degree of automation and therefore objective and reproducible evaluation results following best research practice.

The work on D3.3 *Evaluation Tools – Final Versions* was carried out as planned and there were no significant deviations from the Description of Work. The deliverable itself was on time, allowing dependent tasks to be started on schedule.

While Month 30 marks the deadline for the final deliverable, Task T3.3 continues until the end of the project (M36) in order to provide support, maintenance, and potentially needed minor additions to the collection of tools.

APPENDIX: Tool Documentations

Due to the size of the individual tool documentations these are attached as separate PDF files (in alphabetical order):

- [Aletheia - User Guide.pdf](#)
- [Extractor and Exporter - User Guide.pdf](#)
- [Feature Extractor - User Guide.pdf](#)
- [FineReader Integration - User Guide.pdf](#)
- [Layout Evaluation - User Guide.pdf](#)
- [OCR Evaluation - User Guide.pdf](#)
- [Page Content Ground Truth and Storage Format - Documentation \(Part of the PAGE Format Framework\).pdf](#)
- [PAGE Converter and Validator - User Guide.pdf](#)
- [PAGE Scanner - User Guide.pdf](#)
- [PAGE Viewer - User Guide.pdf](#)
- [Tesseract Integration - User Guide.pdf](#)