

**To:** Cristina Maier (Project Officer)

DG CONNECT G.2  
Creativity Unit  
EUFO 01/162 - European Commission  
L-2557 Luxembourg

**From:** Support Action Centre of Competence in Digitisation

Project acronym: Succeed Project Number: 600555

Project Manager: Rafael C. Carrasco Jiménez

Project Coordinator: Universidad de Alicante

**The following deliverable:**

Deliverable title: Report on the infrastructure and data sets provided for evaluation purposes and their usage

Deliverable number: D5.1

Deliverable date: M21

Partners responsible: USAL

Status:  Public  Restricted  Confidential

**is now complete**

- It is available for your inspection
- Relevant descriptive documents are attached

**The deliverable is:**

- A document
- A website      Url: http://www.primaresearch.org/
- Software
- An event
- Other \_\_\_\_\_

**Sent to Project Officer:** cristina.maier@ec.europa.eu      **Sent to functional mailbox:** CNECT-ICT-600555@ec.europa.eu      **On date:** 17/10/2014







# D5.1 Report on the infrastructure and datasets provided for evaluation purposes and their usage

---

SUCCEED

16/10/2014



## Document information

<b>Deliverable number</b>	D5.1	Start: M4	Due: M21	Actual:
<b>Deliverable name</b>	Report on the infrastructure and datasets provided for evaluation purposes and their usage			
<b>Internal/External</b>	External			
<b>Activity type</b>				
<b>Participant</b>	USAL			
<b>Estimated person months per participant for this deliverable</b>				
<b>Dissemination level<sup>1</sup></b>	Public			

## Document history

<b>Revisions</b>				
<b>Version</b>	<b>Status</b>	<b>Author</b>	<b>Date</b>	<b>Changes</b>
1.0	Draft	Christos Papadopoulos Stefan Pletschacher	18/09/2014	
1.1	Final	Christos Papadopoulos	03/10/2014	Amendments according to internal review comments
1.2	Final	Christos Papadopoulos	16/10/2014	Amendments according to technical project manager's comments
<b>Approvals</b>				
<b>Version</b>	<b>Date of approval</b>	<b>Name</b>	<b>Role in project</b>	<b>Signature</b>
1.0	22/09/2014	Apostolos Antonacopoulos	WP5 Leader	
1.1	03/10/2014	Jesse de Does	D5.1 internal reviewer	
<b>Distribution</b>				
This document was sent to:				
<b>Version</b>	<b>Date of sending</b>	<b>Name</b>	<b>Role in project</b>	
1.0	23/09/2014	Jesse de Does	D5.1 internal reviewer	
1.1	03/10/2014	Isabel Martinez	Technical Project Manager	
1.2	17/10/2014	Isabel Martinez	Technical Project Manager	
	17/10/2014	Cristina Maier	Project Officer	

<sup>1</sup> PU Public; RP Restricted to other programme participants (including Commission Services); RE Restricted to a group specified by the consortium (including Commission Services); CO Confidential, only for members of the consortium (including the Commission Services)

## Table of Contents

<b>1. EXECUTIVE SUMMARY .....</b>	<b>6</b>
<b>2. Evaluation Dataset .....</b>	<b>7</b>
2.1. Content .....	8
2.1.1. Images .....	8
2.1.2. Metadata.....	11
2.1.3. Ground truth .....	20
2.2. Access .....	24
2.2.1. Web Interface.....	24
2.2.2. Direct Access.....	28
2.2.2.1. Authentication .....	28
2.2.2.2. Authorisation .....	29
2.2.2.3. Checking if a document image exists .....	29
2.2.2.4. Accessing an Image.....	30
2.2.2.5. Accessing an Attachment.....	30
2.3. Exploitation .....	32
2.3.1. Uses of the Repository management system .....	32
2.3.2. Integration to other systems .....	32
2.3.3. Dataset .....	32
<b>3. Evaluation Infrastructure .....</b>	<b>33</b>
3.1. Purpose .....	34
3.2. Tools used.....	35
3.2.1. Layout Evaluation Tool .....	35
3.2.2. Text (OCR) Evaluation Tool .....	36
3.3. Platform description .....	37
3.3.1. Web front end .....	37
3.3.2. Evaluation server .....	42
3.3.2.1. Text Evaluation Web Service.....	42
3.3.2.2. Layout Evaluation Web Service .....	45
3.4. Exploitation .....	47
<b>4. Glossary .....</b>	<b>48</b>
<b>5. References.....</b>	<b>50</b>



## Tables

Table 1 — Number of images by content provider .....	8
Table 2 — Distribution of number of document images per century.....	9
Table 3 — Distribution of number of document images per document type .....	9
Table 4 — Distribution of number of document images per language .....	10
Table 5 — Distribution of number of document images per script.....	11
Table 6 — Listing of all metadata .....	13
Table 7 — Listing of all keywords .....	19
Table 8 — Ground truthed regions per type/subtype.....	21
Table 9 — Total ground truth regions .....	22
Table 10 — Evaluation methods supported .....	34
Table 11 — Bag of Words Parameters .....	43
Table 12 — Multiple Bag of Words Parameters .....	43
Table 13 — Text Evaluation Parameters .....	44
Table 14 — Multiple Text Evaluation Parameters .....	45
Table 15 — Layout Evaluation Parameters .....	45
Table 16 — Multiple Layout Evaluation Parameters .....	46

## Figures

Figure 1 — Distribution of number of document images per century.....	9
Figure 2 — Distribution of number of document images per document type .....	9
Figure 3 — Distribution of number of document images per language .....	10
Figure 4 — Distribution of number of released document images per script .....	11
Figure 5 — Keyword tagging interface .....	14
Figure 6 — Sample document images.....	20
Figure 7 — Sample ground truth page, showing region outlines.....	23
Figure 8 — Gallery view of selections of images (either via browsing or searching) .....	25
Figure 9 — Details view for a selected document image. ....	26
Figure 10 — Interactive ground truth explorer. ....	27
Figure 11 — Keywords section of details view, expanded.....	27
Figure 12 — Layout Evaluation Tool .....	35
Figure 13 — File upload and evaluation method selection interface .....	37
Figure 14 — Popup warning.....	38
Figure 15 — Details of files to be uploaded.....	38
Figure 16 — Files queued for evaluation .....	39
Figure 17 — Queued evaluation requests.....	40
Figure 18 — Successful evaluation requests.....	40



Figure 19 — Failed evaluation requests .....	41
Figure 20 — Filtered view of results .....	41
Figure 21 — Preview of evaluation record .....	41
Figure 22 — Comparison chart .....	42



## 1. EXECUTIVE SUMMARY

This report summarises the work and results related to the tasks T5.1 Evaluation Infrastructure and T5.2 Evaluation Datasets which constitute the two technical activities in work package 5 Evaluation, Awards, and Competitions.

The actual outcomes described in the following are two functionality-rich online systems which together form a comprehensive evaluation environment related to document digitisation:

- **The Evaluation Dataset** provides a common point of reference for material that is representative of the holdings of several major European libraries and relevant to their current and near future digitisation efforts. The dataset contains almost 600,000 document images and 45,000 layout and text ground truth files. The underlying system allows for convenient and efficient maintenance and curation of the current content and potential future additions.
- **The Evaluation Infrastructure** implements an interface for carrying out reproducible and in-depth evaluation of OCR results based on well-defined metrics and use-scenarios. It combines the Evaluation Dataset and the Evaluation Tools in a web platform that allows researchers to evaluate their algorithms and compare them with previous results.

The two systems can be considered a major step towards standardisation, automation, and reuse of resources related to evaluation of OCR methods and workflows. Not only will this allow future research and development activities to be more focused and cost-effective, it will also lead to objectively comparable results and a better understanding of the current state-of-the-art and any reported advances on it.

The two systems already proved their usefulness within the scope of other tasks of SUCCEED (most prominently the successful running of two international competitions) as well as in another EU funded project<sup>1</sup>.

Both the Evaluation Dataset and the Evaluation Infrastructure are open for members of the IMPACT Centre of Competence in Digitisation to register for and to use, via the PRImA website (<http://www.primaresearch.org>).

### Interaction with other Work Packages

In terms of technical interaction with other SUCCEED work packages, the outputs of this work package described in this report have been used by work package 2 (Interoperability and infrastructure). Also, work package 6 (Dissemination and community building) contributed towards the dissemination of the repository and evaluation platform.

---

<sup>1</sup> Europeana Newspapers Project [6]



## 2. EVALUATION DATASET

The availability of high-quality datasets is a crucial prerequisite for any research and development activities which aim at improving digitisation and recognition processes. For instance training of OCR classifiers, creation of lexica to aid recognition and post-correction, and general evaluation of new algorithmic approaches, depend all strongly on representative images and ground truth (representing the true content of a page as it would be recognised by the perfect method). Due to the very significant effort of selecting and collecting representative datasets as well as the high costs of creating comprehensive ground truth (covering textual content as well as layout aspects like region outlines) it is highly desirable to make any such resources available as widely and as conveniently as possible. The SUCCEED Evaluation Dataset achieves this for the material that was collected within the EU-funded project IMPACT, facilitating curation, maintenance and use of the hosted resources. As such the Evaluation Dataset provides an invaluable resource for future research by grouping together a variety of documents from major European libraries that are representative of both the challenges and targets in their digitisation efforts.

During the SUCCEED project, the original IMPACT repository has seen a major redesign; taking it from a research dataset to a complete dataset management infrastructure. It allows optimal exploitation of the resources available, both via the web and by integrating and directly accessing images and ground truth files via other applications.

The supporting database infrastructure has been optimised in order to achieve better indexing and more efficient searching and browsing of the images. Also a lot more detail has been added to logging in order to be able to keep track of what resources are accessed by whom. This was greatly improved in comparison to the original dataset (which was a closed system used only by project members, therefore detailed logging was never required).

In terms of access management, a very comprehensive system has been put in place, to grant/restrict access to certain resources for users that are logged in to the system. In order to make this as extendable as possible, a number of different authentication methods (including LDAP support) have been enabled.

A very easy to use interface has been integrated in the repository management system, enabling system administrators to easily tag any image with a set of over 80 keywords, and thus allowing the enrichment of the current dataset at any time.

In the following, the SUCCEED dataset is described in terms of its content, access to it, and real life applications and events where it has been used.



## 2.1. Content

The dataset contains over a quarter of a million representative text-based images compiled by a number of major European libraries. It covers texts from as early as 1500, and contains material from newspapers, books, pamphlets and typewritten notes.

### 2.1.1. Images

The dataset contains a wide variety of document images. The images that are available in this dataset were provided by major European libraries during the IMPACT project and were selected to be representative of both the respective library's holding and their digitisation plans for the near future [1].

A list of the libraries that have contributed to this dataset along with the number of images they provided is presented in Table 1. A number of images and corresponding ground truth files from specific content providers, although added on the repository, are currently not available online, due to administrative processes in legal and copyright matters. These images have been added to the repository but marked as pending, therefore they are reported separately throughout this report.

Library	Country	Released	Pending
Bayerische Staatsbibliothek	Germany	—	67,235
Biblioteca Nacional de España	Spain	60,180	—
Bibliothèque Nationale de France	France	—	96,950
British Library	UK	—	51,451
Koninklijke Bibliotheek	Netherlands	62,935	—
Narodna in Univerzitetna Knjižnica	Slovenia	42,379	—
Národní knihovna České republiky	Czech Republic	75,559	—
Österreichische Nationalbibliothek	Austria	—	110,034
Poznań Supercomputing and Networking Centre	Poland	11,020	—
St. Cyril and Methodius National Library	Bulgaria	4,240	—
<b>Total images</b>		<b>256,313</b>	<b>325,670</b>

Table 1 — Number of images by content provider

In terms of age, the majority of the documents were produced in the 19<sup>th</sup> century, although the collection goes as back as the 16<sup>th</sup> century. A more detailed breakdown of document images per century is available in Table 2 and Figure 1 below.

Century	Released	Pending
15th	—	505
16th	5,761	11,747
17th	47,944	13,895
18th	31,014	12,993
19th	124,143	181,405
20th	46,951	102,067
No info	500	3,058
<b>Total images</b>	<b>256,313</b>	<b>325,670</b>

Table 2 — Distribution of number of document images per century

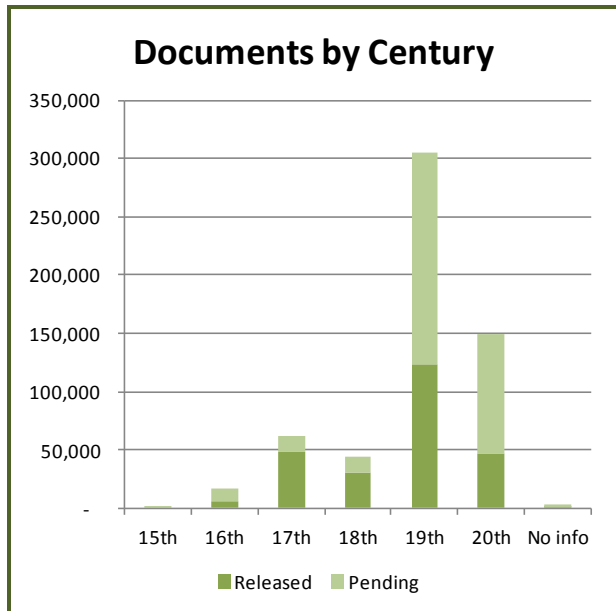


Figure 1 — Distribution of number of document images per century

A breakdown of the document types is provided in Table 3 and Figure 2.

Type	Released	Pending
Book	184,318	155,632
Journal	2,530	17,043
Legal	46,444	33,855
Newspaper	21,378	96,141
Other	1,643	17,576
Unknown	0	5,423
<b>Total images</b>	<b>256,313</b>	<b>325,670</b>

Table 3 — Distribution of number of document images per document type

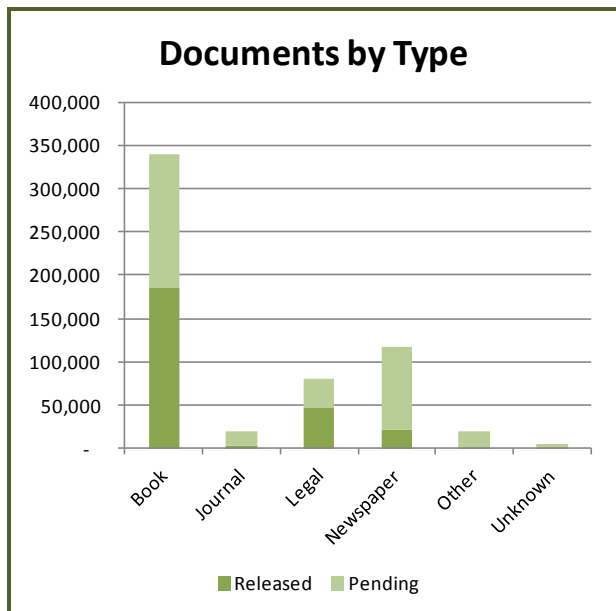


Figure 2 — Distribution of number of document images per document type

A number of different languages and scripts are represented in the dataset. Although the language and script information is not known for a large proportion of the images, a total of 17 languages and 10 scripts are confirmed within the dataset.

A listing of all languages is provided in Table 4 and Figure 3.

Language	Released	Pending
Bulgarian	4,240	—
Catalan	614	—
Czech	75,559	—
Dutch	3,510	476
English	—	11,763
French	—	804
German	—	64,456
Greek	500	—
Hebrew	438	—
Latin	163	4,916
Norwegian	—	446
Old Church Slavonic	500	—
Polish	9,582	—
Portuguese	1,494	—
Russian	—	620
Slovenian	42,379	—
Spanish	56,650	476
No data	60,684	241,713
<b>Total images</b>	<b>256,313</b>	<b>325,670</b>

Table 4 — Distribution of number of document images per language

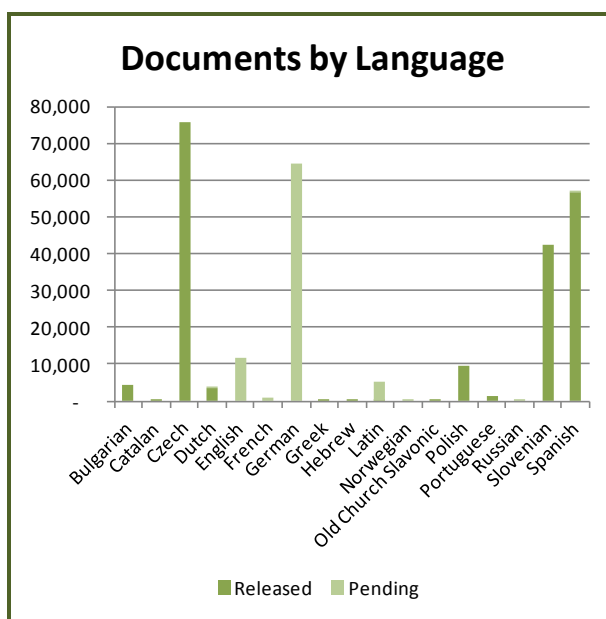


Figure 3 — Distribution of number of document images per language

Script	Released	Pending
Bohoričica	4,018	—
Cyrillic	4,616	620
French	614	—
Gaj	2,012	—
Greek	500	—
Hebrew	438	—
Latin	130,516	10,858
Latin/Gothic	304	—
Old Cyrillic	124	—
Serif	—	417
Not specified	113,171	313,775
<b>Total images</b>	<b>256,313</b>	<b>325,670</b>

Table 5 — Distribution of number of document images per script

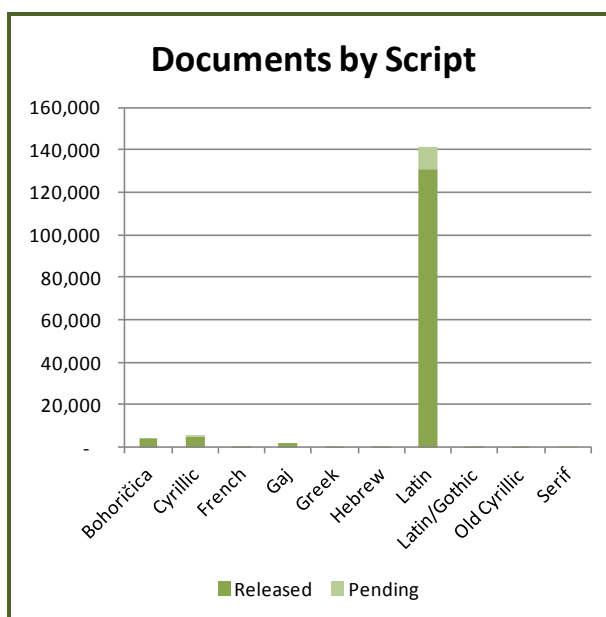


Figure 4 — Distribution of number of released document images per script

Table 5 and Figure 4 illustrate all scripts represented in the repository.

### 2.1.2. Metadata

A wide variety of metadata has been collected and has been made available as part of this dataset. All metadata available in this resource were collected from the contents providers and verified during the IMPACT project.

The metadata provided as part of the dataset are grouped in 6 categories. Most of the available metadata are indexed and searchable via the web interface used to browse the repository.

Table 6 provides a listing of all available metadata grouped in their categories.

Title	Description / Possible values
<b>Bibliographic Information</b>	
Title	Publication title
Author	Publication author
Publication Place	Publication location (city, country etc)
Publication Date (Year)	Publication year

Title	Description / Possible values
Publication Date (Month)	Publication month (if known)
Publication Date (Day)	Publication day (if known)
Date Description	A description of the publication period (if date is not known)
Page Number	Page number (if known/ applicable)
Document Type	Book, Newspaper, etc (see Table 3)
<b>Digitisation Information</b>	
Resolution	In dpi
Colour Depth	In bpp
File Width	In pixels
File Height	In pixels
Scanner Type	Book, Flatbed, Overhead, Camera
Scanner Model	Model of the scanner used
File Type	TIFF, PNG, JPEG, GIF
Compression	Lossy, Lossless, Unknown
Compression Type	
Compression Quality	
Original Source	Paper, microfilm, etc.
<b>Physical Characteristics</b>	
Language	See Table 4
Secondary Language	
Script	See Table 5
Secondary Script	
Default typeface	
Display typeface	
No Columns	Number of columns (1, 2, 3, ...)
<b>Copyright Information</b>	
Institution	See Table 1
Contact Person (Name)	Details of contact person in owner institution.
Contact Person (Email)	

Title	Description / Possible values
Contact Person (Phone)	
Can be published	Whether this image is free to use in scientific research/publications or requires extra licensing from the copyright holder (content provider).
Online Repository	URL of original repository where this image is hosted (if available and provided)
<b>Administrative Information</b>	
ID	The unique ID of the document image (allocated when adding the image in the repository)
Original Filename	The filename of the original image as provided by the library.
Owner Reference	A reference field used by the content provider to identify the image in their collection.
Access Log	Log entries for any access to this image.
<b>Comments</b>	
Physical Appearance	Comment on physical appearance
Paper Quality	Comment on paper quality
Binding	Comment on binding
Physical Layout	Comment on physical layout

Table 6 — Listing of all metadata

As part of the metadata describing an image, a comprehensive set of keywords has also been defined and can be used to describe each document image. For this project, a tagging interface (that was originally developed for the Europeana Newspapers project [6]) was integrated to the repository management system to allow marking images with specific keywords. Although most of the images are not tagged with any of these keywords, the functionality to tag images with an easy to use web based interface (see Figure 5) along with support for searching for images with or without a specific keyword is available. Each image can be marked to either have or not have the characteristic described by each one of the keywords.

Access to this interface is limited to system administrators, for security and consistency reasons.

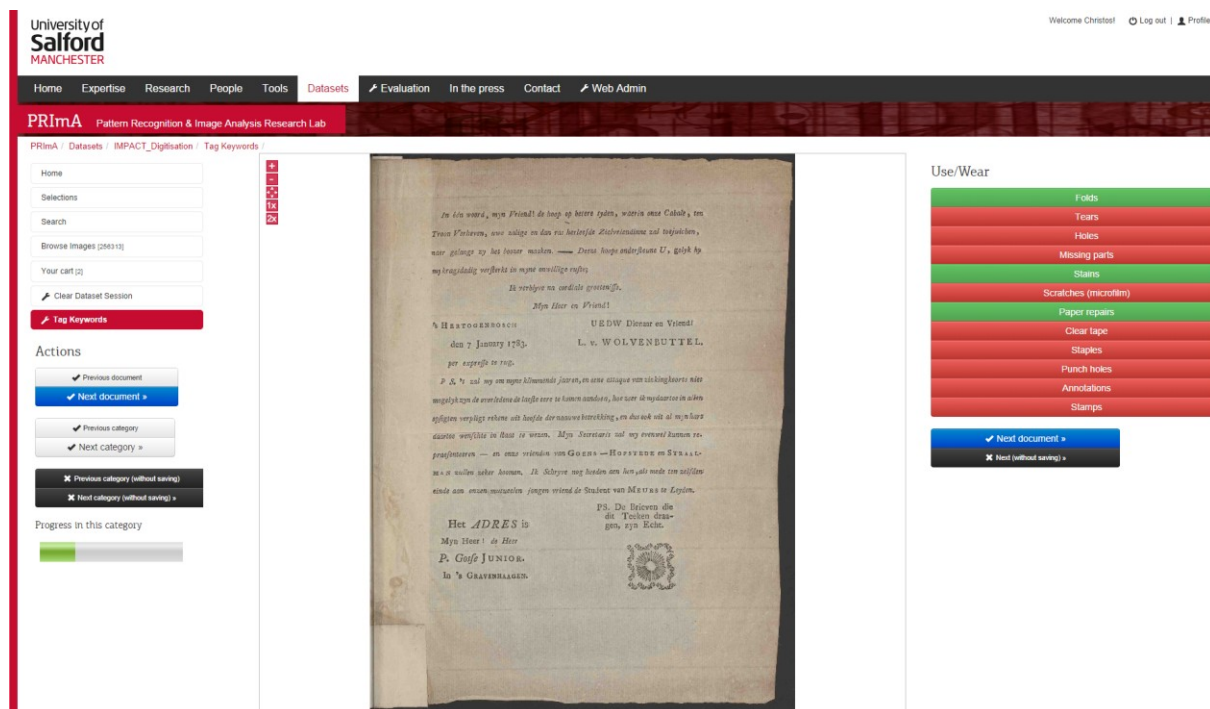


Figure 5 — Keyword tagging interface

Table 7 below, provides a full listing of all the available keywords grouped in 7 categories.

Keyword	Description
<b>Document/Content</b>	
Mixed languages	More than one language used in document – in separate paragraphs but also within the same paragraph (e.g. longer Latin citation in English text).
Illustrations	Figures, drawings, etc.
Photographs	Photographs of any kind.
Tables	Any kind of a table (with or without borders/separators).
Advertisements	Intended to persuade readers to purchase or take some action upon products, ideals, or services.
Charts	Charts or graphs of any kind.
Formulas	Mathematical formulas / equations.
Footnotes	Text placed at the bottom of the page to provide an author's comments on the main text or citations of a reference work in support of the text.



Keyword	Description
Marginalia	Notes and comments in the margins of a book by the author / printer.
Running titles	If the title of a chapter or section of a text is repeated above each (double) page, it is called a 'running title'.
<b>Layout/Formatting</b>	
Pasted clippings	Elements from another page pasted onto the page (substrate and clipping colour might not match).
Mixed typefaces	More than one typeface/font used in document image (e.g. Arial and Times New Roman). See also the 'Black letter' label.
Mixed font sizes	Page contains text of different sizes.
Black letter	Black letter script used (also known as Gothic script; e.g. Fraktur).
Typewritten	Some or all of the text has been produced using a typewriter.
Handwritten	Some or all of the text has been written by hand (cursive or print script). NOT to be used for manuscript, which is handwritten but looks like printed).
Medieval manuscript	Print-like or standardised handwritten text.
Drop caps	Letter(s) at the beginning of a work, a chapter or a paragraph that is larger than the rest of the text.
Decorative drop caps	Drop caps that differ from the remaining text (e.g. with flourishes or drawings).
Decorative borders	Decorative frames or drawings around the content of a page (inherent to the document image; not caused by the scanning process).
Frames	Plain or decorative frames around part of the content.
Multi-column layout	More than one text column per page. Typical for newspapers.
Rotated text	All or part of the text is rotated by 45 degrees or more.
Multiple colours in illustrations	Multiple foreground colours used in illustrations
Multiple colours in text	Multiple foreground colours used in text

Keyword	Description
Reverse video	There is some text with a foreground colour that is brighter than the background colour (e.g. white text on black).
<b>Production Characteristics</b>	
Textured paper	Textured paper was often used in 16th century books and might leads to Salt and Pepper Noise when the image was binarised.
Uneven character spacing	Non-uniform spacing between the glyphs/characters within words.
Multiple colours in annotations	Multiple foreground colours used in annotations
Multiple colours in stamps	Multiple foreground colours used in stamps
Narrow border	Little or no margin around the page content (not to be used when due to scanning artefacts; use 'Tight scan margins' instead).
Low paper to text contrast	Physically caused effect of hardly recognisable text (e.g. by using carbon copies) and would be avoidable when considering it during the printing process (not to be used for scanning related low contrast or fading ink).
Impressions	Impressions / embossing visible through illumination.
Watermarks	Faint background images.
Halftoning	Matrix of dots with different sizes, often used for illustrations/photos in newspapers.
<b>Production Faults</b>	
Uneven ink distribution	Uneven ink distribution within a glyph / character.
Bleed-through	When paper is too thin or the ink applied too heavily the color can bleed or seep through to the other side (not to be confused with show-through which is a scanning artefact).
Ink from facing page	If the ink wasn't dry when two pages were put together, ink from one page can be transferred to the opposite page. Similar to, but not identical with bleed-through.
Broken characters	Separated parts of a glyph that are usually connected. A too low threshold applied in a binarisation process might cause this effect.

Keyword	Description
Faint characters	Faint individual characters (e.g. due to insufficient pressure when typewriting). Not to be used for 'fading ink' which is an ageing/preservation artefact).
Blurred characters	Blurring that has been introduced during the printing/writing process (not to be confused with 'out-of-focus' in scanning).
Smearred Ink	Smearred ink due to the production process.
Filled-in characters	Holes or gaps of glyphs filled in with ink.
Sort shoulder artefacts	Lines or other artefacts around a glyph that are caused by the sort shoulder (the metal body used in type setting) touching the paper.
Horizontaly touching characters	Glyphs/characters within one text line touch a neighbour glyph. Not to be used for cursive handwritten text where this is to be expected.
Vertically touching characters	Glyphs/characters touch glyphs of another text line.
Non-straight text lines	Due to imprecise printing text lines may not appear straight, but corrugated or bent (not to be used for warped paper which is an ageing/preservation artefact).
Use/Wear	
Folds	Visible artefacts on a page that has been folded (e.g. a newspaper that has been folded to save space).
Tears	The paper has been damaged at the sides or ripped open in the centre.
Holes	Unwanted holes in the paper (not punch holes). This not only means the loss of information directly stemming from the hole, but if no underlay paper was used while scanning, the following page will be seen through the hole, thus presenting unwanted information.
Missing parts	Parts of the substrate are missing
Stains	Stains on the paper.
Scratches (microfilm)	
Paper repairs	Repairs of wholes, tears, or weak page matter (there is specialised label for 'clear tape' repairs).
Clear tape	Tears in the paper that are 'repaired' with clear tape.
Staples	Staples visible

Keyword	Description
Punch holes	Punch holes for filing (not to be confused with unwanted 'holes').
Annotations	Any type of annotation that has been added after the original production.
Stamps	Stamps of any kind
<b>Ageing/Preservation</b>	
Warped paper	Arbitrary warping of the paper usually caused by moisture (not to be confused with page curl which is caused by the document binding).
General paper discolouration	
Discoloured paper edges	
Mould	
Non-straight paper edges	The page is not rectangular.
Fading ink	General ink fading on the page (not only individual characters).
<b>Digitisation - Geometric Distortions/Properties</b>	
Skew	During the scanning process pages are not always adjusted according to the print space and therefore images are skewed.
Non-uniform skew	Non-uniform skew (e.g. due to scanner feeder).
Page curl	Smooth curl caused by the document binding.
Perspective distortions	Perspective distortions introduced by scanners (e.g. camera-based overhead scanning).
Incomplete scan	If the scan frame is set improperly, or in extreme cases of narrow bindings, parts of the text may be cut off.
Tight scan margins	Little or no border around the page content due to a tight scanning window.
Double-page	A book is digitised in such a way that one image shows two facing pages, not just one.
Parts of opposite page visible	
Document parts not belonging to page	Cover, binding gutter, or other visible parts of the document not belonging to the page.

Keyword	Description
Scanner background visible	The border of the image shows the background used during scanning (black or white for instance).
<b>Digitisation - Noise/Artefacts</b>	
Show-through	If a paper is relatively thin the ink of the back side may shine through. Very similar to, but not identical with bleed-through (production artefact).
Uneven illumination	Uneven page illumination during scanning.
Out-of-focus	Blurred image due to not properly focused camera.
Noise from scanner	Fluctuations in light intensity in an image that are not to be found in the source document.
Low scan contrast	Low image contrast caused by scanner.
Paper clips visible	Clips to fixate the page during scanning are visible.
Fingers visible	Fingers to fixate the page during scanning are visible.
Salt-and-pepper noise	
Missing information after binarisation	
Noise and remnants after binarisation	
Dithering	Pixel patterns to emulate colours / grey levels. Can occur in binarised or low-colour images.

Table 7 — Listing of all keywords

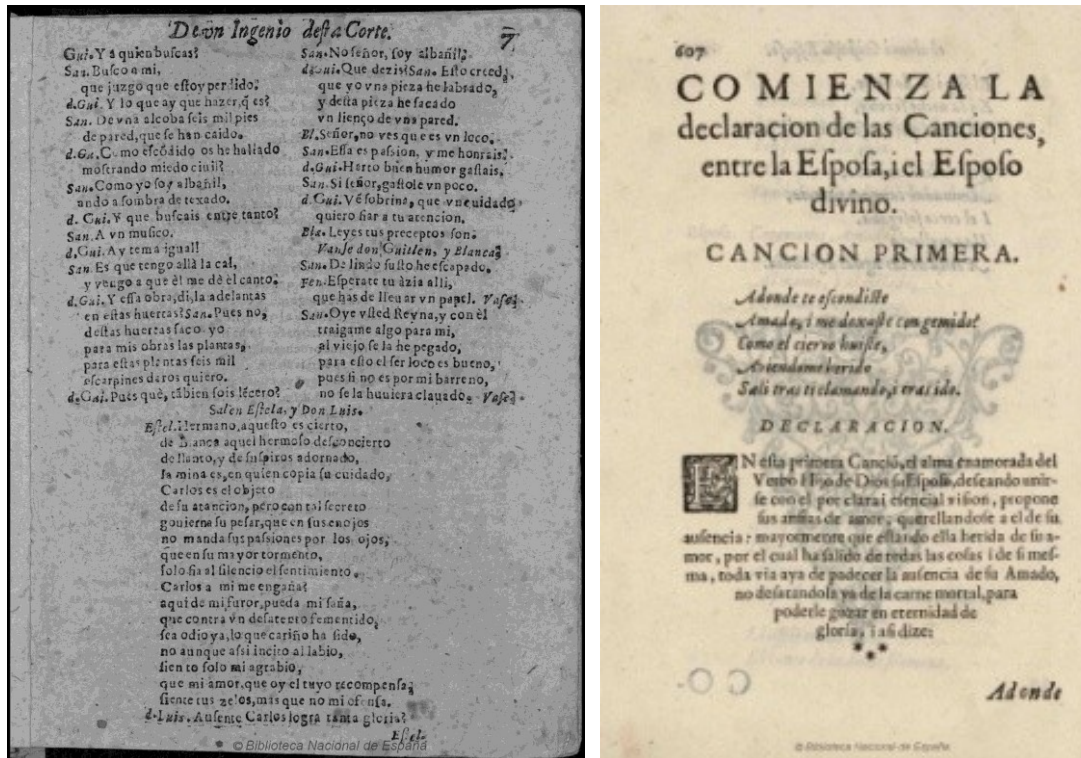


Figure 6 — Sample document images

The ability to tag and search for specific image characteristics makes the Evaluation Dataset a particularly powerful tool for organising targeted experiments related to research and development activities as well as carrying out digitisation pilots in order to estimate the achievable performance for certain types of documents with a given OCR workflow.

### 2.1.3. Ground truth

Layout and text ground truth for a carefully selected subset of these images is also available. This ground truth data allows evaluation of results of any method on either layout analysis or OCR.

A total of 44,458 images are accompanied by ground truth down to region outlines, region text content and reading order. However, 29,762 are released whereas 14,696 are still pending (since they correspond to document images that are still pending). An overview of the different types of regions included in the ground truth data is presented in Table 8.

Region Type	Released	Pending	Total
<b>Text</b>	<b>389,321</b>	<b>150,109</b>	<b>539,488</b>
Heading	23,565	92,705	361,291
Paragraph	268,586	12,715	36,280
Drop capital	3,334	2,169	5,503
Caption	169	125	294
Header	22,995	11,974	34,969
Footer	390	19	409
Footnote	2,091	785	2,876
Footnote continued	156	31	187
Signature mark	7,371	3,261	10,632
Catch word	18,038	2,626	20,664
TOC-entry	2,943	3,274	6,217
Page number	25,146	12,561	37,707
Marginalia	3,434	7,657	11,091
Credit	11,103	204	11,307
<b>Graphic</b>	<b>6,418</b>	<b>2,139</b>	<b>8,673</b>
Logo	1	3	4
Stamp	360	338	698
Handwritten Annotation	1,669	415	2,084
Punch Hole	419	—	419
Signature	13	2	15
Other	3,956	1,199	5,155
<b>Image</b>	<b>387</b>	<b>855</b>	<b>1,242</b>
<b>Line Drawing</b>	<b>5</b>	<b>3</b>	<b>8</b>
<b>Separator</b>	<b>14,982</b>	<b>11,232</b>	<b>26,214</b>
<b>Table</b>	<b>960</b>	<b>454</b>	<b>1,414</b>
<b>Chart</b>	<b>2</b>	<b>2</b>	<b>4</b>
<b>Maths</b>	<b>108</b>	<b>247</b>	<b>355</b>

Table 8 — Ground truthed regions per type/subtype



A small number of the images that have been groundtruthed, contains outlines for text lines and words as well as regions. This results in a total of 10,118 text line and 69,209 word outlines being also available. For more details on all available ground truth regions, see Table 9.

Outlines	Released	Pending	Total
Region (all sub types)	412,392	165,095	<b>577,487</b>
Text Line	3,459	6,659	<b>10,118</b>
Word	24,558	44,651	<b>69,209</b>

Table 9 — Total ground truth regions

The ground truth is stored in the XML format which is part of the PAGE (Page Analysis and Ground truth Elements) representation framework [1]. For each region on the page there is a description of its outline in the form of a closely fitting polygon. A range of metadata is recorded for each different type of region. For example, text regions hold information about language, font, reading direction, text colour, background colour, logical label (e.g. heading, paragraph, caption, footer, etc.) among others. Moreover, the format offers sophisticated means for expressing reading order and more complex relations between regions.

Especially noteworthy is the fact that all text is encoded in Unicode, providing a faithful representation of what can actually be found on the page (diplomatic transcription). This level of detail (for instance knowing whether a character was printed as a ligature or the presence of the long s rather than the modern s) allows for studying not only the abstract content but also the information conveyed in historical spelling and script variations. Current OCR systems are typically not capable of recognising the wealth of historical special characters, making the Evaluation Dataset a most valuable resource for development and training towards more precise recognition of historical documents.

Some example images with ground truth description can be seen in Figure 7. Region outlines are highlighted in blue for textregions, magenta for separators, cyan for images and green for graphics.



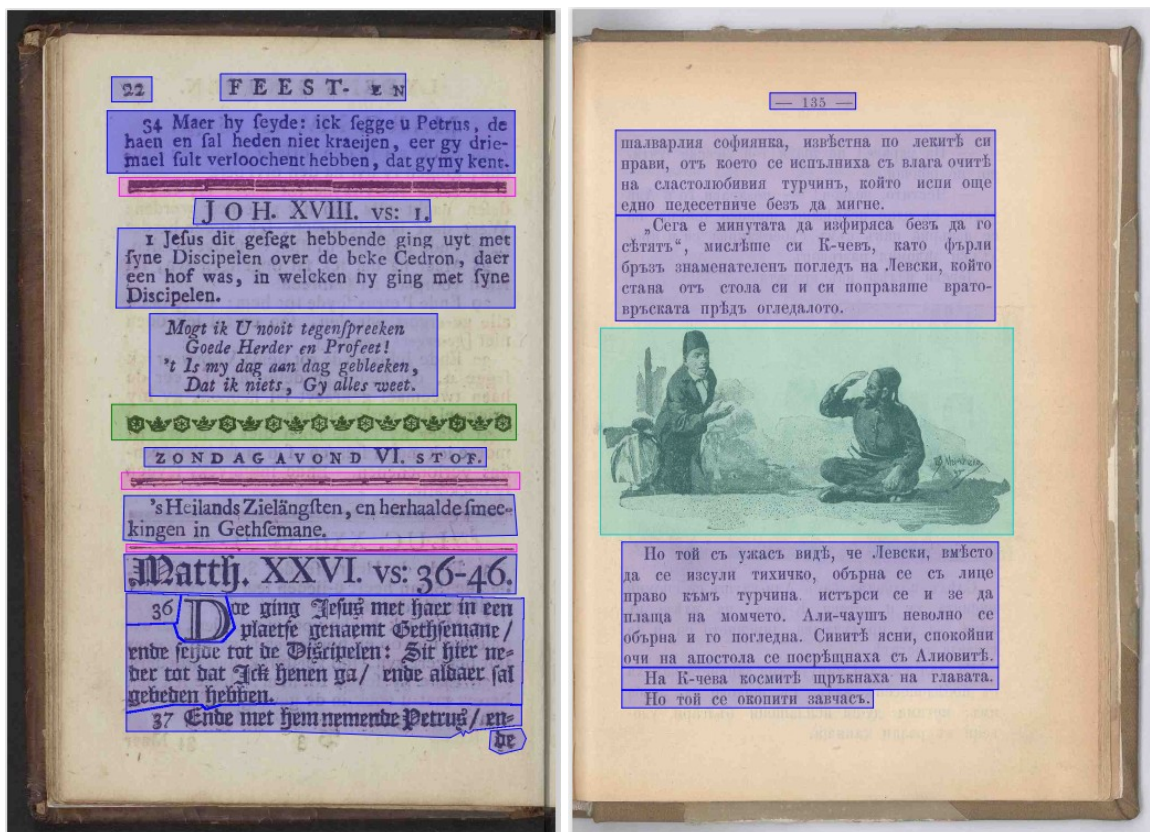


Figure 7 — Sample ground truth page, showing region outlines

## 2.2. Access

Access to the dataset is possible either via a custom built web interface, or via direct API calls for accessing images and ground truth data.

### 2.2.1. Web Interface

To allow access to the repository over the web, a custom built web interface was developed. This was based on the web interface developed and used for the IMPACT Project's [8] exploratory dataset, but was redesigned to make it more extendable and aid rights management and maintenance.

The web interface allows users to access images and/or ground truth files in a very efficient manner. The main options for accessing the document images include:

- **Browse all images:** Allows the user to browse through the whole collection of images.
- **Browse specific subsets:** Allows the user to browse a specific subset of images. These subsets are predefined and aim in assisting the user more quickly browse through smaller sets of images.
- **Search the complete set:** Allows the user to define a number of search criteria (see Table 6 — Listing of all metadata) and keywords which are then applied to the complete dataset.
- **Filter results:** Allows the user to further narrow down the results of any search/browse operation by filtering using any of the fields that are available in the search page.

Figure 8 shows a gallery view from the web interface, showing a set of images (this view is the same when browsing or searching).

Some features worth noting are:

- **Smart page navigator:** Adjusts automatically what range to display so that navigation through a large set of page is as easy as possible,
- **Actions:** Shortcuts for commonly used actions, such as adding/removing all visible images from your cart,
- **Filters:** Dynamically applies a filter based on a list of criteria to help find the image(s) required.

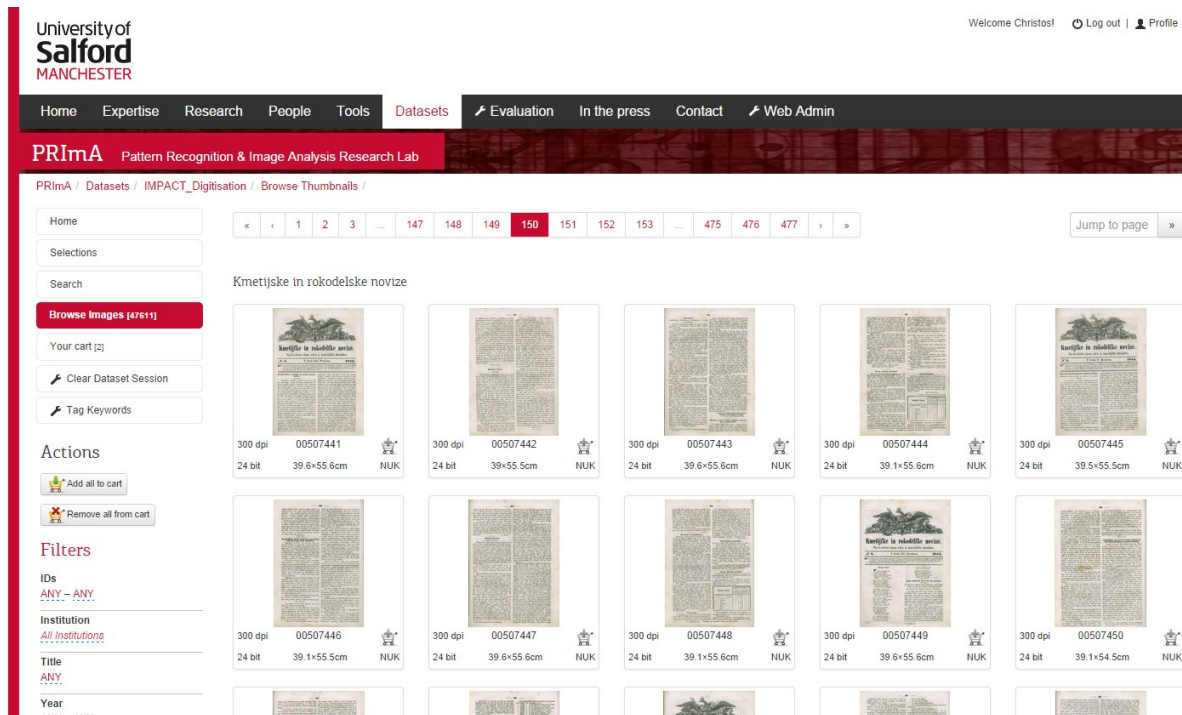


Figure 8 — Gallery view of selections of images (either via browsing or searching)

Once the user selects an image from the thumbnails displayed, they are redirected to a page containing all details about the specific image. These include:

- **Metadata:** All the metadata associated with the specific image (see Table 6 for a complete list).
- **Keywords:** All keywords associated with the specific image.
- **Attachments:** Any attachments linked to the image (this included grounds truth files).
- **Related Images:** Links to other images in the repository that are related to the current image (this could include alternative versions/scans of the same page, left/right pages of double page scans, etc).

For PAGE ground truth attachments, the user is offered the option to view an interactive graphical rendering of the content (see Figure 10).

University of  
**Salford**  
MANCHESTER

Welcome Christos! [Log out](#) | [Profile](#)

Home Expertise Research People Tools **Datasets** Evaluation In the press Contact Web Admin

**PRImA** Pattern Recognition & Image Analysis Research Lab

PRImA / Datasets / IMPACT\_Digitisation / Image Details

Home

Selections

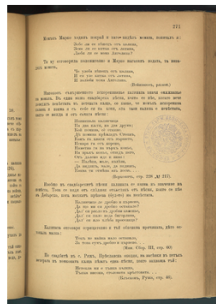
Search

Browse Images [8283]

Your cart [2]

Clear Dataset Session

Tag Keywords



Attachments [1]

PAGE Groundtruth [1]

Attachment ID	Description	Download	Date Added
199756	PAGE Page Content Groundtruth	<a href="#">pc-00541608.xml [9.18 KB]</a>	

### Generic Information

ID	00541608
Institution	NLB: St. Cyril and Methodius National Library (Bulgaria)
Can be published	0

### Document Description

Title	Spisanie Dennica
Publication Date (Year)	1891
Original Source	Paper

### Digitisation

Colour Depth	24 bit
Resolution	300 dpi
Size	18x25.7cm
File Type	UNKNOWN
Compression	UNKNOWN
Compression Type	UNKNOWN

### Physical Appearance

### Physical Layout

Language	Bulgarian
Script	Cyrillic

### Document Types

Types	Journal
-------	---------

### Keywords


Keywords	No keywords available for this document.
----------	--

Number of visitors since : 224291 — Maintained by: Christos Pappadopoulos — © PRImA Research 2004-14

Figure 9 — Details view for a selected document image.



Content: Regions Text Lines Words Glyphs ?



TextRegion	
ID	r16
type	paragraph
production	
primaryLanguage	
secondaryLanguage	
primaryScript	
secondaryScript	
readingOrientation	
readingDirection	
orientation	
leading	
indented	
align	
custom	
comments	

UY SEÑOR MIO. DE LAS BACHILLERIAS de vna conversacion, que en la merced que Vmd. me haze pafaron plaça de viveças naçe en Vmd. el defeo de ver por escrito algunos discursos que alli haze de repente sobre los Sermones de vn excelente Orador, alabando algunas vezes sus fundamentos, otras difintiendo, y siempre admirandome de su fin igual ingenio; que aun sobre sale mas en lo segundo, que en lo primero, por que sobre solidas vafas no es tanto de admirar la hermosura de vna fabrica, como la de la que sobre flacos fundamentos se ostenta luzida, quales son algunas de las propociones de este subtilissimo talento, que es tal su suavidad, su viveça, y energia, que al mismo que disfiende, enamora con la belleza de la Oracion, suspende con la dulçura, y hechiza con la gracia, y cleua, admira, y encanta con el todo, de esto hablamos, y Vmd. gustó ( como ya dixè) ver esto escrito, y por que conozca que le obedezco en lo mas discil, nõ fole de parte de el entendimiento en afumpto tan arduo como notar propociones de tan gran suieto, sino de parte de mi genio repugnanate à todo lo que parece impugnar à nadie, lo hago: aunque modificado este inconveniente, en que assi de lo vno como de lo otro ferà Vmd. fole el restigo, en quien la propria autoridad de su precepto he nestarà los errores de mi obediencia, que à otros ojos pareciera desproporcionada sober via, y mas cayendo

© Biblioteca Nacional de España

Figure 10 — Interactive ground truth explorer.

### Keywords

	<input checked="" type="checkbox"/> Hide True <input checked="" type="checkbox"/> Hide False
Document/Content	<input checked="" type="checkbox"/> Mixed languages <input checked="" type="checkbox"/> Illustrations <input checked="" type="checkbox"/> Photographs <input checked="" type="checkbox"/> Tables <input checked="" type="checkbox"/> Advertisements <input checked="" type="checkbox"/> Charts <input checked="" type="checkbox"/> Formulas <input checked="" type="checkbox"/> Footnotes <input checked="" type="checkbox"/> Marginalia <input checked="" type="checkbox"/> Running titles <input checked="" type="checkbox"/> Pasted clippings
Layout/Formatting	<input checked="" type="checkbox"/> Mixed typefaces <input checked="" type="checkbox"/> Mixed font sizes <input checked="" type="checkbox"/> Black letter <input checked="" type="checkbox"/> Typewritten <input checked="" type="checkbox"/> Handwritten <input checked="" type="checkbox"/> Medieval manuscript <input checked="" type="checkbox"/> Drop caps <input checked="" type="checkbox"/> Decorative drop caps <input checked="" type="checkbox"/> Decorative borders <input checked="" type="checkbox"/> Frames <input checked="" type="checkbox"/> Multi-column layout <input checked="" type="checkbox"/> Rotated text <input checked="" type="checkbox"/> Multiple colours in illustrations <input checked="" type="checkbox"/> Multiple colours in text <input checked="" type="checkbox"/> Reverse video
Production Characteristics	<input checked="" type="checkbox"/> Textured paper <input checked="" type="checkbox"/> Uneven character spacing <input checked="" type="checkbox"/> Multiple colours in annotations <input checked="" type="checkbox"/> Multiple colours in stamps <input checked="" type="checkbox"/> Narrow border <input checked="" type="checkbox"/> Low paper to text contrast <input checked="" type="checkbox"/> Impressions <input checked="" type="checkbox"/> Watermarks <input checked="" type="checkbox"/> Halftoning
Production Faults	<input checked="" type="checkbox"/> Uneven ink distribution <input checked="" type="checkbox"/> Bleed-through <input checked="" type="checkbox"/> Ink from facing page <input checked="" type="checkbox"/> Broken characters <input checked="" type="checkbox"/> Faint characters <input checked="" type="checkbox"/> Blurred characters <input checked="" type="checkbox"/> Smearred ink <input checked="" type="checkbox"/> Filled-in characters <input checked="" type="checkbox"/> Sort shoulder artefacts <input checked="" type="checkbox"/> Horizontally touching characters <input checked="" type="checkbox"/> Vertically touching characters <input checked="" type="checkbox"/> Non-straight text lines
Use/Wear	<input checked="" type="checkbox"/> Folds <input checked="" type="checkbox"/> Tears <input checked="" type="checkbox"/> Holes <input checked="" type="checkbox"/> Missing parts <input checked="" type="checkbox"/> Stains <input checked="" type="checkbox"/> Scratches (microfilm) <input checked="" type="checkbox"/> Paper repairs <input checked="" type="checkbox"/> Clear tape <input checked="" type="checkbox"/> Staples <input checked="" type="checkbox"/> Punch holes <input checked="" type="checkbox"/> Annotations <input checked="" type="checkbox"/> Stamps
Ageing/Preservation	<input checked="" type="checkbox"/> Warped paper <input checked="" type="checkbox"/> General paper discolouration <input checked="" type="checkbox"/> Discoloured paper edges <input checked="" type="checkbox"/> Mould <input checked="" type="checkbox"/> Non-straight paper edges <input checked="" type="checkbox"/> Fading ink
Digitisation - Geometric Distortions/Properties	<input checked="" type="checkbox"/> Skew <input checked="" type="checkbox"/> Non-uniform skew <input checked="" type="checkbox"/> Page curl <input checked="" type="checkbox"/> Perspective distortions <input checked="" type="checkbox"/> Incomplete scan <input checked="" type="checkbox"/> Tight scan margins <input checked="" type="checkbox"/> Double-page <input checked="" type="checkbox"/> Parts of opposite page visible <input checked="" type="checkbox"/> Document parts not belonging to page <input checked="" type="checkbox"/> Scanner background visible
Digitisation - Noise/Artefacts	<input checked="" type="checkbox"/> Show-through <input checked="" type="checkbox"/> Uneven illumination <input checked="" type="checkbox"/> Out-of-focus <input checked="" type="checkbox"/> Noise from scanner <input checked="" type="checkbox"/> Low scan contrast <input checked="" type="checkbox"/> Paper clips visible <input checked="" type="checkbox"/> Fingers visible <input checked="" type="checkbox"/> Salt-and-pepper noise <input checked="" type="checkbox"/> Missing information after binarisation <input checked="" type="checkbox"/> Noise and remnants after binarisation <input checked="" type="checkbox"/> Dithering

Figure 11 — Keywords section of details view, expanded.

### 2.2.2. Direct Access

The direct access approach allows users to directly access specific images or ground truth files hosted on the dataset. This feature can be used to integrate the hosted images (and therefore the dataset) in external applications.

The direct access API has been developed to support completely independent requests for each resource. This means that each time a call is made for accessing a resource, the user needs to authenticate to the system.

Authentication is crucial for checking whether the specific user is authorised to access a specific resource (image, ground truth etc). This is to enforce various different copyright rules that might require certain material available under specific agreements

#### 2.2.2.1. Authentication

Authentication is the process where the user makes their identity known to the server and the server accepts this identity. At the end of this process, the server knows who the user is. It can then proceed to check whether this user is allowed to access the requested resource (authorisation).

There are three different ways for a user to authenticate on the repository system:

- OTAT (One Time Access Token)
- HTTPS
- PHP Session

#### OTAT

By sending a One Time Access Code (64 byte random hashed string). These codes are issued by the system to known users and are valid for accessing a single item one time. The code is used to authenticate a user, but does not guarantee access to the image (this depends on whether that user is authorised to access the requested resource).

```
http(s)://url-for-required-resource/otat=09a0b25c73300b95bf7c6d4a1711f8ec60a76ae  
ee2ad2778630e3a6f2d0dbab5&resourceParam1=123&resourceParam2=456
```

#### HTTPS

It is also possible to provide a user name and password for authentication. These are the same as the credentials used for the web interface. For security purposes (since the password is sent via the network), calls that use this authentication method are only accepted via secure HTTP (HTTPS).

```
https://url-for-required-resource/user=you@email.com&pass=yourPassword&resource  
Param1=123&resourceParam2=456
```

### PHP Session

For web applications that are hosted on the same server as the repository, it is also possible to authenticate using a PHP session (which does not require sending the user's credentials with every request).

| <http://url-for-required-resource/resourceParam1=123&resourceParam2=456> |

#### ***2.2.2.2. Authorisation***

The authorisation step is completely out of the user's control. Once the user is authenticated, the system will retrieve the permissions granted to the specific user and will decide whether access to the requested resource is allowed or not.

A permissions management system has been implemented in order to allow granting and revoking permissions to specific users on defined sets of images.

This is necessary for a variety of reasons, including:

- Implementation of different licensing options for certain images,
- Implementation of different models of commercialising the hosted material (for example providing a limited set to basic members and a more extensive to full members).

#### ***2.2.2.3. Checking if a document image exists***

In order to assist integration to other system, it is possible to query the repository whether a document image (given the Document ID) exists, using the **checkDid** script.

Since this operations does not actually allow access to the image or metadata, it does not require for a user to be authenticated.

#### Resource specific parameters

Did            The Document ID to check

#### Return values

0            Document image with this ID does not exist

1            Document image with this ID exists

HTTP400    Missing/Invalid parameter(s)

HTTP403    Execution not allowed

#### Examples

| <http://www.primaresearch.org/repository/api/checkDid?Did=541608> |

#### *2.2.2.4. Accessing an Image*

In order to access an image from the dataset, the user needs to know the Document ID for the required image. It is then possible to call the **getImage** script directly from any client/application.

##### Resource specific parameters

Did	The Document ID of the required image
type	The type of image required. Possible values are
thumb	150x150 pixels PNG image
view	350x350 pixels PNG image
fullview	Full size JPEG image
orig	Full size TIFF image (using lossless LZW compression)

##### Return values

image	The requested image, if allowed
HTTP400	Missing/Invalid parameter(s)
HTTP403	Execution not allowed, or access to requested resource not allowed

##### Examples

```
http://www.primaresearch.org/repository/api/getImage?otat=0654fdfde87d6766c4a686b2e2f613e4b114c60398769c5a2b63e8fce56b25c7&Did=541608&type=orig
```

```
https://www.primaresearch.org/repository/api/getImage?user=you@email.com&pass=yourPassword&Did=541608&type=thumb
```

#### *2.2.2.5. Accessing an Attachment*

In order to access an attachment from the dataset, the user needs to know either the Attachment ID or the relevant Document ID and the type of attachment required. It is then possible to call the **getAttachment** script directly from any client/application.

If the second set of parameters is used (Document ID and attachment type), the latest attachment of the requested type will be returned (in case there are more than one versions).

##### Resource specific parameters

Aid	The Attachment ID of the required resource
or	
Did	The Document ID of the document image containing the attachment



ATid            The type of attachment required (form the list below)

- 1 Image
- 2 ALTO file
- 3 METS file
- 4 PAGE ground truth
- 5 XML file (any content)
- 6 PAGE ground truth (text content only)
- 7 Text file (result from a method)
- 8 ALTO file (result from a method)
- 9 ABBYY Finereader XML (result)

### Return values

file            The requested attachment, if allowed

HTTP400       Missing/Invalid parameter(s)

HTTP403       Execution not allowed, or access to requested resource not allowed

### Examples

```
https://www.primaresearch.org/repository/api/getAttachment?user=you@email.com&password=yourPassword&Aid=199756
```

```
https://www.primaresearch.org/repository/api/getAttachment?user=you@email.com&password=yourPassword&Did=541608&ATid=4
```



## 2.3. Exploitation

This section outlines applications where this repository has been used, as known to the authors at the time of writing this report, as well as plans for future use.

### 2.3.1. Uses of the Repository management system

The repository management system and web interface that was custom built for hosting and accessing the images has been reused for the management of datasets for other purposes:

- Europeana Newspapers Project [5]: The repository management system, in terms of supporting database and web front-end has been reused for managing the evaluation repository created for the EU funded Europeana Newspapers Project.
- PRImA Contemporary Documents Layout Analysis Dataset [9]: This dataset of realistic documents reflecting the various challenges in contemporary document layout analysis is currently being adopted to make use of this repository management infrastructure and web interface.
- PRImA Natural History Museum Cards Dataset [10]: This dataset containing scans of index cards from the UK's Natural History Museum lepidoptera index, is also being adopted to make use of this repository management infrastructure and web interface.

### 2.3.2. Integration to other systems

As part of this project, it has been integrated to the evaluation platform described later in this report, and also used in the Work Package 2 Interoperable platform.

### 2.3.3. Dataset

Images from the dataset have been used by the PRImA Lab for two International Competitions (also as part of this project for T5.3):

- Competition on Historical Book Recognition 2013 [1]
- Competition on Historical Newspapers Layout Analysis 2013 [2]
- PRImA has successfully ran the ICDAR Page Segmentation competitions since 2001 and will keep running competitions in this theme, aiming at using images and ground truth that was made available via this initiative.



### 3. EVALUATION INFRASTRUCTURE

There are numerous studies on the performance of specific OCR systems achieved on various collections and particular types of documents. When it comes to comparing the findings of such studies, however, it becomes apparent that there is only little standardisation in the way that results are evaluated. The SUCCEED Evaluation Infrastructure is therefore addressing the need for a common point of reference which allows to compare the results of different methods on standard datasets in a reliable and reproducible manner. The approach of a centralised evaluation infrastructure bears numerous advantages over the scattered and hard-to-verify evaluation practice which has been predominant so far:

- Results are reliably recorded and can be reproduced based on extensive metadata (what dataset was used, what results were obtained, was any normalisation applied prior to evaluation, what evaluation metrics and weights were used).
- Datasets and evaluation tools can be tightly integrated.
- Fully automated evaluation processes can be provided.
- The state-of-the-art can be monitored very easily – it will be immediately clear if a new method is superior to pre-existing ones.
- Specific challenges can be defined and scores for different use scenarios maintained (for instance allowing to judge if a method which might not yield an over-all/average improvement can lead to better results for a very specific problem).
- Ongoing competitions can be hosted.
- Research into specific problems can be encouraged by publishing new challenges.

In terms of the SUCCEED project the Evaluation Infrastructure was also a technical requirement for related tasks such as the hosting of competitions.

During this project, numerous improvements and adjustments were made to the tools used for the evaluation infrastructure. Following those, web services were developed to allow the integration of the tools into the new online evaluation platform.

Mainly, the integrated PRImA Layout Evaluation tool was upgraded to support the latest PAGE format that offers more region types, a more robust representation structure and more compact representation of polygons (decreasing ground truth file size by up to 60%). There were also updates in the format used to store evaluation results and profiles to conform to the updated PAGE format and to improve efficiency in storing the results. Also, all evaluation profiles used both in the platform and already in two international competitions have been updated to conform to the new PAGE format and include all new region types defined.

Finally, a number of new OCR evaluation tools (see 3.2.2) were developed and integrated to the evaluation platform, enabling three different approaches in text evaluation.



### 3.1. Purpose

The purpose of the evaluation platform is to provide a web based interface to an array of evaluation tools. Using this platform, researchers are able to upload results from their methods and evaluate them based on pre-defined scenarios. Evaluation results can then be compared with previous methods/results of the same researcher and/or result of state of the art methods for the same document images.

The platform currently supports evaluation for the levels and methods listed in Table 10.

Method	Profile	Expected input
<b>Text Evaluation (OCR)</b>		
Bag of words	No text normalisation	PAGE
	Default Normalisation (v1)	PAGE
Word Accuracy	No text normalisation	PAGE
	Default Normalisation (v1)	PAGE
Character Accuracy	No text normalisation	PAGE
	Default Normalisation (v1)	PAGE
<b>Regions Layout Evaluation</b>		
PRImA Layout Evaluation	Segmentation	PAGE
	OCR Scenario	PAGE
<b>Words Layout Evaluation</b>		
PRImA Layout Evaluation	Default	PAGE
<b>Lines Layout Evaluation</b>		
PRImA Layout Evaluation	Default	PAGE
<b>Glyphs Layout Evaluation</b>		
PRImA Layout Evaluation	Default	PAGE
<b>Border Layout Evaluation</b>		
PRImA Layout Evaluation	Default	PAGE

Table 10 — Evaluation methods supported

### 3.2. Tools used

The main tools used in the evaluation platform were developed by PRImA as part of the IMPACT project [8] and the European Newspapers Project [5]. The two tools integrated are described below.

#### 3.2.1. Layout Evaluation Tool

The Layout Evaluation tool [10] 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.

The layout evaluation tool comes in two versions:

- **Command line executable:** for workflow integration and batch processing. This was used to integrate the tool in the evaluation platform.
- **Stand-alone GUI software:** for inspection of individual results and in-depth analysis of specific problems (see Figure 12).

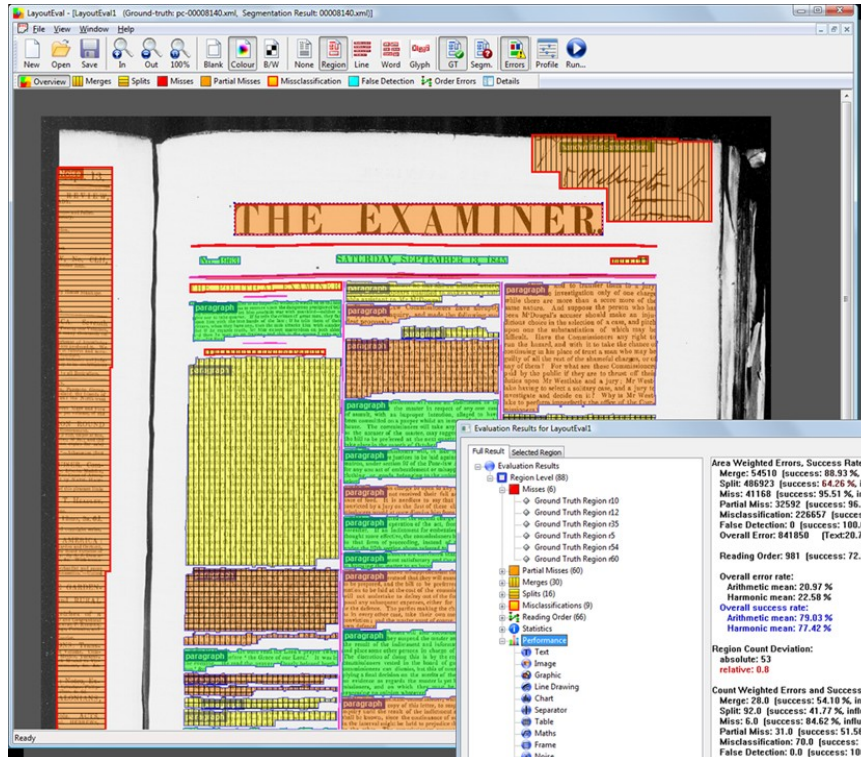


Figure 12 — Layout Evaluation Tool

### 3.2.2. Text (OCR) Evaluation Tool

A completely new command line tool for evaluating the text accuracy of OCR engines was implemented. The 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 and use different normalisation profiled
- **Support of all latest file formats:** PAGE XML, ALTO, ABBYY XML, hOCR.



### 3.3. Platform description

The evaluation platform allows users to upload the result files of a Layout Analysis and/or OCR method in order to evaluate them according to the latest ground truth available. In order for the evaluation to work and produce useful results, the system only accepts files that are already part of the image repository (described previously in this report) and have ground truth associated with them (to evaluate against).

#### 3.3.1. Web front end

The first step from the user's perspective is to log in to the web based evaluation portal and upload result files to the system (see Figure 13). At the same time, they are offered a choice of the available evaluation methods and profiles to choose how they want their results to be evaluated. More detailed information on the available method/profile combinations is listed in Table 10.

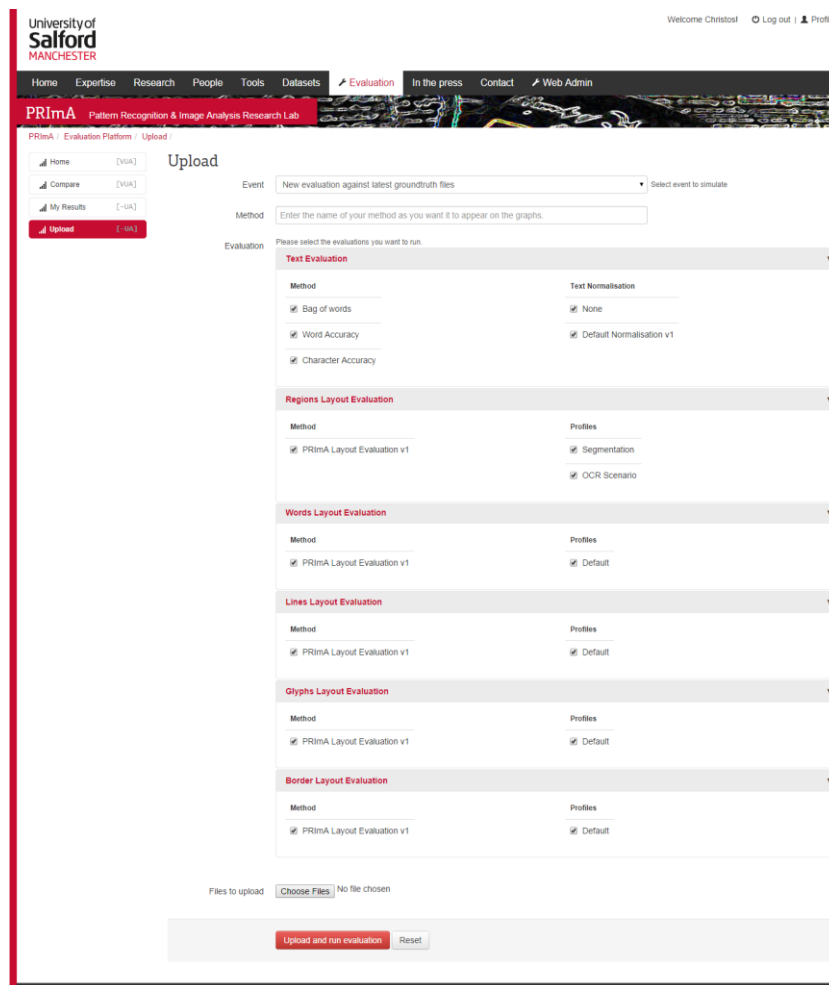


Figure 13 — File upload and evaluation method selection interface

Since it is more than likely the user will try to upload multiple files, the total file size is checked to predict whether the upload is likely to fail. In such case, the user is notified about which files are likely to succeed and which are likely to fail (see Figure 14 and Figure 15). This check is performed using jQuery<sup>1</sup> on the client side, to save upload time and bandwidth.

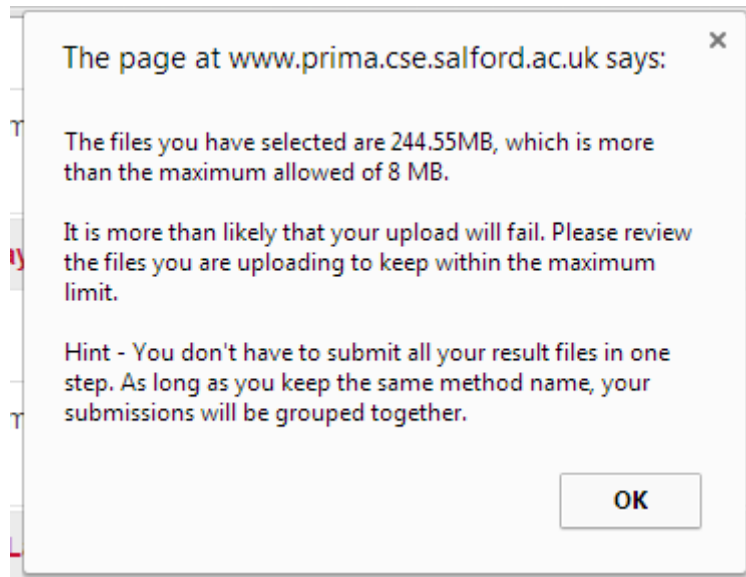


Figure 14 — Popup warning

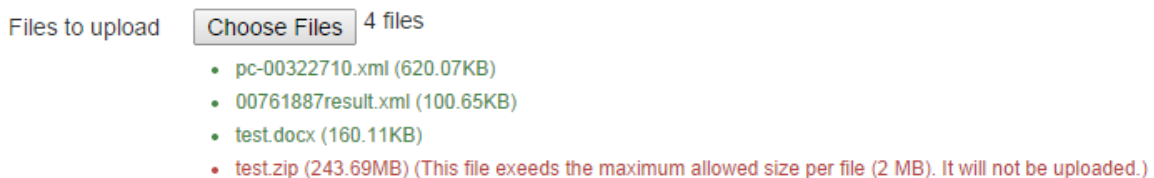


Figure 15 — Details of files to be uploaded

Once the result files are uploaded, the system attempts to link the uploaded result file with a document image that is on the repository in order to obtain the latest ground truth available. This is done by identifying the first eight characters of the filename and using them to query the repository whether a document image with such ID exists.

Filename: pc-00322710.xml	assumes Document ID: <b>pc-00322</b>	<b>WRONG</b>
Filename: 00761887result.xml	assumes Document ID: <b>00761887</b>	<b>CORRECT</b>

<sup>1</sup> jQuery is a cross-platform JavaScript library, designed to simplify client-side scripting in web sites and applications.



Once the system is satisfied that a record in the repository exists for the uploaded document image, it proceeds with queuing the evaluation of the uploaded file using all selected evaluation methods and profiles (see Figure 16).

The following files failed

Filename	Size	Upload Status
pc-00322710.xml	620.07 KB	DID_NOT_FOUND
test.docx	160.11 KB	DID_NOT_FOUND

Files queued for method: PRImA Test

Queue ID	Document ID	Filename	Size	Upload Status	Level	Method	Parameters	Evaluation Status
1260	00761887	00761887result.xml	100.65 KB	OK	Text	Bag of words	Text Normalisation: None	QUEUED
1261	00761887	00761887result.xml	100.65 KB	OK	Text	Bag of words	Text Normalisation: Default Normalisation v1	QUEUED
1262	00761887	00761887result.xml	100.65 KB	OK	Text	Word Accuracy	Text Normalisation: None	QUEUED
1263	00761887	00761887result.xml	100.65 KB	OK	Text	Word Accuracy	Text Normalisation: Default Normalisation v1	QUEUED
1264	00761887	00761887result.xml	100.65 KB	OK	Text	Character Accuracy	Text Normalisation: None	QUEUED
1265	00761887	00761887result.xml	100.65 KB	OK	Text	Character Accuracy	Text Normalisation: Default Normalisation v1	QUEUED
1266	00761887	00761887result.xml	100.65 KB	OK	Regions	PRImA Layout Evaluation v1	Profile: Segmentation	QUEUED
1267	00761887	00761887result.xml	100.65 KB	OK	Regions	PRImA Layout Evaluation v1	Profile: OCR Scenario	QUEUED
1268	00761887	00761887result.xml	100.65 KB	OK	Lines	PRImA Layout Evaluation v1	Profile: Default	QUEUED
1269	00761887	00761887result.xml	100.65 KB	OK	Words	PRImA Layout Evaluation v1	Profile: Default	QUEUED
1270	00761887	00761887result.xml	100.65 KB	OK	Glyphs	PRImA Layout Evaluation v1	Profile: Default	QUEUED
1271	00761887	00761887result.xml	100.65 KB	OK	Border	PRImA Layout Evaluation v1	Profile: Default	QUEUED

Figure 16 — Files queued for evaluation

During the queuing process the system needs to ensure that it passes the required parameters to the evaluation server, so it can access all required files for the evaluation requested. These parameters include:

- URIs for all necessary image and ground truth files. This involves generating One Time Access Tokens (see 2.2.2.1), so that the evaluation server is granted access to the required resources.
- URI for the user uploaded files containing their results
- Call back URL. This is a URL the evaluation server calls when the evaluation is finished to notify the web platform of the fact is has finished and to submit the evaluation results to.

The evaluation server then connects to the image repository to verify the current user is allowed to access the files they are evaluating, and assuming they do it proceeds with retrieving the image (if required) and ground truth file in order to proceed with the evaluation. During this process, the user can monitor the status of their requested evaluations, using the My Results page, where they can see a list of all evaluation requests and whether they are still queued (see Figure 17), completed (Figure 18) or failed (Figure 19).

My Results PRImA Test 2 All Levels Last 30 days Go

There are 72 evaluation results.

Evaluation ID	Document ID	Your Filename	Your Method	Evaluation Level	Evaluation Method	Evaluation Parameters	Evaluation Status	Time Uploaded	Time Finished
1343	761887	00761887resultCopy5.xml	PRImA Test 2	Border	PRImA Layout Evaluation v1	Profile: Default	QUEUED	2014-09-12 10:30:52	
1342	761887	00761887resultCopy4.xml	PRImA Test 2	Border	PRImA Layout Evaluation v1	Profile: Default	QUEUED	2014-09-12 10:30:52	
1341	761887	00761887resultCopy3.xml	PRImA Test 2	Border	PRImA Layout Evaluation v1	Profile: Default	QUEUED	2014-09-12 10:30:52	
1340	761887	00761887resultCopy2.xml	PRImA Test 2	Border	PRImA Layout Evaluation v1	Profile: Default	FINISHED	2014-09-12 10:30:52	2014-09-12 10:31:01
1339	761887	00761887resultCopy.xml	PRImA Test 2	Border	PRImA Layout Evaluation v1	Profile: Default	FINISHED	2014-09-12 10:30:52	2014-09-12 10:30:58
1338	761887	00761887result.xml	PRImA Test 2	Border	PRImA Layout Evaluation v1	Profile: Default	FINISHED	2014-09-12 10:30:52	2014-09-12 10:30:56
1337	761887	00761887resultCopy5.xml	PRImA Test 2	Glyphs	PRImA Layout Evaluation v1	Profile: Default	QUEUED	2014-09-12 10:30:52	
1336	761887	00761887resultCopy4.xml	PRImA Test 2	Glyphs	PRImA Layout Evaluation v1	Profile: Default	QUEUED	2014-09-12 10:30:52	
1335	761887	00761887resultCopy3.xml	PRImA Test 2	Glyphs	PRImA Layout Evaluation v1	Profile: Default	QUEUED	2014-09-12 10:30:52	

Figure 17 — Queued evaluation requests

My Results All Methods All Levels Last 30 days Go

There are 823 evaluation results.

Evaluation ID	Document ID	Your Filename	Your Method	Evaluation Level	Evaluation Method	Evaluation Parameters	Evaluation Status	Time Uploaded	Time Finished
1271	761887	00761887result.xml	PRImA Test	Border	PRImA Layout Evaluation v1	Profile: Default	FINISHED	2014-09-12 09:27:46	2014-09-12 09:28:05
1270	761887	00761887result.xml	PRImA Test	Glyphs	PRImA Layout Evaluation v1	Profile: Default	FINISHED	2014-09-12 09:27:46	2014-09-12 09:28:07
1269	761887	00761887result.xml	PRImA Test	Words	PRImA Layout Evaluation v1	Profile: Default	FINISHED	2014-09-12 09:27:46	2014-09-12 09:28:06
1268	761887	00761887result.xml	PRImA Test	Lines	PRImA Layout Evaluation v1	Profile: Default	FINISHED	2014-09-12 09:27:46	2014-09-12 09:28:06
1267	761887	00761887result.xml	PRImA Test	Regions	PRImA Layout Evaluation v1	Profile: OCR Scenario	FINISHED	2014-09-12 09:27:46	2014-09-12 09:28:23
1266	761887	00761887result.xml	PRImA Test	Regions	PRImA Layout Evaluation v1	Profile: Segmentation	FINISHED	2014-09-12 09:27:46	2014-09-12 09:28:14
1265	761887	00761887result.xml	PRImA Test	Text	Character Accuracy	Text Normalisation: Default Normalisation v1	FINISHED	2014-09-12 09:27:46	2014-09-12 09:27:46
1264	761887	00761887result.xml	PRImA Test	Text	Character Accuracy	Text Normalisation: None	FINISHED	2014-09-12 09:27:46	2014-09-12 09:27:46
1263	761887	00761887result.xml	PRImA Test	Text	Word Accuracy	Text Normalisation: Default Normalisation v1	FINISHED	2014-09-12 09:27:46	2014-09-12 09:27:46
1262	761887	00761887result.xml	PRImA Test	Text	Word Accuracy	Text Normalisation: None	FINISHED	2014-09-12 09:27:46	2014-09-12 09:27:46

Figure 18 — Successful evaluation requests

1177	761887	00761887resultCopy5.xml	dasdsada	Text	Bag of words	Text Normalisation: Default Normalisation v1	FINISHED	2014-08-19 16:56:16	2014-08-19 16:56:17
1176	761887	00761887resultCopy5.xml	dasdsada	Text	Bag of words	Text Normalisation: None	FINISHED	2014-08-19 16:56:16	2014-08-19 16:56:17
1175	8455	00008455.metadata.xml	dasdsada	Border	PRImA Layout Evaluation v1	Profile: Default	FAILED	2014-08-19 16:55:14	
1174	8455	00008455.metadata.xml	dasdsada	Glyphs	PRImA Layout Evaluation v1	Profile: Default	FAILED	2014-08-19 16:55:14	
1173	8455	00008455.metadata.xml	dasdsada	Words	PRImA Layout Evaluation v1	Profile: Default	FAILED	2014-08-19 16:55:13	

Figure 19 — Failed evaluation requests

It is also possible for the user to filter the results view by their method, the level evaluated (text, regions, lines, words, glyphs or border) and the evaluation date (see Figure 20).

My Results PRImA Test Regions Last 30 days Go

There are 2 evaluation results.

Evaluation ID	Document ID	Your Filename	Your Method	Evaluation Level	Evaluation Method	Evaluation Parameters	Evaluation Status	Time Uploaded	Time Finished
1267	761887	00761887result.xml	PRImA Test	Regions	PRImA Layout Evaluation v1	Profile: OCR Scenario	FINISHED	2014-09-12 09:27:46	2014-09-12 09:28:23
1266	761887	00761887result.xml	PRImA Test	Regions	PRImA Layout Evaluation v1	Profile: Segmentation	FINISHED	2014-09-12 09:27:46	2014-09-12 09:28:14

Figure 20 — Filtered view of results

Finally, the user can quickly preview all the details for any completed evaluation (see Figure 21).

Result preview for evaluation 1264 ✕

ARRAY	
Eid	1264
Uemail	chris@cpapadopoulos.co.uk
Mid	13
Did	761887
Qfilename	1410510465.00761887result.xml
Qtimestamp	2014-09-12 09:27:46
EtextNormalisation	none
charsInGroundTruth	10979
charsInResult	10979
characterAccuracy	1

Close

Figure 21 — Preview of evaluation record

Once the evaluation is complete, it sends back the result of the evaluation to the web server, where it gets stored.

The user can at any time monitor the progress of their evaluation queue and see/compare the results of finished evaluations.

Once all queued evaluations are finished, the user can view a graph comparing their own methods (current and previous submissions) along with results from pre-implemented state of the art methods (see Figure 22).

### Compare

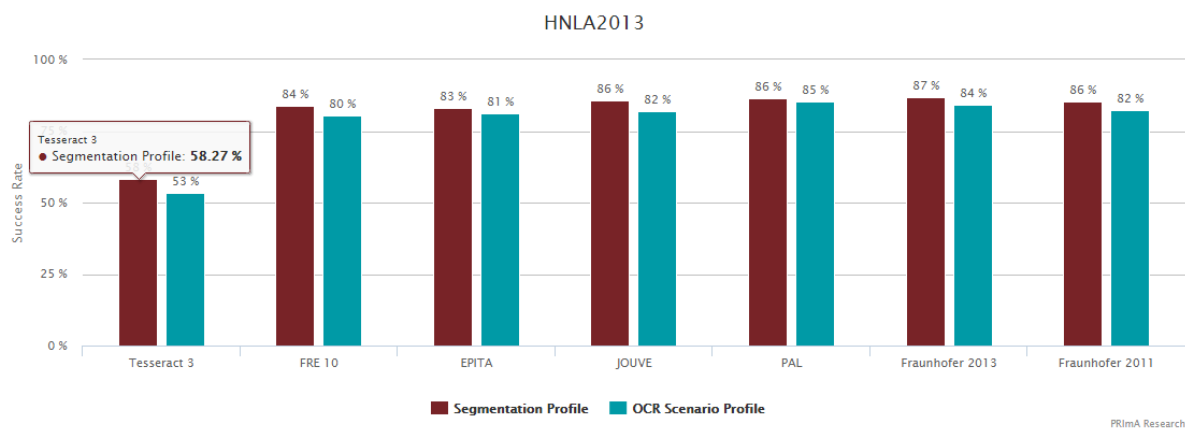


Figure 22 — Comparison chart

### 3.3.2. Evaluation server

The evaluation server is a dedicated machine, providing a web service interface to access all the evaluation methods it hosts (see Table 10).

For each evaluation tool supported, a Java class wrapper has been implemented to provide access to the command line interface of the tool. This class was then used to create a full SOAP 1.2 web service. As a final step, these services were exported as WAR files and deployed on an Apache Tomcat Server [5]. In order to protect the web services from unauthorised access, Apache's basic authentication has been used.

The evaluation server provides two web services that give access to all the evaluation methods described previously; one for the text evaluation methods and one for the layout evaluation methods.

#### 3.3.2.1. Text Evaluation Web Service

This service provides access to all the available text evaluation methods. A summary of the available methods for this service is provided in the following sections.

### Bag of Words using PAGE file

Runs the PRImA text evaluation tool with the "Bag of Words" method and returns CSV output.

Parameter	Type	Description
<b>groundTruth PageFileUrl</b>	String	Ground truth text (PAGE XML file)
<b>result PageFileUrl</b>	String	Result text to be evaluated (PAGE XML file)
<b>text Normalisation RulesUrl</b>	String	[Optional] Text replacement rules for normalisation (.xml file)
<b>callbackUrl</b>	String	[Optional] URL to be called when finished
<b>runAsync</b>	Boolean	If set to true and a callbackUrl is provided, the evaluation is run asynchronously (result content "Task started")
<b>[return]</b>	String[]	Array with result code and comma separated values with headings (or error message)

Table 11 — Bag of Words Parameters

### Multiple files Bag of Words using PAGE file

Runs the PRImA text evaluation tool with the "Bag of Words" method for multiple document images and returns the CSV output.

Parameter	Type	Description
<b>groundTruth PageFileUrl</b>	String[]	Ground truth texts (PAGE XML files)
<b>result PageFileUrl</b>	String[]	Result texts to be evaluated (PAGE XML files)
<b>text Normalisation RulesUrl</b>	String[]	[Optional] Text replacement rules for normalisation (.xml file)
<b>callbackUrl</b>	String[]	[Optional] URL to be called when finished
<b>runAsync</b>	Boolean	If set to true and a callbackUrl is provided, the evaluation is run asynchronously (result content "Task started")
<b>[return]</b>	String[]	Array with result code and result content (or error message). Content: Comma separated values with headings for each document image, separated by two line breaks

Table 12 — Multiple Bag of Words Parameters

### Text Evaluation using PAGE file

Runs the PRImA text evaluation tool and returns CSV output.

Parameter	Type	Description
groundTruth PageFileUrl	String	Ground truth text (PAGE XML file)
result PageFileUrl	String	Result text to be evaluated (PAGE XML file)
text Normalisation RulesUrl	String	[Optional] Text replacement rules for normalisation (.xml file)
bagOfWords	Boolean	Run "Bag of Words" method
wordAccuracy	Boolean	Run word accuracy method
character Accuracy	Boolean	Run character accuracy method
callbackUrl	String	[Optional] URL to be called when finished
runAsync	Boolean	If set to true and a callbackUrl is provided, the evaluation is run asynchronously (result content "Task started")
[return]	String[]	Array with result code and comma separated values with headings (or error message)

Table 13 — Text Evaluation Parameters

### Multiple files Text Evaluation using PAGE file

Runs the PRImA text evaluation tool for multiple document images and returns CSV output.

Parameter	Type	Description
groundTruth PageFileUrl	String[]	Ground truth texts (PAGE XML files)
result PageFileUrl	String[]	Result texts to be evaluated (PAGE XML files)
text Normalisation RulesUrl	String[]	[Optional] Text replacement rules for normalisation (.xml file)
bagOfWords	Boolean	Run "Bag of Words" method
wordAccuracy	Boolean	Run word accuracy method
character Accuracy	Boolean	Run character accuracy method
callbackUrl	String[]	[Optional] URL to be called when finished
runAsync	Boolean	If set to true and a callbackUrl is provided, the evaluation is run asynchronously (result content "Task started")
[return]	String[]	Array with result code and result content (or error message). Content: Comma separated values with headings for each document

image, separated by two line breaks

Table 14 — Multiple Text Evaluation Parameters

### 3.3.2.2. Layout Evaluation Web Service

This service provides access to all the available layout evaluation methods. This includes regions, lines, words and glyphs layouts as well as page border. A summary of the available methods for this service is provided in the following sections.

#### Layout Evaluation using PAGE file

Runs the PRImA layout evaluation tool and returns CSV output. If multiple content levels (regions, text lines, words and glyphs) are evaluated, the result will contain one row of comma separated values per level (to be distinguished by 'Level' column).

Parameter	Type	Description
groundTruth LayoutUrl	String	Page layout ground truth (PAGE XML file)
result LayoutUrl	String	Page layout result that is to be evaluated (PAGE XML file)
imageUrl	String	Document page image (bitonal preferred; .tif/.png/.jpg)
evaluation ProfileUrl	String	Evaluation profile with weights and settings (.evx file)
evaluate Regions	Boolean	Set to <i>true</i> to evaluate on region level
evaluate Lines	Boolean	Set to <i>true</i> to evaluate on text line level
evaluate Words	Boolean	Set to <i>true</i> to evaluate on word level
evaluate Glyphs	Boolean	Set to <i>true</i> to evaluate on glyph level
evaluate Border	Boolean	Set to <i>true</i> to evaluate the page border
callbackUrl	String	[Optional] URL to be called when finished
runAsync	Boolean	If set to <i>true</i> and a callbackUrl is provided, the evaluation is run asynchronously (result content "Task started")
[return]	String[]	Array with result code and comma separated values with headings (or error message)

Table 15 — Layout Evaluation Parameters

#### Multiple files Layout Evaluation using PAGE file

Runs the PRImA layout evaluation tool for multiple document images and returns the CSV output. If multiple content levels (regions, text lines, words and glyphs) are





evaluated, the result will contain one row of comma separated values per level (to be distinguished by 'Level' column).

Parameter	Type	Description
groundTruth LayoutUrl	String[]	Page layout ground truths (PAGE XML files)
result LayoutUrl	String[]	Page layout results that are to be evaluated (PAGE XML files)
imageUrl	String[]	Document page images (bitonal preferred; .tif/.png/.jpg)
evaluation ProfileUrl	String[]	Evaluation profiles with weights and settings (.evx file)
evaluate Regions	Boolean	Set to <i>true</i> to evaluate on region level
evaluate Lines	Boolean	Set to <i>true</i> to evaluate on text line level
evaluate Words	Boolean	Set to <i>true</i> to evaluate on word level
evaluate Glyphs	Boolean	Set to <i>true</i> to evaluate on glyph level
evaluate Border	Boolean	Set to <i>true</i> to evaluate the page border
callbackUrl	String[]	[Optional] URLs to be called when evaluation of document finished
runAsync	Boolean	If set to <i>true</i> and a callbackUrl is provided, the evaluation is run asynchronously (result content "Task started")
[return]	String[]	Array with result code and comma separated values with headings (or error message). Content: Comma separated values with headings for each document image, separated by two line breaks

Table 16 — Multiple Layout Evaluation Parameters

### 3.4. Exploitation

A number of the algorithms used for creating this platform have been used by the PRImA Lab for two International Competitions (also as part of this project for T5.3):

- Competition on Historical Book Recognition 2013 [1]
- Competition on Historical Newspapers Layout Analysis 2013 [2]

The algorithms used included the evaluation tools, profiles and statistics/summary generation for evaluating results of different methods.

The evaluation platform will be available for registered users to upload the results of their methods and compare them with state of the art methods. In addition, it will be used to support the organisation and running of further competitions.

We are currently planning the next competitions on the same series that will run for ICDAR2015 (International Conference on Document Analysis and Recognition).

Apart from the platform, the web services created for the evaluation methods all use a standard SOAP protocol, therefore it can be easily arranged (after agreeing access restriction for security purposes) to integrate them into any other platform that is capable of using standard SOAP web services.

## 4. GLOSSARY

ALTO	Analyzed Layout and Text Object: An open XML Schema to describe OCR text and layout information of printed documents.
API	Application Programming Interface: It specifies a software component in terms of its operations, inputs and outputs and underlying types. Its main purpose is to define a set of functionalities that are independent of their respective implementation, allowing both definition and implementation to vary without compromising each other.
Binary Image	A black and white version of an image (used frequently by layout analysis and OCR methods).
BPP	Bits Per Pixel: Unit for measuring the colour depth of an image.
Colour Depth	The number of bits used to indicate the colour of a single pixel, or the number of bits used for each colour component of a single pixel.
CSV	Comma Separated Values: A file that stores tabular data (numbers and text) in plain-text form, using a comma as the separator character.
DPI	Dots Per Inch: Unit for measuring printing or digitising dot density, in particular the number of individual dots that can be placed in a line within the span of 1 inch.
HTML	HyperText Markup Language: The standard markup language used to create web pages.
Ground truth	Ground truth, in this context, is an exact and formalised reproduction of what is actually present on the physical page or, to put it in other words, what the perfect analysis/recognition method is expected to return as result [3].
GUI	Graphical User Interface: A windows based type of interface that allows users to interact through graphical icons and visual indicators.
hOCR	A format for representing OCR output, including layout information, character confidences, bounding boxes, and style information. It embeds this information invisibly in standard HTML [7].
jQuery	A cross-platform JavaScript library.
LDAP	Lightweight Directory Access Protocol: An open, vendor-neutral, industry standard application protocol for accessing and maintaining distributed directory information services over a network.



OCR	Optical Character Recognition: The mechanical or electronic conversion of scanned or photographed images of typewritten or printed text into machine-encoded/computer-readable text.
OTAT	One Time Access Code: An access code used to identify users, without the need for them to provide a username and password.
PAGE	Page Analysis and Ground-Truth Elements: A XML-based page image representation framework that records information on image characteristics, layout structure and page content [4].
PHP	A server-side scripting language designed for web development.
Pixel	In digital imaging, a pixel, or picture element, is a physical point in a raster image.
Resolution	A measure of the amount of detail in an image. It is measured in dpi.
SOAP	Simple Object Access protocol: A protocol specification for exchanging structured information in the implementation of web services in computer networks.
URI	Uniform Resource Identifier: A string of characters used to identify a name of a resource.
URL	Uniform Resource Locator (also known as a web address, particularly when used with HTTP) is a specific character string that constitutes a reference to a resource.
XML	eXtensible Markup Language: A free opens standard markup language that defines a set of rules for encoding documents in a format that is both human-readable and machine-readable.

## 5. REFERENCES

- [1] A. Antonacopoulos, C. Clausner, C. Papadopoulos, S. Pletschacher, "ICDAR2013 Competition on Historical Book Recognition – HBR2013", Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR2013), Washington DC, USA, August 2013, pp. 1491-1495
- [2] A. Antonacopoulos, C. Clausner, C. Papadopoulos, S. Pletschacher, "ICDAR2013 Competition on Historical Newspaper Layout Analysis – HNLA2013", Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR2013), Washington DC, USA, August 2013, pp. 1486-1490
- [3] C. Papadopoulos, S. Pletschacher, C. Clausner, A. Antonacopoulos, "The IMPACT Dataset of Historical Document Images", Proceedings of the 2013 Workshop on Historical Document Imaging and Processing (HIP2013), Washington DC, USA, August 2013, pp. 123-130
- [4] S. Pletschacher and A. Antonacopoulos, "The PAGE (PageAnalysis and Ground-Truth Elements) Format Framework", Proc. ICPR2008, Istanbul, Turkey, August 23-26, 2010, IEEE-CS Press, pp. 257-260
- [5] Apache Tomcat, <http://tomcat.apache.org/>
- [6] Europeana Newspapers Project, EU Competitiveness and Innovation Framework Programme grant Europeana Newspapers (Ref. 297380), <http://www.europeana-newspapers.eu/>
- [7] hocr-tools, <https://code.google.com/p/hocr-tools/>
- [8] IMPACT Project, EU 7<sup>th</sup> Framework Programme grant (Ref: 215067), <http://www.impact-project.eu/>
- [9] PRImA Contemporary Documents Layout Analysis Dataset, [http://www.prima.cse.salford.ac.uk/datasets/Layout\\_Analysis](http://www.prima.cse.salford.ac.uk/datasets/Layout_Analysis)
- [10] PRImA Layout Evaluation Tool, <http://www.prima.cse.salford.ac.uk/tools/LayoutEvaluation>
- [11] PRImA Natural History Museum Lepidoptera Dataset, <http://www.prima.cse.salford.ac.uk/datasets/NHM>

