The ENP Image and Ground Truth Dataset of Historical Newspapers†

Christian Clausner, Christos Papadopoulos, Stefan Pletschacher and Apostolos Antonacopoulos
Pattern Recognition and Image Analysis (PRImA) Research Lab
School of Computing, Science and Engineering, University of Salford
Greater Manchester, United Kingdom
http://www.primaresearch.org

Abstract— **This paper presents a research dataset of historical newspapers comprising over 500 page images, uniquely representative of European cultural heritage from the digitization projects of 12 national and major European libraries, created within the scope of the large-scale digitisation *Europeana Newspapers Project (ENP)*. Every image is accompanied by comprehensive ground truth (Unicode encoded full-text, layout information with precise region outlines, type labels, and reading order) in PAGE format and searchable metadata about document characteristics and artefacts. The first part of the paper describes the nature of the dataset, how it was built, and the challenges encountered. In the second part, a baseline for two state-of-the-art OCR systems (ABBYY FineReader Engine 11 and Tesseract 3.03) is given with regard to both text recognition and segmentation/layout analysis performance.**

Keywords—image dataset; document analysis; ground truth; historical documents

I. INTRODUCTION

In order to achieve meaningful advancements in image analysis for historical documents, it is important to be aware of all challenges and idiosyncrasies presented by real-world material. Apart from objective performance measures, there is a significant need for representative datasets of images and associated ground truth. Creating such a set requires a careful selection process, since it involves considerable costs as well as limitations to access to historical collections. Newspapers pose particular challenges due to the low quality of print, the large size of images and the very tight layout, among other issues.

The Europeana Newspapers Project (ENP) had as main objectives the aggregation and refinement of historical newspaper collections for *The European Library* [1] and *Europeana* [2], to enhance search and presentation functionalities for their users. The project processed and contributed over 10 million newspaper pages with full OCR text to the above institutions.

Each library participating in the project contributed digitised newspapers and full-texts free of any legal restrictions to Europeana. Special focus was set on newspapers published during World War I, but a wider time period was also included, so that numerous historical events from before and after 1914-1918 linking directly to events within that time period. This would allow people to easily research important local, national, or European events in a broad European context, something that has been out of reach so far. The newspapers are of as high interest to the general public as to researchers, since the material is understandable without historical expert knowledge and has direct links to many European families.

The refinement technologies applied included OCR, layout analysis / article segmentation, Named Entity Recognition (NER), and page class recognition. Quality assurance and quality prediction mechanisms were created and optimised to monitor and control all refinement steps.

In view of the need of the project for performance evaluation and quality assurance, and the availability of data and resources, the authors sought to create a representative and comprehensive dataset that would outlive the project and would be useful for researchers in document analysis and recognition.

The methodology for creating the dataset followed the best practice from previous landmark dataset creation [3][4] placing particular emphasis on the dataset to be:

- *Realistic* – reflecting the library collections with regard to representativeness and frequency of documents.
- *Comprehensive* – including metadata and detailed ground truth.
- *Flexibly structured* – supporting users to search, browse, group etc. and allowing direct access to external systems (OCR workflows or evaluation tools).

To the authors' best knowledge, there is no such other large and comprehensive newspaper dataset with ground truth. The closest dataset of similar nature – that of the IMPACT project [4] – focuses mostly on books.

II. CREATING THE DATASET

The dataset was created in three stages which will be detailed in the next subsections:

1. Aggregation of a broad set of representative images and metadata.
2. Refinement of initial selection to a realistic subset.
3. Ground truth production.

A. Image and Metadata Aggregation

All content providers in the project were asked to review their newspaper collections and to arrive at a selection of images that would reflect the range of documents in their holdings. There was no limit on the number of pages to be included as long as the selection was considered representative. The stipulated approach was to start from an overview of all major titles

† This work has been funded through the EU Competitiveness and Innovation Framework Programme grant Europeana Newspapers (Ref. 297380).

and their frequencies. Next, individual characteristics like language, format, and layout would be considered in order to include all main types of newspapers. To further narrow down the number of individual newspaper issues selected, a sampling approach was recommended (e.g. one full issue per title every 2/5/10 years, depending on the title's frequency). At the same time it was encouraged to include extra issues whenever significant changes (e.g. in the layout of a newspaper) were known to have occurred. This process lead to a broad set of document images - Fig. 1 shows three representative example pages.

Fig. 1. Example images (from the collection of the Berlin State Library)

In terms of image formats and renditions it was decided to collect versions with the best possible quality and/or as close as possible to the original (e.g. service versions with borders removed were accepted but not screen copies). Wherever available, existing OCR results would also be provided.

Essential metadata was collected to allow indexing and searching in the repository, while keeping the effort of converting and mapping potentially complex metadata records to a minimum. Mandatory details were among others: title, primary language, primary script, original source (e.g. microfilm), and publication date. Optionally, information on typeface, scanner model, image artefacts, and comments on the quality of the document could be provided.

The collected resources were examined by the authors and ingested into the repository database system. This involved allocating project-unique IDs for each image, conversion to a common standard image format with lossless compression, generating viewing copies (with lossy compression for smaller file size), parsing of image characteristics and metadata from file headers and the metadata spreadsheet, and linking of attached files (e.g. existing OCR results).

B. *Subselection of Final Dataset*

The selection of a smaller subset to form the final dataset was driven by two major constraints:

1. To narrow the initial selections further down so as to be in line with the available resources (budget).
2. To maintain the representativeness of the individual datasets as far as possible.

Consequently, it was decided to limit the number of images to 50 per institution. In total 600 images were collected, of which some had to be removed for copyright reasons, resulting in 528 in the published dataset. This presents a sufficiently large dataset, considering the greater quantity of information per page, in comparison to books for instance. One page contains over 383 text lines on average, for example.

TABLE I. BREAKDOWN PER PRIMARY LANGUAGE

Language	# Pages	Language	# Pages
Dutch	19	Polish	37
English	50	Russian	5
Estonian	50	Serbian	50
Finnish	31	Swedish	19
French	50	Ukrainian	4
German	169	Yiddish	3
Latvian	41		

TABLE II. BREAKDOWN PER PUBLICATION DATE

Publication Period		Number of Pages
17th century		6
18th century		12
19th century		187
20th century	1900-1925	155
	1926-1950	168

With regard to representativeness, the distribution of languages, scripts, title pages, middle pages, characteristic layouts and time periods was maintained as close to the original selection as possible. Table I illustrates the distribution of languages, while Table II provides a breakdown of the dataset per publication date. It should be noted that, to be able to evaluate realistic digitisation workflow scenarios it was also ensured that a number of full newspaper issues were included in the dataset. All page images in the dataset are either 300dpi or 400dpi and there is a broad distribution of grayscale, bitonal and colour pages (see Fig. 2). Irrespective of the original image source files, all images in the dataset are stored as TIFF files (with lossless compression). Finally, there is a wide range of page characteristics and image artefacts recorded, the most frequent of which can be seen in Fig. 3.

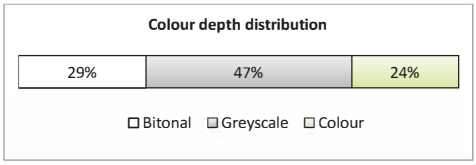

Fig. 2. Distribution of images by colour depth

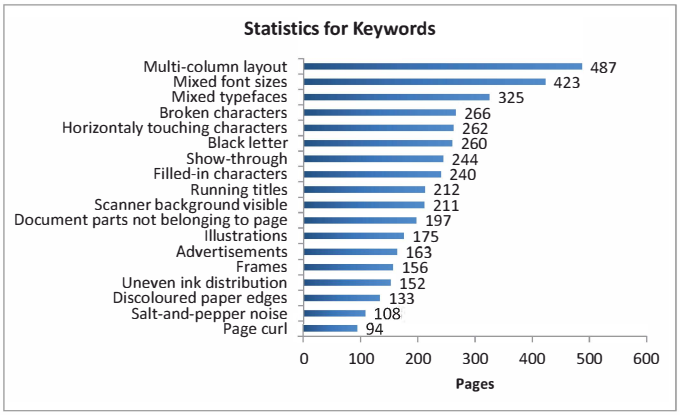

Fig. 3. Most frequent page charactersitics and image artefacts.

C. Ground Truth Production

Ground truth can be described as the ideal outcome of the perfect OCR workflow. As such it is crucial to have for evaluating the output of document analysis methods to what would have been considered the correct result. The creation of ground truth is typically a manual or (at best) semi-automated task due to the fact that current OCR engines are still far from being perfect (especially for historical documents).

Ground truth production was outsourced to commercial service providers in order to speed up the process and to obtain uniform results of high quality (aiming at 99.95% accuracy).

The specified ground truth to be created comprised:

- Precise region outlines.
- Region type labels.
- Full text (Unicode encoded, including special characters such as symbols and ligatures).
- Reading order.

All ground truth files were created in the established PAGE (Page Analysis and Ground Truth Elements) format [5] as recommended by the IMPACT Centre of Competence in Digitisation [6] and used in other large-scale EU projects.

In order to enhance productivity, the service providers were provided with preliminary-processed OCR output files in PAGE (produced by the authors, using a PAGE exporter tool based on the ABBYY FineReader Engine 10) which they could either correct or, depending on the quality of the material, discard and create all ground truth manually. They were also provided with a customised version of Aletheia [7] (a widely-used semi-automated ground truth production system, developed by the authors, see Fig. 4) and detailed instructions on how to interpret and represent specific content elements.

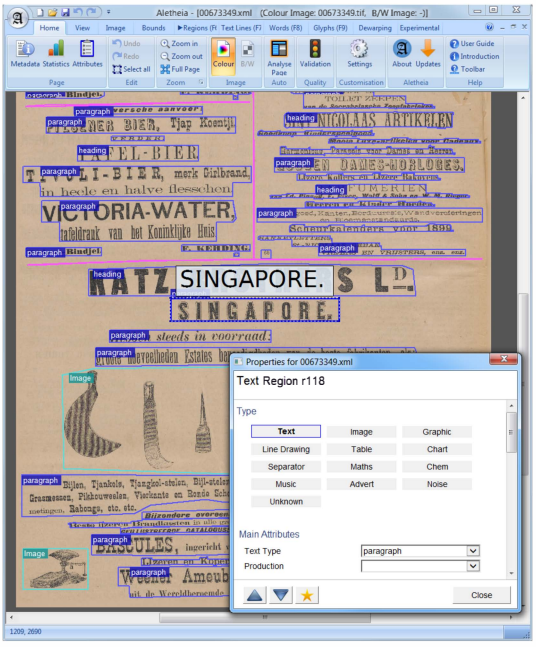

Fig. 4. Aletheia ground truth editor and OCR result viewer

For quality assurance, a three-stage process was established. The first stage took place at the service providers' end by applying the *ground truth validator* (a system implemented and provided by the authors). It performs automated checks against all ground truth rules that can be verified programmatically (e.g. do all regions contain text, are all regions included in the reading order, are there overlapping region outlines etc.).

TABLE III. CONTENT OF GROUND TRUTH SET (528 PAGES)

Page Content	Count
Regions (blocks/zones)	61,619
Images/Graphics	1,497
Tables	208
Text Regions	46,889
Text Lines	202,524

Once all automated checks were passed, ground truth files were subjected to manual quality assurance. Layout-related elements (region outlines, type labels, and reading order) were inspected first. If the layout was approved, files were sent to the respective content provider (library) for the actual text to be verified. Minor problems were typically rectified straightaway during quality control whereas more severe deficiencies were referred back to the service provider.

TABLE III. presents the distribution of a variety of page content objects in the ground truth.

III. ONLINE REPOSITORY

In order to fully unlock the potential of this unique dataset of newspaper images with corresponding ground truth it was decided to follow the best practice from previous dataset collection work of the authors [3][4] and to make it accessible through a web-based online repository. This has proven very useful for establishing a common point of reference among all project partners (resources have unique IDs), searching and identifying documents with specific characteristics, building subsets for individual evaluation experiments, and allowing other technical systems (e.g. workflow systems and evaluation tools) to access resources directly. The underlying technology is based on work from the EU-funded projects IMPACT [8] and SUCCEED [9], with enhancements and customisations tailored to the needs of the ENP project.

A. Web User Interface

The web presence of the dataset comprises five sections:

- Introduction – Overview of the content with statistics on the hosted material and its usage.
- Dataset – Entry point for browsing the dataset per contributing institution and specific subsets defined.
- Advanced Search – Entry point for searching by metadata, image characteristics, and attachments.
- Cart – A tool for managing and exporting selections of images and ground truth files.
- Login – Management of user details and password.

The actual content can be accessed on four levels:

- Lists of thumbnails (subset browsing, search results).
- Individual document details (larger preview image, metadata, links to attachments).
- Full resolution preview (JPEG viewing copy).
- Download of original images, ground truth, and other attachments (individually through the document details view or batch download through the cart).

Fig. 5. Interactive online ground truth viewer and editor showing region outlines and reading order.

A detailed search interface enables users to quickly find images of newspaper pages that fulfil certain criteria. Besides a conventional metadata based search there is a number of additional features tailored to support development and evaluation activities. The *Randomiser* feature, for instance, can be used to generate sets containing a specified number of randomly selected images (while still complying with any given search criteria). Searching by image characteristic is another powerful tool that can be used to select images for studying the impact of specific types of artefacts on a given method or full workflow.

Other existing repositories suffer from the fact that there is no simple way for visualising the corresponding ground truth. For the examination of specific challenges, workflow setups, and evaluation results it is however crucial to be able to quickly view the actual ground truth data. To this end, an interactive ground truth viewer/editor (developed by the authors; see Fig. 5) has been embedded in the online repository, which renders ground truth files in PAGE format to be displayed directly in the browser.

B. Web Services

The resources of the dataset can be accessed directly through web service interfaces. This functionality can be used to integrate the hosted images and ground truth (or other attachments) in external applications. A direct access API (application programming interface) has been developed to support completely independent requests for each resource.

Three main operations are currently supported via this method:

- Check that a document exists.
- Retrieve a document image (thumbnail, viewing copy, or original).
- Retrieve an attachment of a document image (e.g. PAGE ground truth).

Every attempt for accessing a specific resource requires authentication, which is crucial for checking whether the specific user is authorised to access the requested resource. The authorisation is based on a comprehensive permissions management system which allows granting and revoking permissions to specific users on defined sets of images. This is to enforce various different copyright scenarios that might require certain material becoming available under specific agreements (specified by the contributing institutions).

IV. Evaluation of State-of-the-art OCR

Having established such a representative baseline dataset, the natural next valuable step is to evaluate state-of-the-art OCR workflows against it. ABBYY FineReader 11 was chosen as the OCR system to be used in the Europeana Newspapers production workflow for numerous technical reasons. Being a commercial product, however, it might not always be a viable choice for some users. In order to explore also other options, a comparison with Tesseract (version 3.03), an open source OCR engine, was carried out.

A. Text-Based Performance Evaluation

While word accuracy [10] is a good measure for documents with a simple (linear) text structure, it becomes unreliable for documents with complex layouts and non-linear reading order (such as newspapers) due to ambiguities when serialising the text (that evaluation method attempts to match ground truth words to words in the OCR result). Completely independent newspaper articles, for instance, may occur in varying order. There is no single correct solution. To circumvent this problem it appears appropriate to carry out a Bag of Words analysis, which disregards the particular order of words.

Figure 6 shows the performance of FineReader in recognising text in three different font situations (Gothic, Normal and Mixed) as they were used in the production workflow.

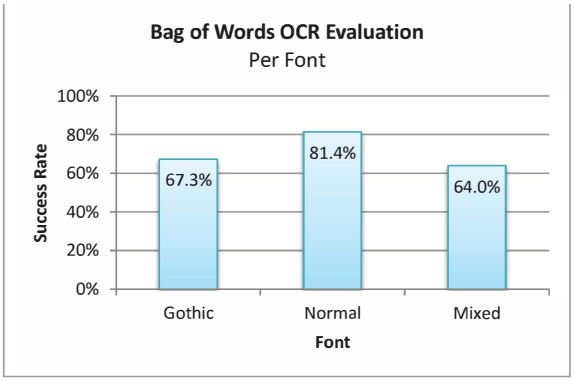

Fig. 6. Bag of Words evaluation for FineReader results – per font type.

As expected, normal (Antiqua) fonts are recognised best. This can be explained as a result of commercial OCR products traditionally focusing on modern business documents. It should be noted that the performance on the very difficult Gothic text is comparatively very good, due to ABBYY FineReader having a new Fraktur module (as a result of the EU-funded project IMPACT). What used to be near random results for such documents is now close to 70% which is considered by many the threshold for meaningful full text search. Documents with mixed content (which basically requires the OCR engine to apply all classifiers and then to decide which result to use) are naturally harder to recognise.

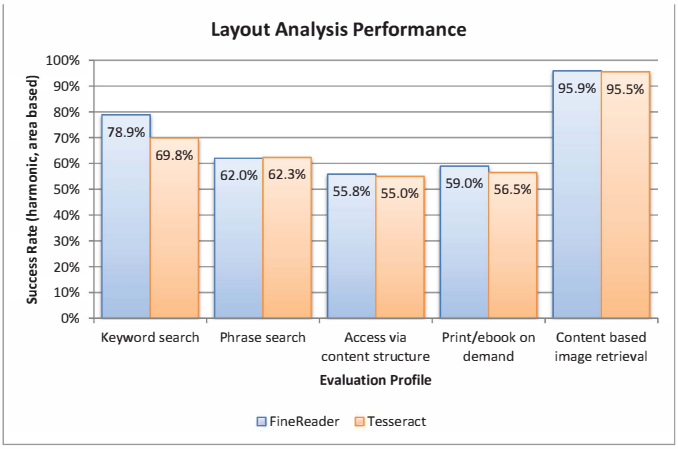

Fig. 7. Performance evaluation results for ABBYY FineReader Engine and Tesseract OCR (different use scenarios).

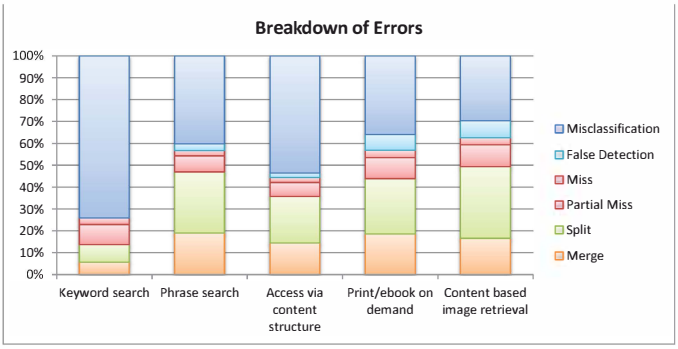

Fig. 8. Breakdown of FineReader layout analysis (weighted) errors for different scenarios.

B. *Scenario-Based Layout Analysis Performance Evaluation*

In addition to textual output, it is important to evaluate the results of OCR workflows in terms of layout and logical structure. This comprises segmentation (location and shape of distinct regions on the page), region classification (type of the content of regions), and reading order.

The measures to be used and how much impact they should have on the overall result is specified in evaluation profiles. These include weights for segmentation errors (merge, split, miss, and false detection), misclassification errors, and reading order errors. Depending on the profile, the overall success rate for an OCR result can vary significantly [11].

The motivation of scenario-based evaluation comes from the observation that abstract error metrics must be put in context of the intended use in order to obtain meaningful scores. Typical examples highlighting this are *keyword search* and *phrase search in full text*. While both rely on text recognition results to be of sufficient quality, phrase search has far greater requirements on the layout being recognised correctly as well. For instance, if two columns on a newspaper page were erroneously merged, the individual words would still be accessible for keyword search but phrase search would fail on any portions of text lines that now span the two merged columns.

The ENP project identified five use scenarios as particularly significant, according to content holders and users:

- Keyword search in full text.
- Phrase search in full text.
- Access via content structure.
- Print/eBook on demand.
- Content-based image retrieval.

Each scenario defines an evaluation strategy that includes settings and weights, which are then to be applied to the evaluation metrics.

Figure 7 shows the evaluation results for FineReader and Tesseract. The success rates vary considerably for the different scenarios, but even more interesting is that there is not a clear winner, suggesting that there are use cases where Tesseract outperforms the commercial FineReader. Figure 8 provides an example breakdown of the different (weighted) error types that are measured by the performance analysis system.

V. Concluding Remarks

The dataset presented in this paper is a valuable resource for researchers and digitisation initiatives. Its usability and the wide and representative range of content, as well as the detailed ground truth underline its uniqueness.

Reflecting on the creation of the dataset, it required several hundreds of person-hours from a variety of people in different organisations to select, pre-process, manually correct (this is the most easily quantifiable cost: €15,000), verify, ingest and categorise the page images and ground truth content. In addition it required strict and detailed specifications and rigorous quality control in a distributed setup. The cost is part of the story though. A significant aspect that made this effort unique was the collaboration of the large number of national and other main European libraries in making this dataset both representative and free to access.

An illustration of the usefulness of the dataset was given by using it to evaluate state-of-the-art page analysis systems, providing valuable insights. The dataset is available free-of-charge for researchers: *www.primaresearch.org/datasets*

References

[1] The European Library, http://www.theeuropeanlibrary.org

[2] Europeana, http://www.europeana.eu

[3] A. Antonacopoulos, D. Bridson, C. Papadopoulos, S. Pletschacher, "A Realistic Dataset for Performance Evaluation of Document Layout Analysis", *Proc. ICDAR2009*, Barcelona, Spain, pp. 296-300.

[4] C. Papadopoulos, S. Pletschacher, C. Clausner, A. Antonacopoulos, "The IMPACT Dataset of Historical Document Images", *Proc. 2nd Int. Workshop on Historical Document Imaging and Processing (HIP2013)*, Washington DC, USA, August 2013, pp. 123-130.

[5] S. Pletschacher and A. Antonacopoulos, "The PAGE (Page Analysis and Ground-Truth Elements) Format Framework", *Proc. ICPR2008*, Istanbul, Turkey, August 2010, pp. 257-260.

[6] IMPACT Centre of Competence in Digitisation, http://www.digitisation.eu

[7] C. Clausner, S. Pletschacher and A. Antonacopoulos, "Aletheia - An Advanced Document Layout and Text Ground-Truthing System for Production Environments", *Proc. ICDAR2011*, Beijing, China, September 2011, pp. 48-52.

[8] IMPACT (Improving Access to Text): http://www.impact-project.eu.

[9] SUCCEED (Support Action Centre of Competence in Digitisation): http://succeed-project.eu.

[10] S. V. Rice, "Measuring the Accuracy of Page-Reading Systems", Doctoral Dissertation, University of Nevada, Las Vegas Las Vegas, NV, USA 1996, ISBN:0-591-26287-8.

[11] C. Clausner, S. Pletschacher, A. Antonacopoulos, "Scenario Driven In-Depth Performance Evaluation of Document Layout Analysis Methods", *Proc. ICDAR2011*, Beijing, China, September 2011, pp. 1404-1408.