The IMPACT Dataset of Historical Document Images†

Christos Papadopoulos, Stefan Pletschacher, Christian Clausner 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

{C.Papadopoulos, S.Pletschacher, C.Clausner, A.Antonacopoulos}@primaresearch.org

ABSTRACT

Representative and comprehensive datasets are a prerequisite for any research activity, from studying specific types of problems through training of algorithms to evaluating results of actual implementations. This paper describes an invaluable resource which is the result of a large scale effort undertaken in the EU funded project IMPACT. A number of challenges faced during the creation phase but also the significant benefits and potential of this collection of printed historical documents are described. The dataset contains over 600,000 document images that originate from major European libraries and are representative of both their respective holdings and digitisation plans for the near to medium term. It is truly unique with regard to the very substantial amount of high-quality ground truth which is available for approximately 45,000 pages, capturing detailed layout, reading order and text content. The dataset is publicly available through the IMPACT Centre of Competence (www.digitisation.eu).

Categories and Subject Descriptors

H.2.8 [**Database Applications**] Image databases

I.7.5 [**Document and Text Processing**] Document Capture–Graphics recognition and interpretation.

General Terms

Management, Standardization.

Keywords

Dataset production, Ground truth production, Historical documents

1. INTRODUCTION

In order for any research and development effort to be successful it is crucial to thoroughly analyse the underlying problems and to specify objective measures for evaluating the effectiveness of potential approaches and eventually the quality of the delivered results. IMPACT (Improving Access to Text) was a large scale project funded by the European Commission aiming at significantly improving access to historical text and reducing barriers that stand in the way of mass digitisation of the European cultural heritage [1]. At the beginning of the project, while the overall goals and methodologies to be applied were clearly defined in a comprehensive description of work, there was only vague knowledge, let alone a systematic overview, of what the combined material (with associated digitisation issues) of so many major European libraries looked like and where the greatest impact could be achieved with the given resources. Creating a central dataset was therefore a crucial step towards establishing a common understanding of the material with regard to aspects like relevance for digitisation, frequency, complexity and technical challenges. It was soon realised that a dataset of this dimension, stemming from so many high-profile libraries, would not only satisfy certain requirements for the work in the project but could also be an invaluable resource for research and development beyond IMPACT. The scope of this task was therefore greatly extended, including the creation of ground truth which can only be described as unprecedented in terms of volume, completeness and precision.

The approach taken for building the dataset followed the best practice from previous activities [2] giving particular emphasis for the dataset to be:

- *Realistic* – reflecting the actual library holdings with regard to representativeness and frequency of documents
- *Comprehensive* – including metadata and detailed ground truth
- *Flexibly structured* – supporting all stakeholders to search, browse, group etc. and allowing other technical systems (such as workflow systems and evaluation tools) to interface directly.

The use of the dataset within the project was wide-ranging: From analytical activities (identification of particular types of image artefacts, scripts, languages and/or spelling variations) through research and development tasks (in-depth analysis of specific example images and training of algorithms) to performance evaluation of intermediate results as well as of the final deliverables. Due to its broad and rich content it is expected that the IMPACT image and ground truth dataset will continue to be the basis for cutting-edge research related to digitisation and OCR (Optical Character Recognition).

It should be noted that, apart from digital repositories maintained by the libraries themselves, there is a number of other collections and datasets which may be of interest for development and performance evaluation in this field. The Internet Archive [3] for instance is a rich source of scanned books and OCRed text. The quality of the available texts, however, is not always sufficient for more sophisticated research tasks and certainly not adequate for

HIP '13 August 24 2013, Washington, DC, USA

http://dx.doi.org/10.1145/2501115.2501130

† This work was funded by the EU 7th Framework Programme grants IMPACT (Ref: 215064) and SUCCEED (Ref. 600555).

evaluation purposes. Project Gutenberg [4] offers over 42,000 free eBooks which were thoroughly proofread by volunteers. While the text quality of this collection is generally good, there is no information about the correspondence between the encoded text and the scanned page. The RETAS OCR Evaluation Dataset [5] is an attempt to overcome this problem by aligning text from Project Gutenberg with page images from the Internet Archive. The combination of these two sources leads to the increased quality of text ground truth but is still lacking layout information. The PRImA Layout Analysis Dataset [2], on the other hand, is mainly devoted to layout elements with very precise region outlines but focuses on contemporary documents and does not contain text. The UvA Dataset [6] is similar in the sense that it focuses on layout and structure without including any text content. Various datasets are hosted at the University of Washington [7], among others the Pacific Northwest Historical Documents Database which ranges from photographs to handwritten and printed documents. While all of the above datasets are very valuable resources, none of them were created with the goal of being representative of the actual holdings of major libraries (and therefore not truly representative of current digitisation challenges).

In the following sections the creation of the IMPACT dataset is described, its contents, how it is hosted, how it was used and how it can be of use for those interested.

2. BUILDING THE DATASET

The two core activities for building up the content of the dataset were selection and delivery of images and metadata (individually by each library partner), followed by ground truth production for specific subsets (centrally for the whole project).

2.1 Image Aggregation

The image aggregation process for creating a dataset of such size had to be very clearly and strictly defined and followed. The first step was *image selection*, where the content providers identified representative titles from within their holdings. This was done taking into consideration what material had already been digitised, in order to expedite the process of generating the dataset by minimising the need to digitise new material. It was however clear that even though the selection would be mostly from already digitised material, it should be representative of the complete holdings and the future digitisation plans of a given library, not just what had been already digitised for whatever reason (e.g. a specific bequest). This, in most cases, meant that libraries could not rely on random image selections, but had to devote significant time and effort to manually select which titles, issues and pages to include.

The image selection was followed by *metadata collection*. To standardise this process, a dedicated spreadsheet with fields for the required metadata was filled in by the libraries. This basic metadata formed the basis for indexing the images on the online repository. The specification of metadata required at this stage was compiled using input from libraries, technical partners and language specialists (see below).

Once all images were selected and retrieved from a library's storage facilities and all required metadata collected, both images and metadata were uploaded to the dataset server, where *validation* and *pre-processing* took place. First, all images were checked for validity — it was not uncommon for libraries to upload service (viewing) copies of images instead of higher quality master files. In such cases, the libraries were contacted to either re-upload higher quality images or to adjust their image selection by providing alternative images. The second step of the verification process involved the metadata. It was ensured, for all images, that metadata was present and complete (taking into account any alterations during the first step of this verification process).

Once all available data was of satisfactory quality, the *pre-processing* of the images could be carried out in order to be ingested into the dataset. This step involved allocating a unique ID to each image and conversion to standard TIFF format with LZW compression. The source images were provided in varying formats, such as uncompressed TIFF, JPEG, JPEG2000. It was therefore necessary to standardise to an open and easy-to-use format so that images in the dataset could be easily used by a wide range of different tools. As part of the pre-processing of the images, also a number of viewing copies (in lower quality JPEG) and thumbnails were generated for more efficient transmission and display over the web interface of the dataset.

Following the pre-processing, the *image characteristics* were retrieved from the files themselves. Image headers were parsed for size (in pixels), resolution, colour depth and scanner information (where available). Then, a *secondary verification* related to image size was performed since in some cases the image resolution information was incorrect, leading to misleading image sizes. Given an image width of 2,000 pixels, for example, the differences in resolution could vary the actual width of the page considerably from 17cm (at 300 dpi) to 70 cm (at 72 dpi). This was a very common issue, especially for scans that where obtained from microfilm as an intermediate step without adjusting the resolution in relation to the (paper) original. Other checks performed at this stage included detecting inverted images (again common for images scanned from microfilm) and removing the alpha layer (present in many colour scans).

Several libraries provided a variety of additional files along with the images, containing supplementary metadata and layout information in different formats, text ground truth and raw OCR results. Any such files were at this point identified and logically linked to their corresponding images. This step was performed manually (a relationship had to be inferred from the filename or folder structure).

Once all validations were complete and all additional files processed, the final stage was to commit images and metadata to the repository and make them available to all authorised users of the system.

2.2 Ground Truth Production

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. Consequently, it has to be created (or at least verified) by a human. The format of the ground truth varies thereby depending on the task it is related to (such as region outlines for segmentation or Unicode text for OCR).

Ground truth is a crucial asset not only for developing and training new methods (such as adaptive text line segmentation or lexicon-based post correction) but also for performance evaluation as well as tuning of end-to-end digitisation workflows (ranging from image enhancement, segmentation, region classification, reading order detection to the actual text recognition and post correction) and/or any of those individual intermediate steps.

In order to keep this manual task manageable and at the same time the selection representative, the number of document pages to be ground truthed was fixed to several thousand per partner (taking

into account the size of their collection). While the creation of high quality ground truth is a challenging task at any rate, doing this for tens of thousands of pages is unparalleled.

2.2.1 Production Process

Given the immensity and complexity of the historical material, in-house production by the libraries was prohibitive in terms of both cost and training requirements. Accordingly, preparations were made in order to outsource the actual production task to a service provider. In fact, several service providers had to be commissioned in parallel in order to handle the sheer amount as well as the particularities of different languages, scripts etc.

Ground-truthing guidelines and a detailed specification of requirements were established jointly by technical partners and libraries in order to ensure uniformity and quality of the final deliverable. Based on the valuable feedback from the service providers it soon became apparent that the guidelines had to be complemented with an illustrated FAQ (Frequently Asked Questions) so as to deal with the growing number of captured special cases. Due to the nature of the historical material there was a constant stream of unprecedented cases, requests for clarification, discussions by the technical work group, involvement of external experts like historians or linguists in case of very specific questions and updates to the guidelines and the FAQ. Several mechanisms (like splitting of batches and "pending characters", see below) had to be put in place in order for the production not to be stalled by this clarification process.

The semi-automated ground truth production tool Aletheia [8] and a comprehensive documentation of the software were made available to all service providers and appropriate training was arranged via video conferencing. Initial tests showed that making an appropriate choice between different production approaches can have a huge impact on the overall efficiency, depending on the quality of the material. The first approach was to start from scratch (i.e. the image only), drawing first region outlines then labelling regions, annotating reading order and finally entering the actual text. The second approach was based on starting from a decent quality OCR result (segmented lines and text output) and then checking and/or correcting potential errors. While this second approach worked well for "newer" material it was not an option for some of the very old or low quality documents. All service providers were supplied with pre-produced material (images plus OCR results by ABBYY FineReader Engine 9) but the choice of which approach to apply was left with the individual ground-truther on a case-by-case basis.

2.2.2 Quality Assurance

Several levels of quality assurance (QA) were put in place in order to ensure the best possible quality.

The first level of QA was established at the service providers' end, requiring any material, before delivery, to be scanned for violation of the guidelines. For this purpose, an automatic validation module was implemented and integrated in Aletheia, which was used to check against all rules that can be verified programmatically (such as region outlines not overlapping, text entered for all regions, all relevant regions included in reading order etc.)

The second level of QA was done by project partners and involved double checking of the above guidelines for whole batches (using a command line version of the aforementioned validation module) as well as manual checking of the actual content. Layout and reading order were handled by technical partners whereas the accuracy of the text content was checked by the respective libraries.

It should be noted that although ground truth is generally defined as a perfect representation, it has to be acknowledged that mistakes in the manual production process, for text in particular, cannot be entirely eliminated. The target word accuracy for text content was therefore set to 99.95% or better, which is realistically achievable and yet precise enough for the intended usage of the ground truth.

Whenever ground truth files failed any of the above checks this was recorded and the material referred back to the service provider for correction. In some difficult cases this took several iterations of checking and correcting.

2.2.3 Content

The following three aspects of ground truth are useful for a wide range of methods and processing steps in a digitisation pipeline and were chosen for the actual ground truth production: *layout*, *reading order*, and *encoded text*.

Layout, in this case, comprises region outlines in the form of precise polygons (not only bounding boxes) and labels for the overall type of each printed region and, where applicable, subtype (such as a *text region* with subtype *heading*). The primary rules for marking the outlines of regions were to fully contain all connected components (i.e. black pixels of the corresponding black-and-white image) of a coherent region, not to include any connected components belonging to another region and to avoid overlap between regions (which would not have been possible with simple bounding boxes). In general, the distinction between coherent regions was to be made based on changes in layout characteristics from one to the other (for instance changes in number of columns, indentation, font type of whole lines etc.).

The *reading order* was captured using the novel feature of ordered and unordered groups, introduced with the PAGE format (see also below), and allows for a more suitable representation of elements for which a definite sequential order is not appropriate (such as adverts in newspapers or articles with no particular order). For practical reasons it was decided only to include textual regions in the reading order and moreover, to exclude certain text subtypes which are not relevant for reconstructing the actual content (for instance page numbers).

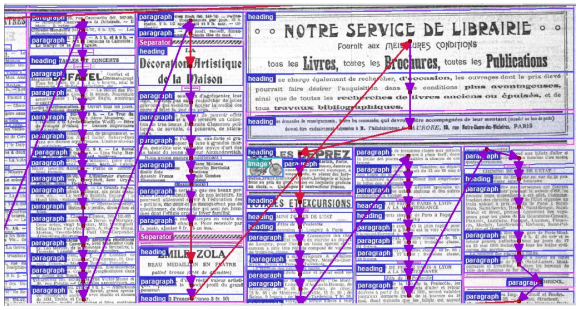

Figure 1. Reading order example

Entering *text* might appear as a straight forward task but, due to the magnitude of this venture and the very nature of historical material, required considerable effort and coordination in order to obtain uniform and accurate results. The consistent use of Unicode was crucial for dealing with multilingual and multi-script text as well as historical and/or unprecedented characters. Numerous decisions had to be made in order to standardise the representation of special characters, ligatures, historical versions of characters, illegible text etc. The general approach followed

was thereby to capture what is on the page as precisely as possible. For the above issues this meant researching the correct Unicode entry for any special character, using the Unicode code points for ligatures rather than (two or more) individual characters, no transcription of any sort (for instance the historical long s was to be entered as such and not to be replaced by the modern s) and marking illegible characters with a replacement character. In cases where previously unseen characters were discovered they had to be temporarily entered as application specifically defined (Unicode private use areas) "pending characters" in order for the overall production process not to be stalled. After clarification by the ground truth working group or in difficult cases by external experts those instances were revisited and corrected. In a number of cases no corresponding Unicode entries could be found. The procedure for dealing with such characters involved then a cross-check with other recommendations like those of the Medieval Unicode Font Initiative (MUFI) [9], where applicable using their recommended code point, or eventually defining an IMPACT-specific code point from within a Unicode private use area not interfering with the aforementioned MUFI recommendations. For special characters defined in IMPACT a specific font was maintained and provided to service providers with regular Aletheia updates to enable convenient editing and correct rendering of the entered text.

For all ground-truthed images in the dataset, the minimum content is the above three aspects. For a smaller number of pages and for some more specific applications, additional types of ground truth were created in addition to the above (examples include: print space, page border, text lines, words and glyphs).

3. DATASET DESCRIPTION

This section provides a description of the dataset content in terms of the images and metadata available, the amount and level of ground truth associated with those images as well as the relationships and links between the images.

3.1 Images

The dataset contains a wide variety of documents reflecting both the holdings of major European libraries, but also their digitisation plans for the near and medium term (5 to 10years). The dataset comprises over 600,000 images originating from 10 different national and prominent libraries across Europe. As a guideline, each library was asked to provide approximately 50,000 representative images. However, each library was free to vary that amount as long as they were satisfied they provided a sufficient number of page images that represented its holdings and priorities. A list of all the content providers along with numbers of contributed images is presented in Table 1.

Table 1. Dataset breakdown by library

Library	Country	Number
British Library	UK	48,515
Biblioteca Nacional de España	Spain	60,180
Bibliothèque Nationale de France	France	96,950
Bayerische Staatsbibliothek	Germany	66,627
Koninklijke Bibliotheek	Netherlands	88,192
Národní Knihovna České Republiky	Czech Republic	75,559
St. Cyril and Methodius National Library	Bulgaria	4,240
Narodna in Univerzitetna Knjižnica	Slovenia	41,313
Österreichische Nationalbibliothek	Austria	110,034
Poznań Supercomputing and Networking Centre	Poland	11,020
Total images		**602,630**

A breakdown of the types of documents is provided in Table 2. Even though the contents of the dataset cover a wide variety of document types, it is mostly focused on books and newspapers.

Table 2. Dataset breakdown by document type

Type	Number of documents
Book Page	335,640
Newspaper Page	142,748
Legal Document Page	80,289
Journal Page	19,573
Other Document Page	18,957
Unclassified Page	5,423
Total Pages	**602,630**

In terms of age, the majority of the documents in the dataset – almost 80% – were produced in the 19th or early 20th century (especially in the case of newspapers). Further 17% originate from the 17th or 18th century, with the rest of the images ranging as far back as the 15th century. A detailed breakdown of the age of the original documents in combination with the document types is presented in Table 3. There is a very small percentage of images where the publication year is not known, or was not made available to us (marked as "?" in the table below). Although all possible steps were taken to avoid such issues, in a dataset of this size it is inevitable that some metadata will be incomplete (also unknown to the libraries).

Table 3. Document type and production century distribution

Century	Book Page	Newspaper Page	Legal Document Page	Journal Page	Other Document Page	Unclassified Page	TOTAL
15	338	0	0	0	1	0	**339**
16	17,384	119	0	0	5	0	**17,508**
17	58,633	1,119	0	0	8	280	**60,040**
18	39,139	2,707	1,297	0	144	132	**43,419**
19	194,682	64,642	29,821	10,216	13,556	4,061	**316,978**
20	23,038	73,745	49,171	9,357	5,243	234	**160,788**
?	2,426	416	0	0	0	716	**3,558**

Since the documents in the dataset originate from different libraries across Europe, a large number of languages and scripts are included. While there is detailed language and script metadata for a large proportion of the dataset, such information is not available for a number of images (in such cases, an assumption is made for the data to be consistent with the default of the particular library). Confirmed are a total of 18 languages represented in 10 different scripts (see Table 4 and Table 5). Figure 2 contains sample book and newspaper pages.

Table 4. Listing of languages of the dataset

Bulgarian	German	Polish
Catalan	Greek	Portuguese
Czech	Hebrew	Russian
Dutch	Latin	Slovenian
English	Norwegian	Spanish
French	Old Church Slavonic	

Table 5. Listing of scripts of the dataset

Bohoričica	Hebrew
Cyrillic	Latin
French	Latin/Gothic
Gaj	Old Cyrillic
Greek	Serif

3.2 Metadata

In order to enable users to efficiently search the repository, a large set of metadata is stored as part of the dataset. All available metadata is indexed and can be used as search parameters to access specific images or sets of images within the overall dataset.

The metadata has been organised under the following structure:

- *Bibliographic information* — title, author, publication date and location, document type, page number,
- *Digitisation information* — resolution, bit depth, image dimensions, scanner used, file type, compression algorithm and quality, source of digitisation (paper, microfilm, etc),
- *Physical characteristics* — language, script, typeface, number of columns,
- *Copyright information* — copyright holder, contact details, publishing permissions, content provider's reference,
- *Administrative information* — original filename, access log,
- *Comments* — comments for any information not captured by the above metadata (on physical appearance, paper quality, binding, typefaces used and physical layout.

In addition to image-related metadata, part of the dataset has a number of additional keywords associated with each image. That list of keywords was compiled during the IMPACT project with input from both libraries and technical partners in a way that it provides additional useful search possibilities for different users of the dataset. The keywords are grouped into three logical sets as outlined below:

- *Condition related* — stains, holes, missing parts, tears, folds, etc.
- *Document related* — impressions, filled in characters, broken characters, blurred, faded, etc.
- *Scanning related* — skew, parts of adjoining page, fingers, paper clips, copyright notices, etc.

3.3 Ground Truth

As mentioned earlier, one of the crucial advantages of this dataset is the substantial amount of detailed and high quality ground truth available and the scope for its use. There are over 45,000 images that have been ground-truthed down to region outlines, region text content and reading order (see Section 2.2.3). Table 6 presents an overview of the variety and number of regions in the dataset by type and subtype. different Furthermore, for another 300 images there is additional detailed ground truth down to text lines and words (outlines and text content) – totalling about 70,000 word outlines and corresponding encoded text. This detailed ground truth can be used for more focussed research.

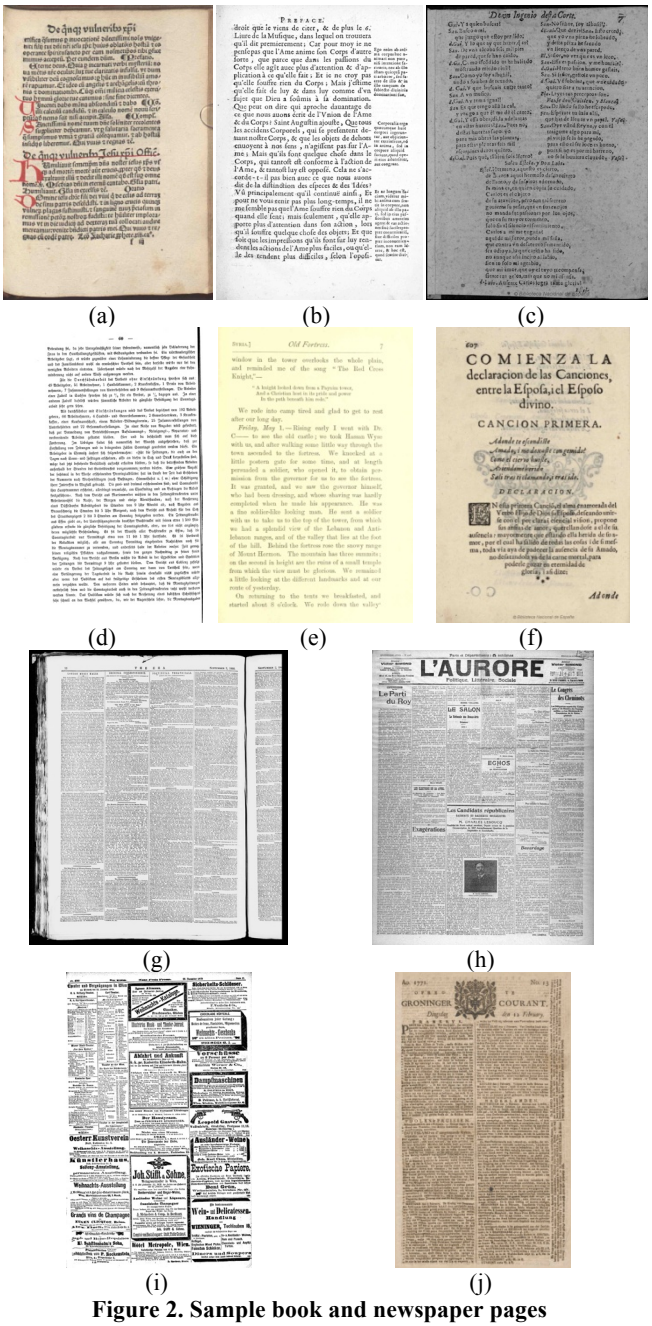

Figure 2. Sample book and newspaper pages

(a)-(f): Books, (g)-(j): Newspapers

The ground truth is stored in the XML format which is part of the PAGE (Page Analysis and Ground truth Elements) representation framework [10]. 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. A sample image with ground truth description can be seen in Figure 3.

Table 6. Total ground-truthed regions per type and subtype

Region type/subtype	Number
Text	**573,725**
Heading	42,345
Paragraph	388,636
Drop capital	6,211
Caption	294
Header	35,023
Footer	409
Footnote	2,897
Footnote continued	187
Signature mark	10,642
Catch word	20,678
TOC-entry	6,217
Page number	37,727
Marginalia	11,091
Credit	11,307
Graphic	**10,151**
Logo	4
Stamp	937
Handwritten annotation	2,343
Punch hole	419
Signature	15
Other	6,135
Image	**1,312**
Line Drawing	**8**
Separator	**30,998**
Table	**1,558**
Chart	**5**
Maths	**355**

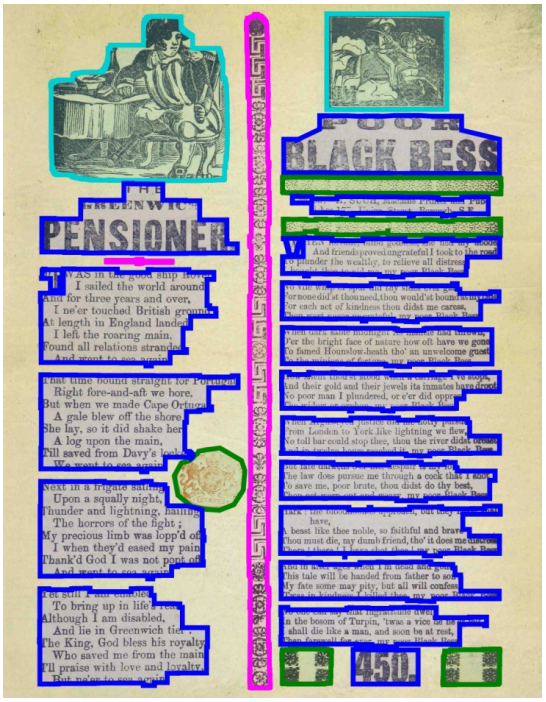

Figure 3. Sample ground truth page, showing region outlines (blue: text, magenta: separator, cyan: image, green: graphic)

3.4 Relationships

In order to index the dataset in a more usable way, a number of relationship models have been defined. Each image in the dataset may be logically linked to other related images or files, as explained below.

3.4.1 Related images

Links between different document images constitute highly useful information. For example, since the unit of reference in the dataset is a single page scan, it is very common during the data collection to have to split double page scans, which are common especially in digitisation of microfilmed pages. After splitting the original images into the corresponding left and right pages, both the original and the new images are stored in the dataset and linked to each other. This approach has two benefits:

- It allows to correctly ground-truth the two pages separately, in terms of layout and text,
- It keeps the original image and the resulting split pages linked together so they can be used to evaluate a page splitting algorithm.

Figure 4 shows a screenshot of the web interface, illustrating the linking between an original image and the corresponding left and right pages.

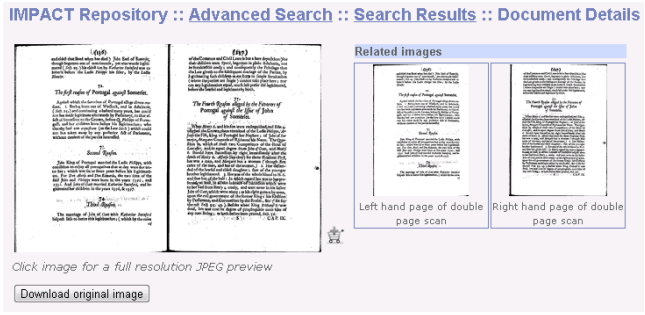

Figure 4. Double page scan linked to left and right pages.

The functionality of linking images from the dataset to each other is utilised to represent several possible links between page images, such as:

- Double page scan with left and right pages (as explained above),
- Clippings of a page that need to be treated as separate images (especially useful in large newspaper pages),
- Rescans of the same image that are linked together but still treated as different images (since they have different metadata at least regarding the scanning parameters),
- Scans of the same book pages from different providers (for example a book that has been digitised by two different content providers),
- Scans of different editions of the same book.

3.4.2 Attachments for images

Another useful facility provided by this dataset is the flexibility to also allow files of various types to be attached to the document images. This has been used to accommodate linking various different types of files to the images available. These include:

- Ground truth files for layout and text content (PAGE),

- Other versions of the same image (binarised, dewarped, with removed page border, etc.), that might be useful as alternative or secondary input for some tools,
- Layout segmentation results (PAGE and other XML)
- OCR Results (PAGE, ALTO, plain text)
- Any other file with ground truth or results for the image (ALTO, METS and plain text files are supported)

Figure 5 illustrates an image that is linked to two different attachments, a generic XML file and a PAGE ground truth file.

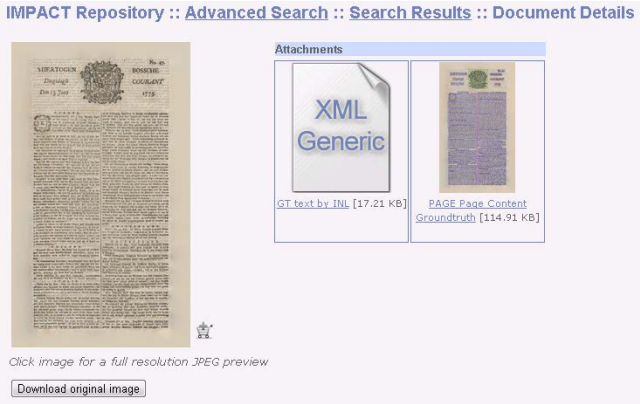

Figure 5. Image with attached files.

3.5 Subsets

A further feature of the dataset is the flexibility in creating and working with subsets. This is especially beneficial for such vast resource (ease of use and access).

Apart from dynamic subsets (search results generated using the web interface) such as images from specific document types, with specific characteristics etc., there are several existing subsets. Among others these cover suggested subsets for training and evaluation tasks and groups of images intended for specific experimental tasks.

4. TECHNICAL INFRASTRUCTURE

This section describes the technical infrastructure (software and hardware) that was required and developed for creating and maintaining the dataset.

4.1 Web Interface

The main means of accessing the dataset is through the dedicated web interface, a custom-built platform designed specifically to access the images and/or ground truth files in an efficient and user-friendly manner.

Options for accessing the images include:

- *Browsing all images* — allowing the user to browse through all available images.
- *Browsing specific subsets* — allowing the user to browse a set of images that have been grouped together already, according to some specific criterion (e.g. typewritten images).
- *Searching through the complete set* — allowing the user to specify a set of search criteria and retrieve only the images that satisfy them.

Searching is the most useful way of accessing the dataset, providing a multitude of search options that can be used individually or in any combination. The available search options are:

- *Search by metadata* — all available types of metadata can be searched for. These include, but are not limited to: publication title, author, publication location and year, language, script, resolution and bit rate.
- *Search by keywords* — any keywords attached to images are fully indexed and searchable. These describe issues and characteristics associated with the images.
- *Search within subsets* — the subsets defined are also fully searchable.
- *Search by attachment* — allowing to restrict search results to images with specific types of attachments only (i.e. PAGE ground truth).
- *Random selection* — giving a random selection of images that satisfy all other search criteria.

Figure 6a shows a screenshot of the search interface available to the users of the dataset. Figure 6b and Figure 6c demonstrate the web interface with search results displayed as thumbnails and the document preview and metadata respectively. Figure 6d shows the interactive ground truth viewer that is available to preview PAGE files (both ground truth files and processing results) without the need to download both the PAGE and corresponding image file and open them offline.

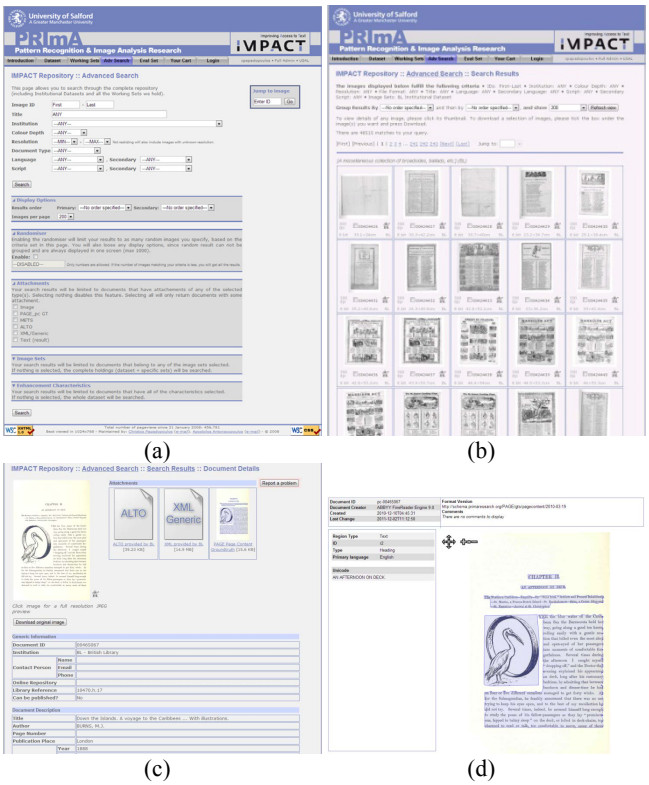

Figure 6. Screenshots from the web interface

(a) Advanced search interface, (b) Search results displayed as thumbnails, (c) Document preview with metadata, (d) Interactive ground truth viewer

4.2 Integration

In addition to the web interface, a number of web services have been developed in order to access images and ground truth files in a secure manner, bypassing the web interface.

These include plain HTTP, secure HTTP and SOAP services that provide authentication and retrieval methods which can be called via third party applications and workflow managers in order to retrieve specific files.

4.3 Hardware and Storage

In order to store and make available a dataset of such size, specialist equipment had to be acquired and set up.

In terms of storage requirements, the dataset is powered by a fully backed-up 26TB disk array, which is used to store over 3,000,000 image and ground truth files. In the background, it is supported by a database which currently is just over 900MB in size and contains in excess of 12,000,000 records.

With regard to traffic, in the past 31 months, over 8,000,000 images or approximately 5TB of image data were served via the web interface and services combined.

5. USE AND AVAILABILITY

The dataset has been used as part of the IMPACT project for development, evaluation and demonstration purposes. Subsets of the images have been used for a variety of purposes by project partners ranging from more technical – such as binarisation, page border detection, dewarping – to more theoretical – such as generation of lexica [11] –. It is worth noting that, with ABBYY as partner in the project, lexica developed via this dataset and training data from the ground truth set have helped to further improve FineReader to better support historical documents. The authors have used the dataset to organise three performance evaluation competitions on historical documents (both books and newspapers) [12] [13] [14].

While parts of the dataset are already available online through the respective libraries, access to the complete set is available via the IMPACT Centre of Competence [15].

6. CONCLUSIONS

Well maintained comprehensive datasets constitute crucial assets for any kind of research and development activities. In this paper, a large-scale effort with numerous challenges but also very substantial benefits was described. The IMPACT image repository is a unique resource which enables and fosters research beyond and well after the end of the project.

The dataset provides a collection of over 600,000 historical document images that originate from Europe's main libraries and are representative of both their respective holdings and digitisation plans for the near and medium term. In addition to the vast collection of page images, approximately 45,000 of them have been enriched with detailed layout, reading order and text ground truth.

7. ACKNOWLEDGMENTS

The authors would like to acknowledge the significant contributions of Günter Mühlberger and Que-Anh Ha from the University of Innsbruck related to managing the production of ground truth as well as all the libraries and other project partners in the IMPACT project for their support.

8. REFERENCES

[1] IMPACT project: http://www.impact-project.eu

[2] 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

[3] Internet Archive – Text Archive: http://archive.org/details/texts

[4] Project Gutenberg: http://www.gutenberg.org/

[5] I. Z. Yalniz, R. Manmatha, "A Fast Alignment Scheme for Automatic OCR Evaluation of Books", *Proc ICDAR2011*, Beijing, China, pp. 754-758

[6] Leon Todoran, Marcel Worring, Arnold Smeulders, "The UvA color document dataset", International Journal of Document Analysis and Recognition (IJDAR), 2005, Vol.7(4), pp.228-240

[7] University of Washington, University Libraries – Datasets: http://www.lib.washington.edu/types/datasets/

[8] 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

[9] Medieval Unicode Font Initiative: http://www.mufi.info/

[10] S. Pletschacher and A. Antonacopoulos, "The PAGE (Page Analysis and Ground-Truth Elements) Format Framework", *Proc. ICPR2008*, Istanbul, Turkey, August 23-26, 2010, IEEE-CS Press, pp. 257-260

[11] Jesse de Does, Katrien Depuydt, " Lexicon-supported OCR of eighteenth century Dutch books: a case study", Proc. SPIE 8658, Document Recognition and Retrieval XX, 86580L (February 4, 2013); doi:10.1117/12.2008423

[12] A. Antonacopoulos, C. Clausner, C. Papadopoulos and S. Pletschacher, "Historical Document Layout Analysis Competition", *Proc. ICDAR2011*, Beijing, China, Septempber 2011, pp. 1516-1520

[13] A. Antonacopoulos, C. Clausner, C. Papadopoulos and S. Pletschacher, "Historical Book Recognition Competition – HBR2013", *Proc. ICDAR2013*, Washington DC, USA, August 2013

[14] A. Antonacopoulos, C. Clausner, C. Papadopoulos and S. Pletschacher, "Historical Newspaper Layout Analysis Competition – HNLA2013", *Proc. ICDAR2013*, Washington DC, USA, August 2013

[15] Impact Centre of Competence: http://www.digitisation.eu/

[16] SUCCEED project: http://succeed-project.eu/