succeed

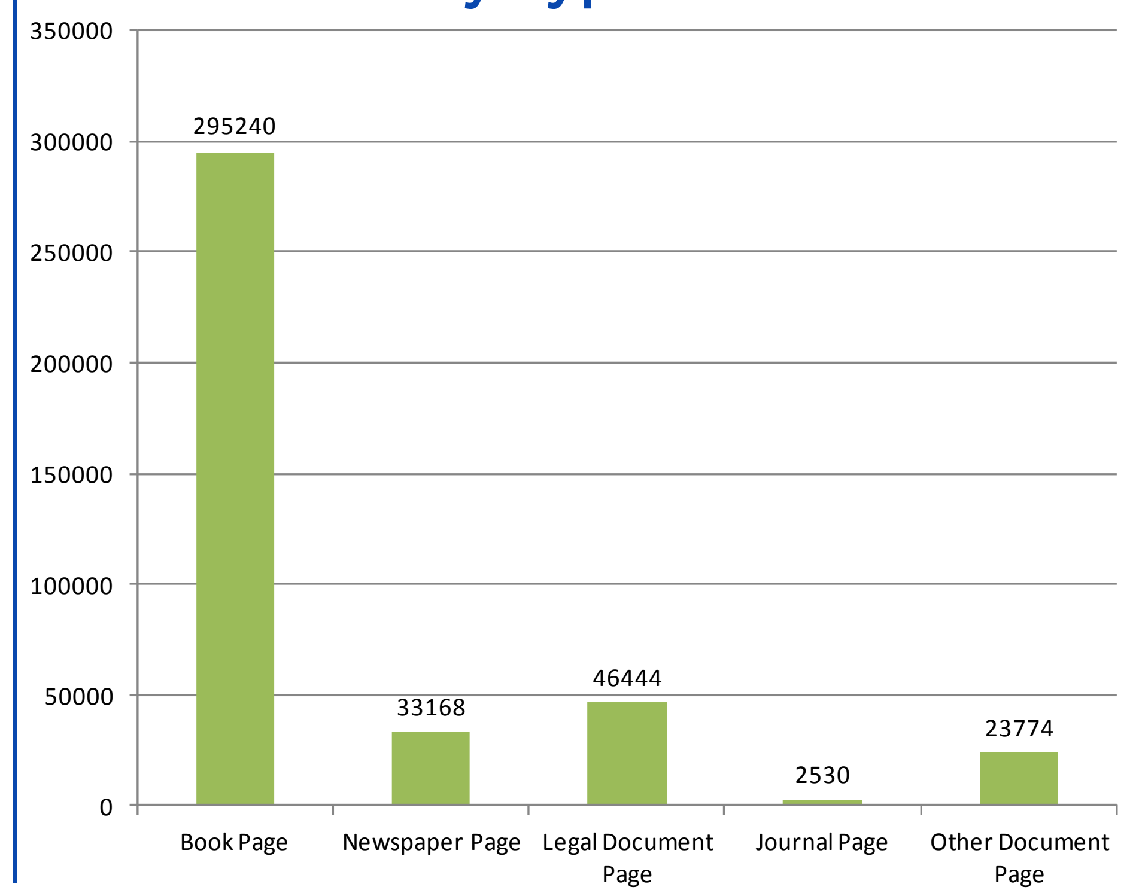
The IMPACT Digitisation Image Repository

Pattern Recognition and Image Analysis (PRImA) Research Lab, School of Computing, Science and Engineering, University of Salford, Greater Manchester, United Kingdom

The Dataset

Part of the IMPACT Centre of Competence in Digitisation (digitisation.eu), it contains more than half a million representative historical text-based images compiled from major European libraries. Covering texts from as early as 1500, and containing material from newspapers, books, pamphlets and typewritten notes, the dataset is an invaluable resource for future research into imaging technology, OCR and language enrichment.

Documents by Type



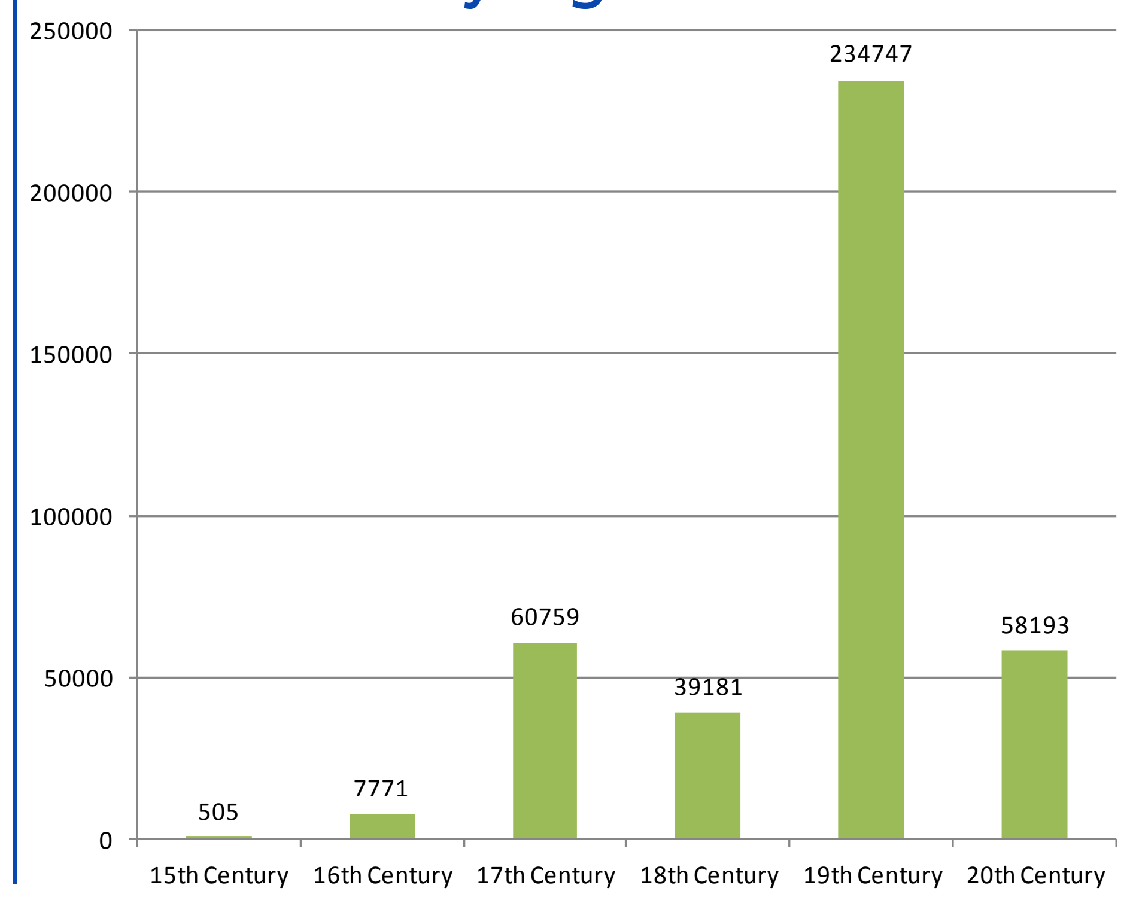
Languages

Français
Nederlands
Čeština
English
Polski
Español
Slovenščina

Access

Access to the dataset is provided via membership of the IMPACT Centre of Competence. Once granted access, users can browse and search the dataset via a web interface, which allows to view detailed document metadata, preview images and ground truth files, and download individual images or collections. It is also possible to access images and/or ground truth files directly, using HTTP web service calls.

Documents by Age



Scripts

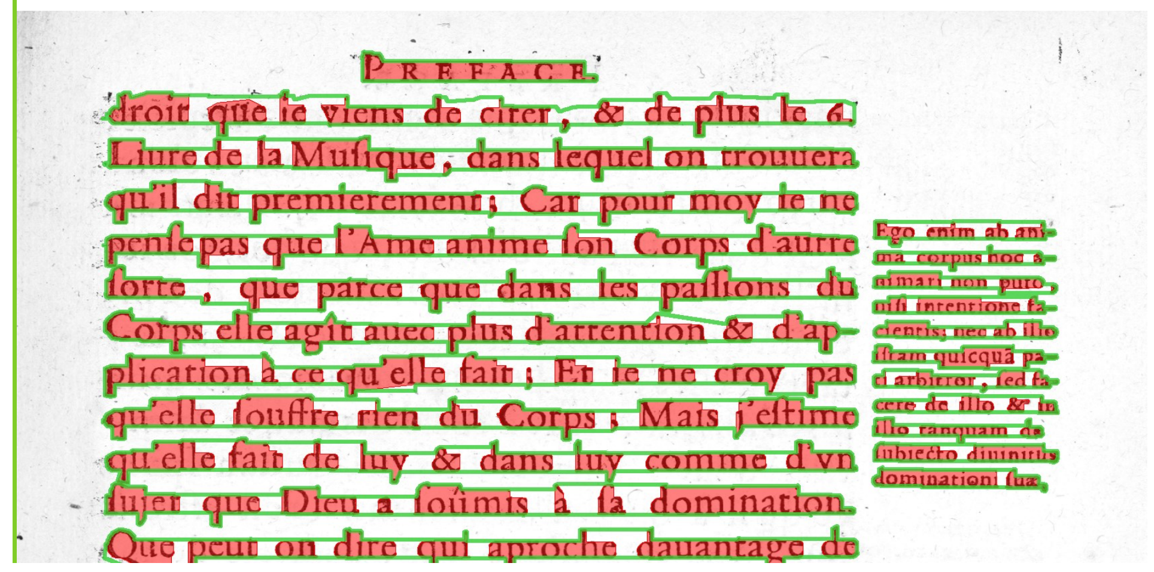
Greek
Cyrillic
Gaj
Latin/Gothic
Old Cyrillic
Hebrew
Latin
Bohoričica

Ground Truth

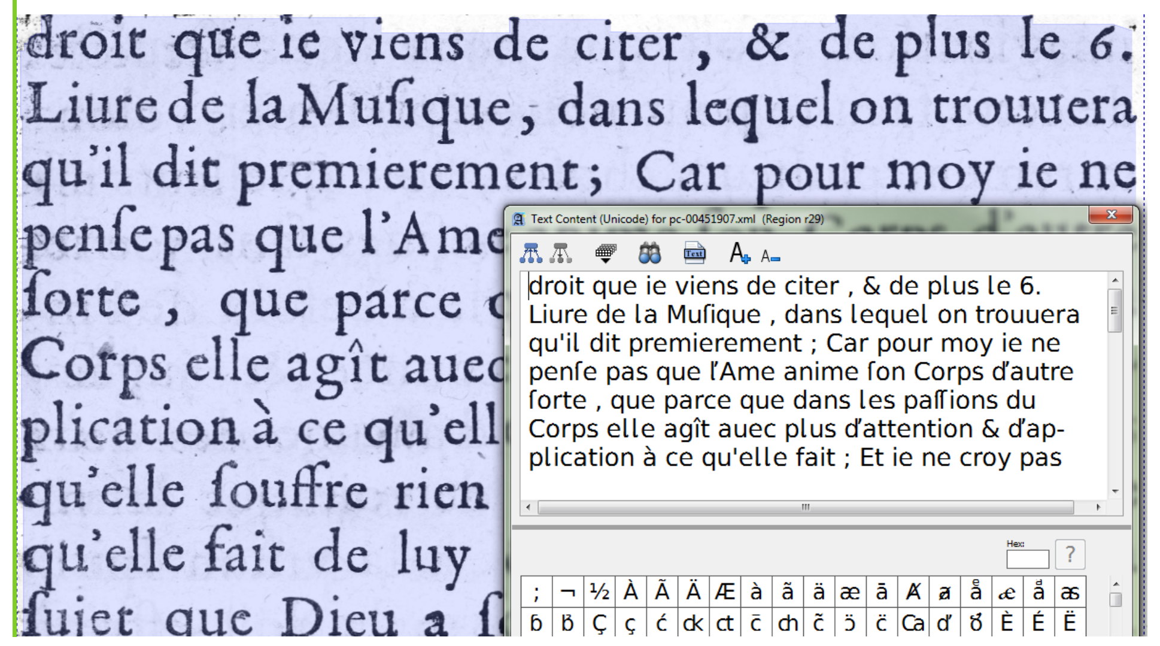
A carefully selected representative subset of images from the dataset has been enriched with ground truth. In digital imaging and OCR, ground truth is a representation of the actual (correct) content as found on the page. For a document image's text content, for instance, it is the complete and accurate transcription produced by a human. Ground truth is used for evaluating the accuracy of automated systems and/or training new methods.

Line and Word Outlines

About 300 images have even more detailed ground truth, leading to more than **5,000 text line outlines** and **70,000 word outlines** with corresponding text.



Text Content



droit que ie viens de citer , & de plus le 6.
Liure de la Muſique , dans lequel on trouuera
qu'il dit premierement ; Car pour moy ie ne
penſe pas que l'Ame anime ſon Corps d'autre
ſorte , que parce que dans les paſſions du
Corps elle agît auec plus d'attention & d'ap-
plication à ce qu'elle fait ; Et ie ne croy pas

Layout Regions

45,000 images have region outlines, text content and reading order.

573,725 text regions, including:

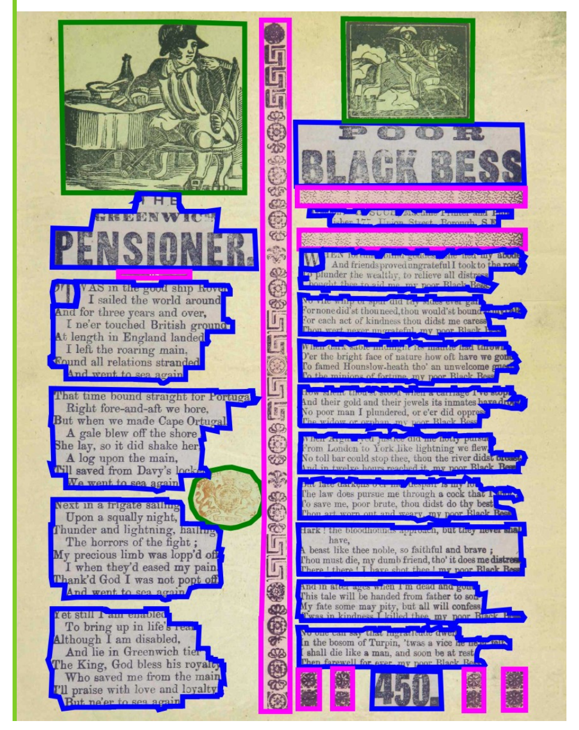
388,636 paragraphs,

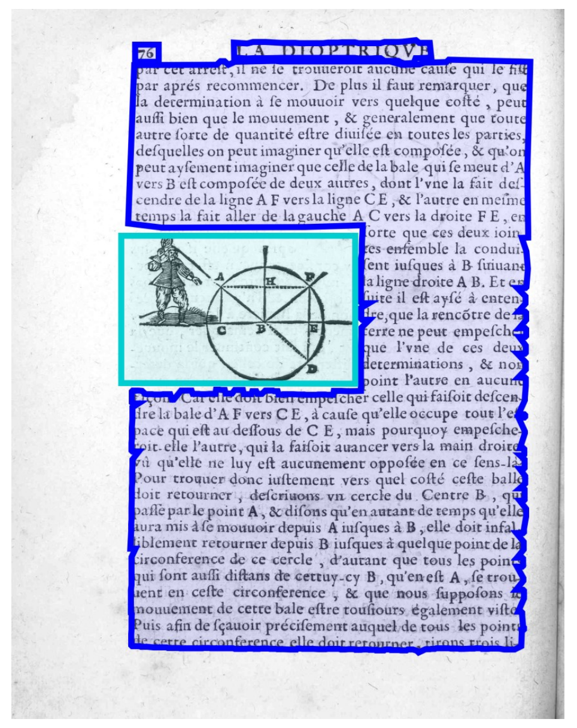
42,345 headings,

6,211 captions

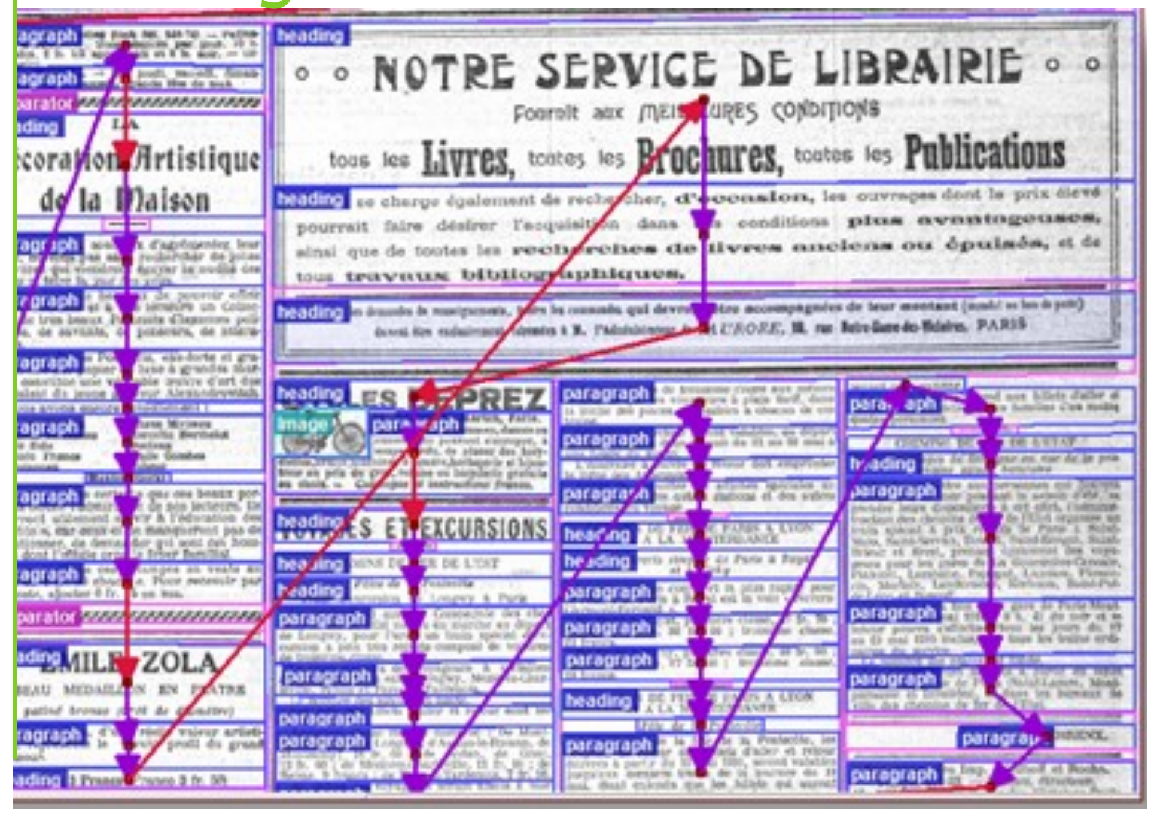
In addition, there are headers and footers, footnotes, signature marks, catch words, table of content entries, page numbers, marginalia, and credits.

10,151 graphic regions such as logos, stamps, handwritten annotations, punch holes, signatures, etc.





Reading Order



SUCCEED is supported by the European Union under FP7 -ICT and coordinated by Universidad de Alicante

University of Salford
MANCHESTER

PRImA
Research Lab

www.primaresearch.org/datasets

www.succeed-project.eu