

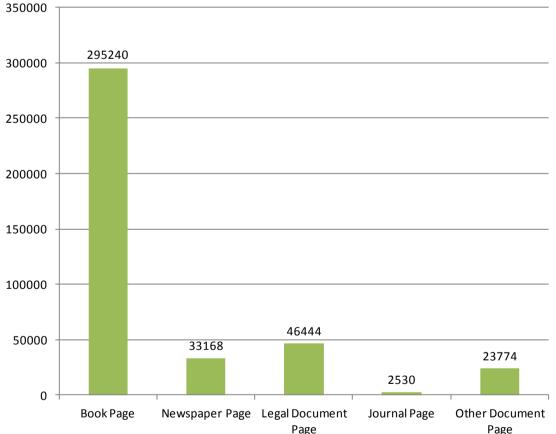
The IMPACT Digitisation Image Repository

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



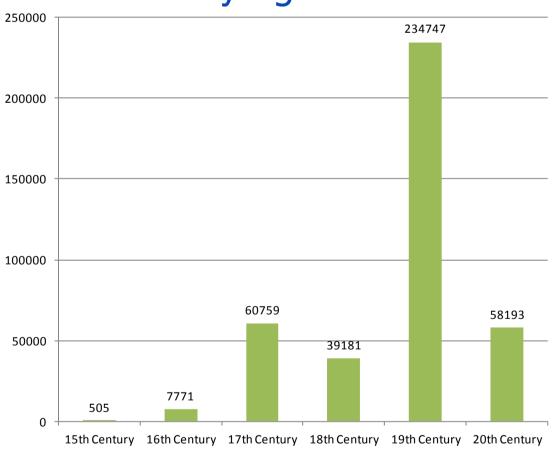
Languages



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



L'AURORE MOTERNATIONE MAINTE AND ADMINISTRATION ADM DISCOVRS TROISIESME. Exagérations peut on dire qui aproche dauantage de le nous auons écrit de l'Union de l'Ame

ENZALA e la dotrina que el Hijo de Dios vino a ense. n de las Canciones, ces & des Idées tinuë ainfi , Et Et ne longum fasposa, iel Esposo capazidad para poder recebir el Espiritu de Dios en pu transformacion. Desto tenemos figura en el libro del Exodo, donde se lee, que no dio la Magestad de Dios el manjar del cielo, q era el Manà: Ecce ego pluam vobis pa-nes de cœlo, a los hijos de Ifrael, hasta que les falto la ha-rina que ellos avian traido de Egipto. Dando por esto a font fur luy ren-ON PRIMERA. aciles, ou qu'eltiam, non eam la tere, & hoc est. felon l'oposiquod sentire dicitur. ntender, que primero conviene renunciar todas las co as, porque este manjar de Angeles no es ni se da al par,que quiere tomar sabor en el de los hombres. I no olamente se haze incapaz del Espiritu divino el almaque le apacienta i detiene en otros estraños gustos: mas Sali trasticlamando, i eras ido. an mucho a la Magestad divina, los que precen iendo el manjar de Espiritu, no se contentan con solo DECLARACION. Dios, sino que quieren entremeter el apetito i aficion de otras cosas. Lo cual tambien se echa de ver en la mis

enuoyent à nos sens , n'agissent pas sur l'A- in anima

Verbo Hijo de Dios su Esposo, descando unir-se con el por clara i esencial visión, propone Num. 11 ma Escritura, donde se dize: Quis dabit nobis ad vescendit sus ansias de amor, querellandose a el de su, ausencia: mayormente que estando ella herida de su amor, por el cual ha falido de todas las cofas i de si mes. ma, toda via aya de padecer la ausencia de su Amado,

there was no man that durste abyde about hym, nor come nyghe hym to holde his horse; and he hadde no seruante to serue hym, but let his horse stande there at the gate, and drewe out of his swerde, whiche was all swerde in his hande, she was sore aferde, and wolde haue flede a way frome hym, for she knewe wel his condycyons. Robert, seynge that every body dyde flee from hym, and that his owne moder wolde haue fledde in lykewyse, he called unto her pyteously afarre, and stande styl tyl I haue spoken with you, and flee not from me in the worshyp of Crystes passyon." Than Roberte's herte beynge full of thought and repentaunce, wente nygher her, saynge thus: "Dere lady moder, I praye and requyre you tell me how and by what maner or wherby cometh it that I am soo vycyous and curste for I knowe wel I haue it other by you or of my fader wherefore incontynent I hertly desyre and praye you that ye shewe me the trouth hereof." How the Duchesse desyred Robert her sone to smyte of

her hede, and than she tolde hym howe she had gyuen hym to the deuyll in his concepcyon. THE duchesse had gretly meruaylynge whan she herde her sone speke these wordes; and piteously wepynge, with a sorrowful herte saynge thus to hym: "My dere

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.

droit que le viens de citer, & de plus le 6. Liure de la Musique, dans lequel on trouuera qu'il dit premierement; Car pour moy le ne pensepas que l'Ame anime son Corps d'autre Corps elle agit auen plus d'arrention & d'ap cientis ner abilio plication à ce qu'elle fait; Et le ne croy pas d'arbitror, sed fa qu'elle souffre rien du Corps; Mais j'estime flo ranquam de qu'elle fait de luy & dans luy comme d'yn subjecto diumnits domination sus jet que Dieu a soûmis à la domination

Text Content droit que le 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 a Text Content (Unicode) for pc-00451907.xml (Region 129) droit que ie viens de citer, & de plus le 6. forte, que parce de la Musique, dans lequel on trouuera Corps elle agît auec qu'il dit premierement ; Car pour moy ie ne pense pas que l'Ame anime son Corps d'autre plication a ce qu'ell Corps elle agît auec plus d'attention & d'application à ce qu'elle fait : Et je ne croy pas qu'elle souffre rien qu'elle fait de luy

Layout Regions

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

573,725 text regions, including:

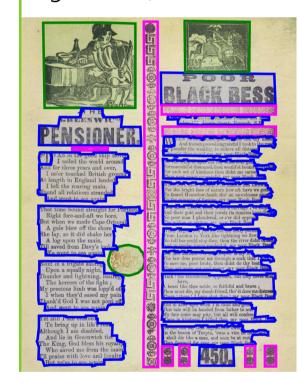
388,636 paragraphs,

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









