Recognition and Enrichment of Archival Documents

D8.12
Large Scale Demonstrators

Keyword Spotting in Registry Books P3

ABP
Hannelore Putz, Eva M. Lang, Wolfgang Fronhöfer,
Andrea Fronhöfer, Elena Mühlbauer

Distribution: Public

http://read.transkribus.eu/

READ
H2020 Project 674943

This project has received funding from the European Union's Horizon 2020
research and innovation programme under grant agreement No 674943

Project ref no.	H2020 674943
Project acronym	**READ**
Project full title	**Recognition and Enrichment of Archival Documents**
Instrument	H2020-EINFRA-2015-1
Thematic Priority	EINFRA-9-2015 - e-Infrastructures for virtual research environments (VRE)
Start date / duration	01 January 2016 / 42 Months

Distribution	Public
Contractual date of delivery	31.12.2018
Actual date of delivery	22.12.2018
Date of last update	05.12.2018
Deliverable number	D8.12
Deliverable title	Passau – Keyword Spotting in Registry Books
Type	Report
Status & version	Final
Contributing WP(s)	WP5, WP6, WP7
Responsible beneficiary	ABP
Other contributors	NLE, CVL, UPVLC, URO, DUTH
Internal reviewers	NLE; CVL; NAF; StAZH
Author(s)	Eva M. Lang
EC project officer	Martin Majek
Keywords	Large Scale Demonstrator, Ground Truth, Archives, Reference Data, Handwritten Text Recognition, Key Word Spotting, Document Understanding, Table Matching, Table Processing, Information Extraction, Registry Books

Table of Contents

Executive Summary

The Archive of the Catholic Diocese of Passau (ABP) is one of the Large Scale Demonstrators within the READ project. During the 2018 reporting period, the main efforts of the ABP have been the application and testing of the methods developed by the technical partners.

This report summarizes the achievements and results of the last three years (Section 1), presents evaluation of the key word spotting and information extraction on our sources, compares the applications provided by different partners (Section 2), and gives a prospective application to other archival institutions and sources (Section 3). In Section 4, we report about the different working groups and activities supporting WP 8 and list publications and datasets in Section 5.

All efforts were carried out by the interdisciplinary team at ABP together with the various technical partners within the READ project.

1. Achievements and results

The ABP team consisted of Computer Scientist Eva Lang, Media Expert Elena Mühlbauer, Senior Archivist Wolfgang Fronhöfer and Historians Andrea Fronhöfer and Hannelore Putz. Given this cross-disciplinary project setup, we were able to focus on providing expert transcription leading to good HTR results and constantly working towards user-friendly implementation of Transkribus X and the web interfaces.

As reported in 2017, we directed our efforts to a subset of all 800,000 available registry book scans. The team provided a set of 1,200 images and transcripts of death records in expert transcription mode. This was mainly done by outsourcing segmentation and transcription to a subcontractor and correcting results to expert standard in-house. The full set of death records from 1847-1878 (in total 26,579 images) was select as one of the test frameworks for automatic table processing, document understanding and for enhancement of keyword spotting techniques.

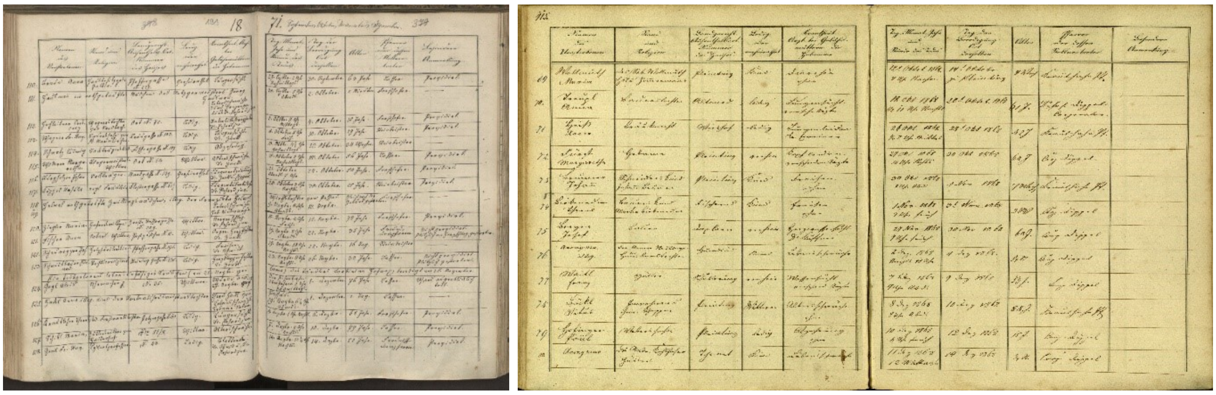


Figure 1 Two sample images from the ABP death records collection ABP_S_1847-1878

Thanks to the support of technical READ partners at URO (HTR / HTR+, KWS), UIBK (general infrastructure), UPVLC (indexing KWS), DUTH (KWS by example), NLE (table processing and information extraction), CVL (table matching and writer identification) we were able to experience and test the results of cutting-edge research and development carried out on our data in Year 3.

Our chosen dataset with more than 590 hands proved to be very complex for scholars, researchers and technical partners. The data proved very useful for the development of table processing and document understanding use cases (WP 6), key word spotting applications (WP 7) as well as for developing general guidelines for table segmentation and user-friendly interfaces (WP 4).

2. Use Cases

ABP followed an interdisciplinary approach, which aimed at honing and developing various applications. Due to the high interest in our selected dataset and the advancement of tools developed especially in Y3, ABP was able to test the applications such as Key Word Spotting and Information Extraction provided by the project partners.

2.1. Key Word Spotting

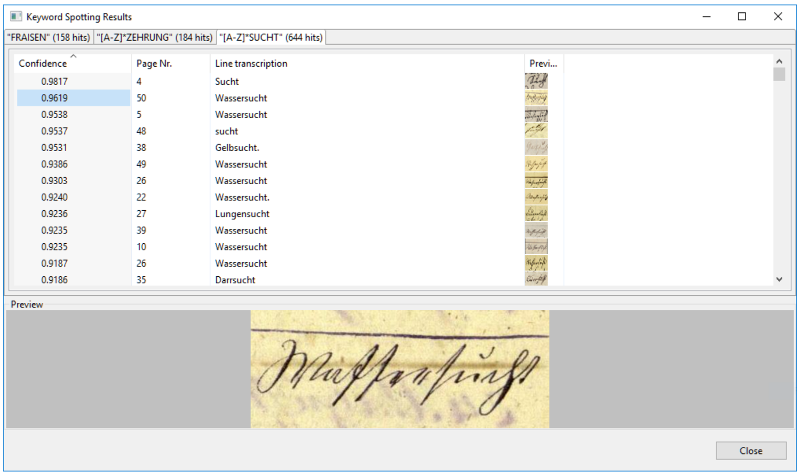
The partners provided three different interfaces and methods for searching our data. While the URO search tool is directly integrated in Transkribus, the demonstrators provided by UPVLC and DUTH are at time of writing stand-alone applications on external servers.

All three methods used the same set of images which we had prepared for test cases in the first year. Technical details can be found in the WP 7 reports.

2.1.1. URO approach: Keyword Spotting in TranskribusX

The tool offered by the University of Rostock (see also deliverable D 7.9) was integrated into TranskribusX in Y2. It offers a simple search interface and is easy to use. Searching for a specific key word or combination of search phrases works effectively and with a high rate of correct results on a given Transkribus document, e.g. a volume specific to death records in a parish, selected pages for one hand etc.

This opens avenues for questions dedicated not only to general history, but also to social, economic and population history. Therefore, we can serve scholarly researchers looking for statistical data, for example to search for different forms of illnesses, names of families, and



Confidence	Page Nr.	Line transcription	Previ...
0.9817	4	Sucht	
0.9619	50	Wassersucht	
0.9538	5	Wassersucht	
0.9537	48	sucht	
0.9531	38	Gelbsucht.	
0.9386	49	Wassersucht	
0.9303	26	Wassersucht	
0.9240	22	Wassersucht.	
0.9236	27	Lungensucht	
0.9235	39	Wassersucht	
0.9235	10	Wassersucht	
0.9187	26	Wassersucht	
0.9186	35	Darrsucht	

locations, etc. Through the connection between the key word and the actual location of the written word, KWS serves similar to an index on the current Transkribus document.

Figure 2 Search for any type of "sucht" (addiction)

2.1.2. UPVLC approach: Keyword Spotting by indexing

The approach by Polytechnic University of Valencia was already described in depth in the previous report (D 8.11). Our main contribution in 2018 was to produce sample queries for statistical evaluation by the UPVLC, which were presented at the ICFHR conference (see also deliverable D 7.3).

The UPVLC KWS platform provides a very time-efficient tool to search and present results, spanning thousands of documents. The tool is independent of differing hands and covered time-spans. Even several centuries can be covered if necessary. It is therefore proof of application on large scale. The technique also offers searching for combination of keywords and therefore provides entry points for complex and elaborate research questions.

2.1.3. DUTH approach: Keyword Spotting by example

The colleagues at the Democritos University of Thrace set up a demonstrator using ABP test data (for details see deliverable D 7.15). The platform allows searching for visual information after the user provides an image snippet. This approach can easily be transferred to other datasets as it only requires baselines to start with, no transcript needs to be supplied.

The end-user can work with this method to check different forms of writing style for a given key word and therefore assist as a method to learn different hands.

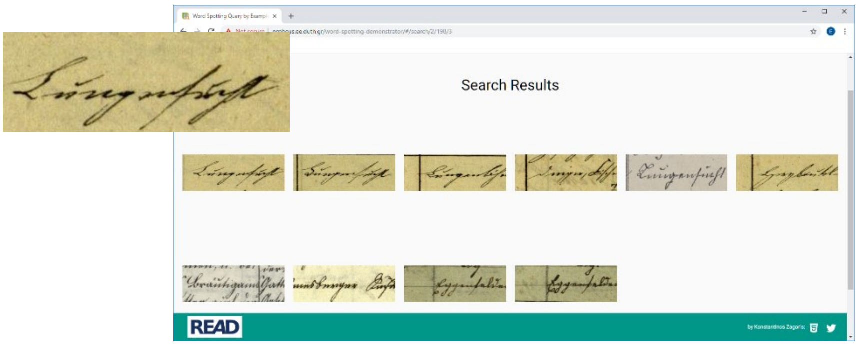


Figure 3 Query by example search for "Lungensucht" using snippet on the left

2.2. Information Extraction

In the course of the READ project, the HTR quality improved very much. The use of dictionaries dedicated to specific entries such as dates, locations, names etc. seems to be useful on single or few hands with little training material.

Thanks to the efforts of the URO and UIBK teams, a first optimized HTR+ model was provided by URO for tests in June 2018 with a CER of 6.98 %. In September, we trained new models for our selected scribes and achieved a CER of 8.00 % and WER of 20.44 % on these 40 hands. On a separate evaluation set of 50 pages consisting of hands not known to the HTR+ engine, we reached a CER of 10.67 % and WER of 29.98 %.

The end-to-end workflow from image to data records had first been tested using the old implementation of the HTR in early 2018, the full automation on our death records dataset is yet to be tested. Results are forthcoming and described in depth in deliverable D6.15.

2.3. Writer Identification

Due to the pre-selection based on the curriculum of the actual scribes of our parish registry books and the already existing index database at ABP, a new dataset for Writer Identification of historic hands was created and provided for the scientific community on zenodo (see Section 5.3 below). Details can be found in deliverable D 5.10 – for technical details see also deliverable D 7.18.

2.4. Prospects and trends

In a side project, an intern of ABP segmented example pages of the Ordinariatsprotokolle, a typical court minutes structure. The student scanned the pages with the help of the ScanTent and the DocScan app, checked and uploaded the images to Transkribus. Segmentation was done with the aid of the baseline finder tool (Layout Analysis by CITlab, URO) in Transkribus.

50 pages were transcribed in an iterative process where two HTR engines were trained, applied to the segmented pages, corrected and results looped back into a re-training of the net. This project also served as one of the first HTR+ projects after the tool was integrated into Transkribus in the second half of the year.

Overall results are very good showing a CER of 6.35% on the evaluation set.

The selected sources also aroused great interest in the community – nearly every archive has similar sources and records, so the information extraction on these sort of tables is key. The workflow from image to information, as developed together with READ partners NLE and CVL, should be applicable and transferable to similar sources.

Archival READ partner StAZH applied this HTR+ model to their own sources and received promising results of around 8% CER.

3. Short evaluation and lessons learned

From ABP perspective, the READ project brought together different areas of expertise and opened avenues for research questions which users had not been able to ask before. The team benefited from the archival knowledge and long-term experience with queries from archival users.

What we had underestimated at the beginning, was the time expenditure for the production of high-quality ground truth. Time effort and resources were spent for selecting the data, producing transcripts and tagging and expanding abbreviations. The direct feedback loop between transcribers and developers increased the usability of the software as well as the quality of transcriptions.

Making use of external subcontractors for transcription and tagging, while doing the quality control and corrections in house, sped up the process and time for ground truth production and reduced not only production time but also production costs.

The algorithmic enhancements provided by the technical partners and the opening of the methods in year three, gave us the chance to directly see and evaluate the results on the data we had provided.

An overall HTR quality of 8-12 % CER on our death record test sets is trend-setting for further archival documents and sources. Sample tests have also shown that writer-specific training can optimize recognition results. Due to the multitude of hands and the complexity of the writing and layout, however, the general model is the most promising way and already provides access to a large collection of documents which had not been accessible to the audience for search.

Thanks to the methods developed and tested within the READ project, research studies based on the evaluation of large-scale datasets or dedicated studies based on a specific aspect in a large dataset, can now be carried out on the selected death records. Applications of the technology to baptism and marriage records are forthcoming as well using the trained HTR+ nets on the same period of writing.

4. Activities supporting WP 8

As stated in the WP description (see GA), ABP was focusing on the experience of the archival users. Due to regular feedback from the subcontractor and from our transcription experts, several requests for module development were passed on to the Transkribus developers, some of which were also taken on and developed by ABP.

4.1. Software Development focusing on the user perspective

ABP involvement in Software development and functionality enhancement of TranskribusX modules and concentrated on providing development and input to the following tasks

- Enhancing the essential upload functionality for PDF documents, which now allows the user to work with large files
- Enhancing the usability of the tagging framework
- Developing a user friendly interface for marking graphical lines in tables
- Testing and debugging of web interface functionalities representing the archival role
- Testing and debugging the python framework supporting the automatic processing of images to information workflow by NLE
- Providing feedback for and implementing sub tasks related to general table processing

4.2. Dissemination Activities

ABP participated in the regular conferences of the dissemination working group. For written publications by ABP, see Section 5 below. General dissemination activities within the READ project are presented in WP 2.

4.3. Inter-Archival Working Group

ABP regularly communicated with the core READ archival partners ABP, NAF, StAZH to streamline processing and exchange best practices.

4.4. Table Competition Working Group

ABP played an active role in setting up the dataset for the planned ICDAR 2019 Table Competition (proposal submitted) together with partners from NLE and CVL. Further details about the work of the group the READ project can be found in WP 3 and WP 6.

5. Papers and Publications

5.1. Papers and presentations

- 04.12.2018, Passau, Universität Passau, Oberseminar Lst. Digital Humanities, Prof. Malte Rehbein, Automatische Handschrifterkennung; Transkribus – eine Einführung (Eva Lang)
- 07.-08.11.2018, Wien, Universität Wien, Transkribus User Conference 2018: Automated recognition using Transkribus in the Passau Diocesan Archives(Eva Lang)
- 04.-05.10.2018, Rostock, Universität Rostock, Bibliotheca Baltica 2018: READ technology on Large Scale; Use cases, applications and results from READ demonstrators (Eva Lang)
- 13.07.2018, Passau, Archiv des Bistums Passau, Oberseminar Lst. Bayerische Landesgeschichte, Prof. Ferdinand Kramer (LMU): Automatische Handschrifterkennung; Transkribus – eine Einführung (Elena Mühlbauer)
- 05.06.2018, Passau, Archiv des Bistums Passau, Lst. Kunstgeschichte / Bildwissenschaft, Prof. Jörg Trempler: Automatische Handschrifterkennung; Transkribus – eine Einführung (Eva Lang)

5.2. Publications

- Lang Eva Maria, Alte Handschriften spielerisch lernen. in: Passauer Bistumsblatt 83 (2018) H.40, S. 16.
- E. Lang, J. Puigcerver, A. H. Toselli, E.Vidal, Probabilistic Indexing and Search for Information Extraction on Handwritten German Parish records. in: Proceedings of the 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), (forthcoming), 2018.
- F. Kleber, M. Diem, H. Dejean, J.-L. Meunier, E. Lang, Matching Table Structure, of Historical Register Books using Association Graphs. in: Proceedings of the 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), preprint, 2018, p. 217-222.
- S. Clinchant, H. Dejean, J.-L. Meunier, E. Lang, and F. Kleber, Comparing machine learning approchaes for table recognition in historical register books, in: Proceedings of the 13th IAPR International Workshop on Document Analysis Systems (DAS 2018), 2018, p. 133-138.
- Mühlbauer Elena: How to use Transkribus in 10 steps (updated), https://www.youtube.com/watch?v=GjChcDExshU (tutorial video by ABP)
- Lang Eva: How To Process Tables in Transkribus, https://transkribus.eu/wiki/images/1/14/HowToProcessTables.pdf (written tutorial by ABP)
- Mühlbauer Elena: How To use Transkribus eLearning (written tutorial, to be published)
- Mühlbauer Elena: How To Process Tables in Transkribus (tutorial video, to be published)

5.3. Datasets

- Déjean Hervé, Lang Eva, & Kleber Florian. (2018). READ ABP Table datasets (Version 1.1) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.1243098

- Fiel Stefan, Kleber Florian, Lang Eva-Maria & Fronhöfer Wolfgang. (2018). READ ABP WI Dataset - Writer Identification over decades [Data set]. Zenodo. http://doi.org/10.5281/zenodo.1421600
- Lang Eva-Maria, Fronhöfer Wolfgang, Putz Hannelore (2018). READ ABP S 1847-1878 [death records dataset used project internally]

RECOGNITION & ENRICHMENT
OF ARCHIVAL DOCUMENTS

D8.12 (Annexe A)

Large Scale Demonstrators. Keyword Spotting in Registry Books

Large-scale Probabilistic Word Indexing of Handwritten Text Images (UPVLC)

Alejandro H. Toselli, Enrique Vidal
UPVLC

Distribution: http://read.transkribus.eu/

READ
H2020 Project 674943

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 674943

Project ref no.	H2020 674943
Project acronym	READ
Project full title	Recognition and Enrichment of Archival Documents
Instrument	H2020-EINFRA-2015-1
Thematic priority	EINFRA-9-2015 - e-Infrastructures for virtual research environments (VRE)
Start date/duration	01 January 2016 / 42 Months

Distribution	Public
Contract. date of delivery	21.11.2018
Actual date of delivery	31.12.2018
Date of last update	31.12.2016
Deliverable number	D8.12 (Annexe A)
Deliverable title	Large Scale Demonstrators. Keyword Spotting in Registry Books
Type	report
Status & version	in process
Contributing WP(s)	WP6
Responsible beneficiary	ABP
Other contributors	UPVLC
Internal reviewers	Günter Mühlberger
Author(s)	Alejandro H. Toselli, Enrique Vidal
EC project officer	unknown
Keywords	probabislitic indexing, information extraction

Contents

1 Introduction

The accuracy of handwritten text transcription (or "Handwritting Text Recognition", HTR) has dramaticaly improved in the last few years. Impresive Word Error Rate (WER) as low as 5-10% is often reported for well-defined experimental handwritting datasets, with simple layout, well detected text lines and sufficiently clean images. However, when real bulk data, from *complete*, *large* collections is considered, transcription accuracy generally drops dramatically.

In many cases, the main target application of HTR is to allow searching for textual information in the considered collection of text images. For these applications, rather than insisting in trying to achieve perfect transcripts for human reading, one can instead attempt to allow searching for textual information in the original images, without relaying on explicit transcripts of the images.

As discussed below, with this idea in mind, UPVLC researchers have developped the concept of *Probabilistic Indexing of Text Images* and the associate indexing, search and retrieval technologies. These technologies have reached in READ a high degree of maturity which allows to undertake very large collections of scanned manuscripts in a fully automatic way.

As part of the commitment of UPVLC to Work package WP8 "Large Scale Demonstrators", two large collections of paramount interest to READ partners or MOU partners have been very successfully indexed during 2018. As a result, these two paradigmatic collections are now fully and efficiently searchable, to the great satisfaction of the scholars and general public interested in the information contained in these collections.

The work carried out to achieve these objectives is reported in the following sections.

2 Probabilistic Word Indexing (PWI)

A *probabilistic word index* (PWI) of a text images is a probability map which asignes to each image pixel the probability with which each word or character sequence (called "pesudo-word") appears in a word-sized region of the images which containes this pixel. Fig. 1 ilustrates this concept. Efficient technlogy to compute PWIs has been developped by UPVLC researchers in the framework of READ and other related projects.

3 Practical Considerations for Large-Scale PWI

Searching for information in untranscribed text image collections can be achieved by means of keyword spooting (KWS). KWS and, in particular *query by string* (QbS) KWS, nicely lends itself to searching under the *precision-recall tradeoff model*: it allows the users to somehow specify in each query whether they need the results with more precision or more recall.

3.1 Keyword Spotting and the Need for Word Image indexing

In general, to achieve acceptable query response times, all large-scale information retrieval tasks rely on a preparatory, off-line phase where adequate *indices* are *precomputed* [4][1]. This

[1] Free on-line version at `http://nlp.stanford.edu/IR-book/html/htmledition/irbook.html`

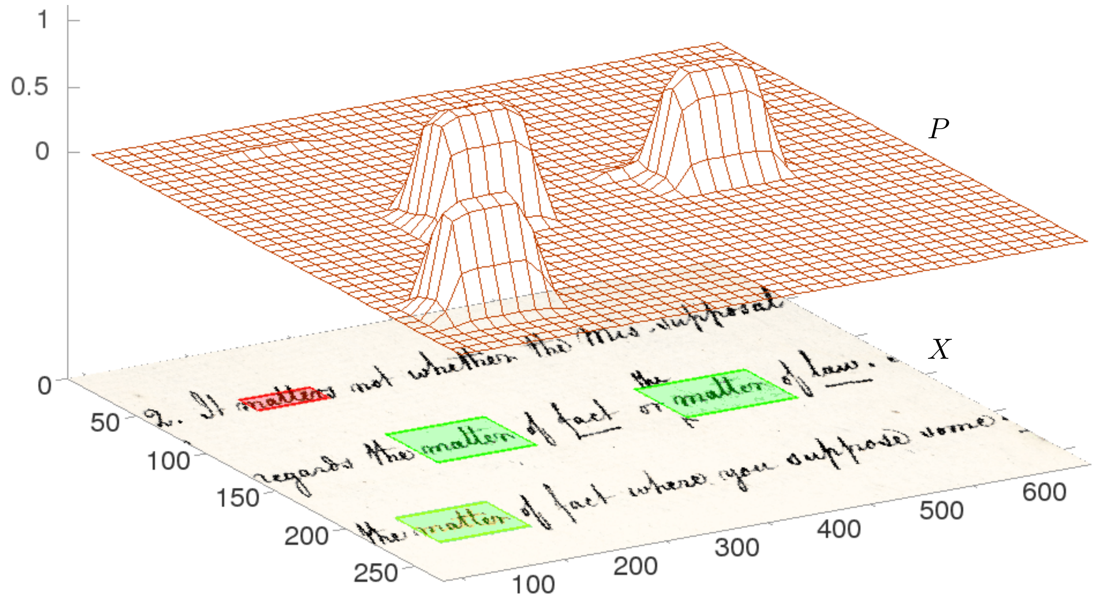


Figure 1: Probability map P of an image X for the word “matter” and PWI for this word.

off-line precomputation is still more necesssary in our case, since direct KWS, even using rich data structures extracted from images (such as word or character lattices, confMat’s, etc.), is computationaly expensive [9]. Therefore, in large collections of, say, tens or hundreds of thousands text images, the only reassonable way to support responsive keyword queries is by using *precomputed keyword indices*.

A keyword index should contain, for each image region of interest in the text image collection, a *list of keywords* which probably appear in that region, along with the corresponding *confidence scores* and *locations*. By stockpiling this information into an adequate database or data structure, fast response times (under a fraction of a second) can be easily achieved, with little dependence on how large is the indexed collection.

The *confidence score* of a keyword should be a real number proportional, or at least directly related, to the probability that the keyword appears in the corresponding region and location. Using scores bounded in $[0, 1]$ which can be properly interpreted as *relevance probabilities* has many advantages, as it will be seen throughout this document.

The *location*, on the other hand, should inform about the geometrical position of the keyword in the image region. In its simplest form, a location can be just the pair of estimated X-Y coordinates of the keyword *center*; in addition, the *width* and *height* of a bounding box which may contain the keyword image can help nicely displying query results.

To actually provide the intended service, the number of indexed keywords needs to be (very) large – say, tens or even hundreds of thousands of keywords. These keywords can be obtained from external linguistic resources and/or derived from the indexed images themselves. On the other hand, an index should not be too large. As a rule of thumb, the size (e.g. in Kbyts) of the (compressed) index part corresponding to a text image of the indexed collection should not be larger than one half of the size of the image itself [4].

Taking all these considerations into account, UPVLC has developped the Probabilistic Word Indices introduced un the previous section.

3.2 Architecture and Workflow

The proposed architecture is shown in Fig. 2. The block labelled *"KWS & indexing tool"* corresponds to the off-line pre-computation of the keyword index. The *"Ingestion"* process which creates the actual database is also an off-line task, althogh its computational cost will be generally neglieable, as compared with the cost of KWS and indexing. *"Keyword search"*, on the other hand, is in charge of analizing the user queries, finding the requested information in the database and present the retrived images. These three steps are to be performed on-line, with response times as short as possible.

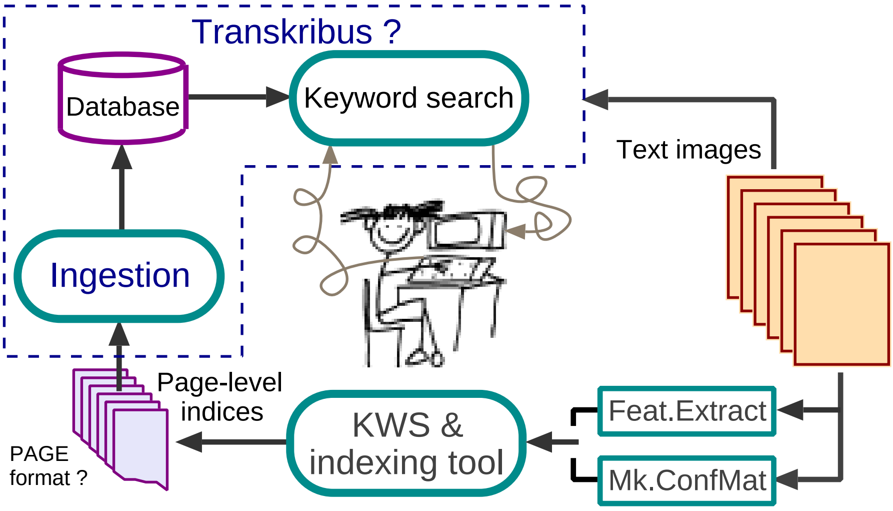


Figure 2: Proposed architecture and workflow.

Some open questions are:

- Should the PAGE format be used to store indices?
  - Pros: indices can be seen as metadata which acompany the raw images; PAGE is already a familiar metadata container.
  - Cons: PAGE files are becoming large an complex; a lighter format such as JSON, or a specialized plaintext representation, could be significantly more efficient in terms of storage (disk) requirements
- Should the "Keyword search" subsystem (including query front-end analysis and presentation of retrived images) be provided by Transkribus itself? Or should we implement a specialized system? (cf. Sec. 3.9)

These questions are so far still being pondered and no final decision has made so far.

3.3 Index Data

Assume the image regions of interest are just full page images. For a word w which appears in an image location l, let us define $g(w, l)$ as the set of pixels wich render (depict) w at l. For each page, the proposed index is compossed of an adequate number of entries or "spots" such as:

```
KWord, Score, GLoc, [SSize,] [Other ...]
```

`KWord`: a keyword ID of w; typically a character string. Several spots may have the same `KWord` field, provided that their `GLoc` (l) is different

`Score`: a real number; many advantages if bounded in $[0, 1]$ and interpreted probabilistically

`GLoc`: geometrical location (l) within the image. Ideally sould be defined as the two first-order *normalized geometrical moments* of $g(w, l)$ [5]; that is, as the X–Y coordinates of the center of mass of $g(w, l)$. Several spots, with different `KWord` and `Score` values, are typically expected for the same or similar `GLoc`,

`SSize`: size and shape of $g(w, l)$. Typically the width and height of a rectangular bounding box which taightly contain the pixels in $g(w, l)$. But it would be advantageous if it is rather defined as three parameters (mass, orientation and excentricity) derived from the first-order and the three second-order *central moments* of $g(w, l)$ [5]. In any case, I see these data as optional, because in most applications it is more than enough to present the retrived results just as points or small greyed zones where the query words may approximately appear in the images

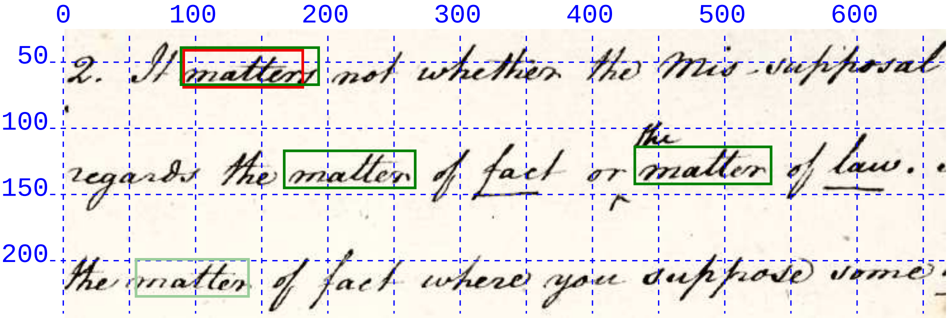
`Other`: additional data which may prove handy depending on the application-dependent search task. For instance, an ordinal number specifying the relative position of this word with respect to words in other index entries of this page, would be useful to simplify multiword phrase search (cf, Sec. 3.10). Another possibly useful information would be a bit ($\{0, 1\}$) specifying whether this spot can or cannot be considered *"ground truth"*; that is, whether its correctness has been manually verified. Such information could be used for experimental purposes, or in KWS-based, collaborative (crowd-sourcing) transcription endeavours such as those proposed in [10, 12].

3.4 Considerations About PWI Formatting

The spot data proposed above should be computed for each page image of the collection to be indexed. Then all these spots entries, along with the corresponding page image IDs, must be stored in some specific data structure, or "ingested" into an adequate databse which supports the kind of queries involved in our *precision-recall tradeof model.*

For average page and script sizes, a good index may contain something ranging from 5K to 50K spots per page. Using plain, uncompressed ASCII text, these spots will take about 100-1000 Kbytes, which is well within the rule-of-thumb size limits commented in Sec. 3.1).

If rather than plain ASCII, an XML PAGE-style representation is adopted, the uncompressed size per page would be roughly twice as large, but still well under our rule-of-thumb size limits. Therefore, only the pros and cons stated in Sec. 3.2 may help us deciding whether it is convenient to include indices into PAGE files. Nevertheless, even if we decide not to "merge" index (meta-)data with other types of metadata already included in PAGE, a separate PAGE representation of indices can obviusly be chosen if considered convenient. A small exeample of plaintext index file is shown in Fig. 3.


0
100
200
300
400
500
600
50
100
150
200

```
# pageID="Bentham-071-021-002-part"   REGARDS  0.857    5 115   84 31        THE  0.990    1 198   28 31
#     keyword relPrb     bounding box  UGARDS  0.138    5 115   80 31     MATTER  0.934   61 198   64 31
#                                         THE  0.993  110 115   43 31         OF  0.988  141 198   28 31
            2  0.929    1  36   20 31  MATTER  0.998  160 115   93 31       FAST  0.367  182 198   62 31
           21  0.064    1  36   24 31      OF  0.996  271 115   23 31        FAR  0.186  182 198   36 31
           IT  0.982   33  36   27 31    FACT  0.999  306 115   49 31        ...  ...      ...      ...
           IF  0.012   33  36   26 31      OR  0.973  377 115   37 31       FACT  0.017  182 198   46 31
      MATTERS  0.998  160 115   93 31      ON  0.021  377 115   42 31         AS  0.142  200 198   29 31
       MATTER  0.011   77  36   93 31  MATTER  0.990  425 116  100 31        HAE  0.022  200 198   29 31
          NOT  0.999  216  36    7 31      OF  0.995  542 115   25 31      WHERE  0.992  255 198   90 31
      WHETHER  1.000  256  36   99 31     LAM  0.407  575 115   30 31        YOU  0.761  365 198   45 31
          THE  0.997  389  36   33 31    BIMR  0.175  575 115   55 31        YOW  0.030  365 198   45 31
 MIS-SUPPOSAL  1.000  455  36  193 31     ...  ...      ...      ...       GOUS  0.064  372 198   47 31
                                          LAW  0.032  575 115   36 31    SUPPOSE  0.975  429 198  120 31
          THE  0.927  430  88   30 31    TAUE  0.031  575 115   55 31   SUPFROSE  0.024  429 198  125 31
          LHE  0.056  434  88   25 31     ...  ...      ...      ...        SOME  0.834  570 198   78 31
          ...  ...      ...      ...     LANE  0.012  575 115   59 31      SONER  0.016  576 198   83 31
                                                                             OME  0.109  580 198   65 31
                                                                              ME  0.022  620 198   22 31
```

Figure 3: PWI of an image region enphasizing entries for the words "matter" and "matters".

3.5 Indexing Through Query by Example (QbE) KWS

It is unclear how query-by-example (QbE) KWS techniques (such as, e.g., those described in [2] or [12]) could be used for large-scale indexing and retrival applications. Probably the index structure and formats discussed in the previous section are totally inadequate in the QbE framework and something completely different might be needed.

Nevertheless, recent work shows that QbS KWS approaches (such as those implicitly considered so far in this document) can be advantageously used for, or seamless merged with QbE KWS – see [11], and [1, 7], respectively.

3.6 PWI Preparatory Empirical Assessment

Several technologies, most of them based on KWS methods, are available to implement an indexing tool. The quality expected for the indices produced by an indexing tool can be easily estimated through relatively simple experiments that compute precision-recall curves, along with scalar measures such as the *average precison* (AP) and the *mean average precision* (mAP). To carry out these experiments only a relatively small but representative set of (50-100) *transcribed* images is needed. Before starting a large-scale indexing project, it is strongly advisable to perform this kind of experiments in order to compare the index quality estimated for several tools, and choose the one wich proves most promissing.

3.7 PWI Tools

The primary input of an indexing tool are text page images, possibly with PAGE metadata. For each image, the tool should produce an index of that page. The output may be written into a plain ASCII file or perhaps into PAGE (the same PAGE file of the image if it was avaialble, or a different file). An important paramenter is the *indexing density* aimed at. It can be given as a relevance score threshold, or as a number specifying how many spots per page, per image region, or per running word should be indexed.

The most effective KWS approaches generally require training data to estimate the paramenters of their statistical models and/or the weights of their neural network units. In these cases, the training system is generally a different piece of software, not included in the indexing tool proper. But the trained models do constitute another input to this tool. Moreover, indexingt tools may be designed not to work directly on the raw images. Instead they may take as input rich data stuctures derived from the images, such as word or character lattices or a confMat's. Additional, optional input data and paramenters for an indexing tool are:

- Keyword dictionary
- Layout analysis markup of text blocks (from PAGE files)
- Line detection or extraction markup (from PAGE files)
- Score normalizing and/or smoothing parameters
- etc.

In general, indexing tools are computationally (very) demanding. Therefore, for large image collections, decent (or even large) computational resources will be needed.

3.8 Hierarchical Indexing

Either if the index of a large collection is provided as a single, huge file, or it is split into (metadata) files associated to each image of the collection, the overall size will be very large. To manage such a large amount of information, it will hardly be possible to load the full set of data into main memory and the data-base tools and search engines will need to relay on adequately storing the index data into secondary storage.

Most large collections are not "flat"; instead they are typically structured hierarchically, as illustrated in Fig. 4. In these (usual) cases, one can leverage the natural hierarchy to provide a natural way to adequately store the index data (as well as other metadata and the images themselves) into secondary storage.

Moreover, in these cases, it becomes natural to use an *"aggregate index"* for each level of the hierarchy. Aggregate indices may allow the search process to start at the highest levels of the hierarchy and retrive candidate high-level elements such as bundles or books, rather than specific images. Then the user may chose some of these elements and go down to retrive lower-level elements such as chapters or sections. And so on down to the page images. These aggregate indices may consist of entries very similar to those proposed in Sec.3.3. For example, at a *collection* level, index entries may be something like:

```
KWord, Score, BookID, [Other ...]
```

And at a *book* level:

```
KWord, Score, PageID, [Other ...]
```

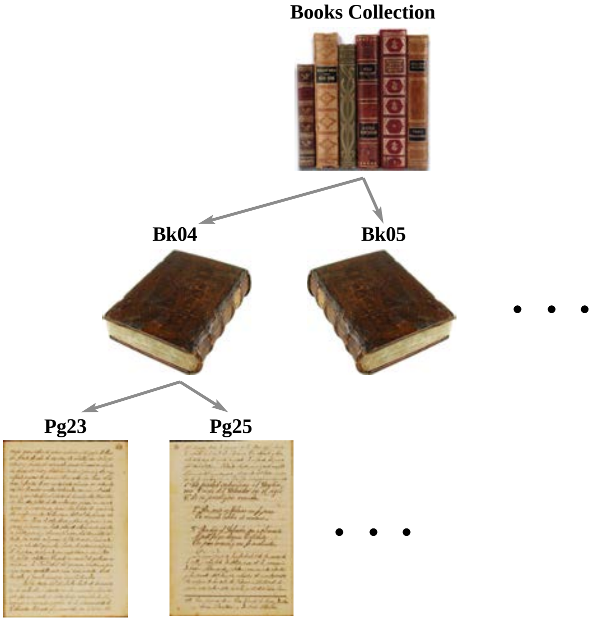


Figure 4: A hierarchical indexing and search model for handwritten text image collections. The top level in this illustration is a collection of books and the lowest level are page-sized images. The specific levels of a hirarchy should be defined according to the characteristics of the document collection and search task considerd.

For the keyword scores to be useful at the different levels, they must be *homogeneous* and properly *normalized* across the hierarchy. To this end, as previously commented, scores bounded in $[0, 1]$ are clearly advantageous. In addition, if the scores can be properly interpreted as true *relevance probabilities*, then it is feassible to compute good approximations to the true relevance probability of a keyword at one level, using the relevance probabilities already available for this keyword at the immediate lower level of the hiararchy [9, 8].

3.9 Search and Retrieval: Search Engine

The subsystem in charge of keyword search (see Fig. 2) encompasses three main modules: user *query analysis*, *search engine* and *presentatiion of retrived images* (see Fig. 5).

Query analysis is trivial for single-word queries, but it becomes significant for the kind of queries discussed in Sec. 3.10. In that case this module has to provide the user with graphical and/or textual means to specify several words *and* the possibly nested relations between them (such as "sentence", "AND", "OR", etc.).

In addition, this module should provide graphical and/or textual means to specify the precision-recall tradeoff desired for the query. At least two, complementary ways have to be supported: a *confidence threshold* and a *maximum number of retrived spots*. If scores are bounded in $[0, 1]$, the confidence thresold is just a number in this interval. Otherwise, some score normalization will be needed. In any case, it seems more natural to express this threshold as a percentage (in $[0, 100]$).

Depending on how the database has been created by the ingestion module of Fig. 2, The

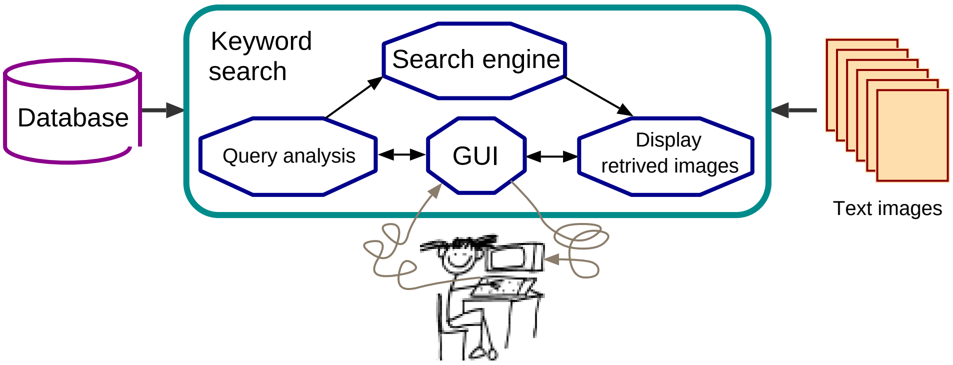


Figure 5: Search engine internals.

search engine can be implemented by means of calls to the corresponding database search engine. Using simple, open-source database software such as MySQL can easily and properly support both database creation and search for single-word queries and flat (non-herarchical) search. However, this solution can become exceedingly complex or even inappropriate to provide probabilistically consitent support for multi-word queries and/or for hierarchical search (see sections 3.10 and 3.8). Therefore developping a specilized database structure and search engine will probably be a simpler and much more effective and efficient solution. UPVLC has a working prototype of such an engine, which can be used as a starting point.

The module labelled "Display retrieved images" is in charge of preparing the images and/or image regions which should be presented to the users as a result of their queries. There are many ways the query results can be presented to the user and it is difficult to make a proper choice without taking into account the specific task or application of the keyword search system. Perhaps the two more "natural" ways are: a) dsiplay the retrived images in the natural order they appear in the image collection and b) display them in a decreasing order of their query relevance score or probability. In addition, the location of the query words in each retrieved image must be shown in some friendly maner which allows the users to focus their attention on the intersting zones of the text images, but do not blemish the image or hinder the reading of that zone.

Finally, the graphical user interface (GUI) provides the obvious, but not any less important function of allowing friendly bidirectional communication with the user.

3.9.1 Keyword Search in Transkribus?

The overall architecture and the query, search and retrieval modules discused in sections 3.9 and 3.2 are perfectly compatible with PAGE and the general Transkribus architecture. Therefore, endowing Transkribus with indexing and search capabilities might be useful only for relatively small collections. In that case, many of the issues discussed in Sec. 3.9, and others, become simpler and the implementation effort could be fairly low.

For a large-scale projects, however, using Transkribus, with its general purposeness, and its myriad of tools, options and features, might *not* a good idea. In this case, a specialized system, possibly based on a narrowed version of Transkribus and a home-made database and search engine, will probably be the right choice.

3.10 Multiple-word Queries

The primary purpose of indexing a text image collection is to allow the users to find *textual information* in this collection. While, for this purpose, just single-word queries will certainly entail a major break through in digital humanities, the next obvious step is to allow the users to formulate the same kind of multi-word queries they are used to in conventional plain-text retrieval applications like *Google*. Usual multi-word query combinations include:

- Boolean (AND/OR/NOT) queries
- Proximity (NEAR) queries
- Phrase or word-sequence queries
- Regular expresions?

From a statistical point of view, each of these types of multi-word queries would require computing the query relevance probabilities directly on the images (or on rich data structures derevid from the images, such as word/char lattices, confMats, etc.). This would allow us to take full advantage of the probabilisitic dependences between the involved words in the given word combination but, unfortunately, this is not compatible with (simple) indexing.

Nevertheless, the data we propose to be included in an index (Sec. 3.3) does provide the flexibility needed to support multi-word queries, using only the single-word relevance probabilities included in the index. In fact, very accurate relevance probabilities for boolean word combinations can be easily obtained using the approximations proposed in [6, 8]. Moreover, these approximations require only very little extra computation to be done by the search engine.

Boolean queries and search can be used as a basic building block to support other types of queries. Thus, *proximity queries* can be honored by first issuing an AND query, and then filtering out those retrived spots which are not geometrically close to each other. Cleary, geometrical proximity can be evaluated using the location information of the retrieved spots. The same idea can be used to honor *phrase* or *word-sequence queries*. In this case, the filtering after the AND retrieval would consist in checking whether the words in the retrieved spots can be considered to appear one ofter the other in the image. Again, this can be heursitically determined from the location and size information of the retrieved spots.

General *regular expression queries*, finally, do not seam to be easyly supported by precomputed single-word relevance probabilities. Therefore this remains an academically interesting open problem. Nevertheless, the above types of multi-word queries will probably be more than enough to cover most of the search demands which normally appear in information searching tasks on large handwritten text image collections.

3.11 Auxiliary Tools and Other Use Cases of PWIs

Tools which can work on indices and provide modified versions of these indices:

- Sanity checking,
- Computing index “density” (spots per running word)
- Index prunning,
- Score normalization and/or smoothing.
- Quality evaluation (computing R-P curve and AP/mAP).

The indices obtained for keyword search may have other applications:

- PDF export,
- Estimatomg the amount of text (running words) in a collection,
- Information retrieval (e.g. from tables, se Deliverable D6.9 and [3]),
- Textual-content-based image classification,
- . . . etc.

4 Large Collections Studied for PWI Demonstration

Three collections were considered as candidates for large-scale PWI demonstration: Passau, the Bentham Papers, and the Spanish "Teatro del Siglo de Oro" (TSO).

- *PASSAU Collection*. XVI-XVIII century collection of historical records. Hundreds of Thousands of images, written in German. Contain data about the baptized, married and death parishioners of the various Passau's Diocese parishes. Provided by the Passau Diocesan Archives, a READ partner.

- *Bentham Papers*. XVIII-XIX century documents and drafts by Jeremy Bentham, with about 100 000 images, most casually written by many hands. Provided by the Universitu College of London (ULC, a READ partner) and the British Library (BL).

- *TSO collection* (Spanish Teatro del Siglo de Oro). XV-XVII century manuscripts of Spanish comedies, with more than 100 000 images, written by many hands. Provided by the Biblioteca Nacional de España (BNE) and PROLOPE, both READ MOU partners.

5 Preparatory Experiments

The preparatory processes needed to index each of the collections considered entailed more or less strightforward work devoted to data organization, definition of adequate transliteration tables, etc. and, most importantly, empirical work to estimate the expected search and retireval performance. These experiments and their results are described in Deliverable D7.15.

6 PWI of Selected Collections

From the collections considered in Sec. 4, two were finally selected for actual complete indexing; namely, Bentham Papers and TSO. While all the preparatory work was carried out for the Passau collection (see Deliverable D7.15), the decision to go ahead with the actual indexing of this collection was not made so far.

6.1 Bentham Papers PWI features

The whole collection, including all the 153 UCL and 20 BL "boxes" was succesfully indexed. The process required about 2 months of multi-core computation and the resulting probabilistic index contains about 198 million entries and requires about 10 gigabytes of storage. During this process, about 6 million lattices were generated, then used to compute the probabilistic index entries, and finally discarded. All in all, this workflow involved handling about 550 gigabytes of data during a time span of about 3 months.

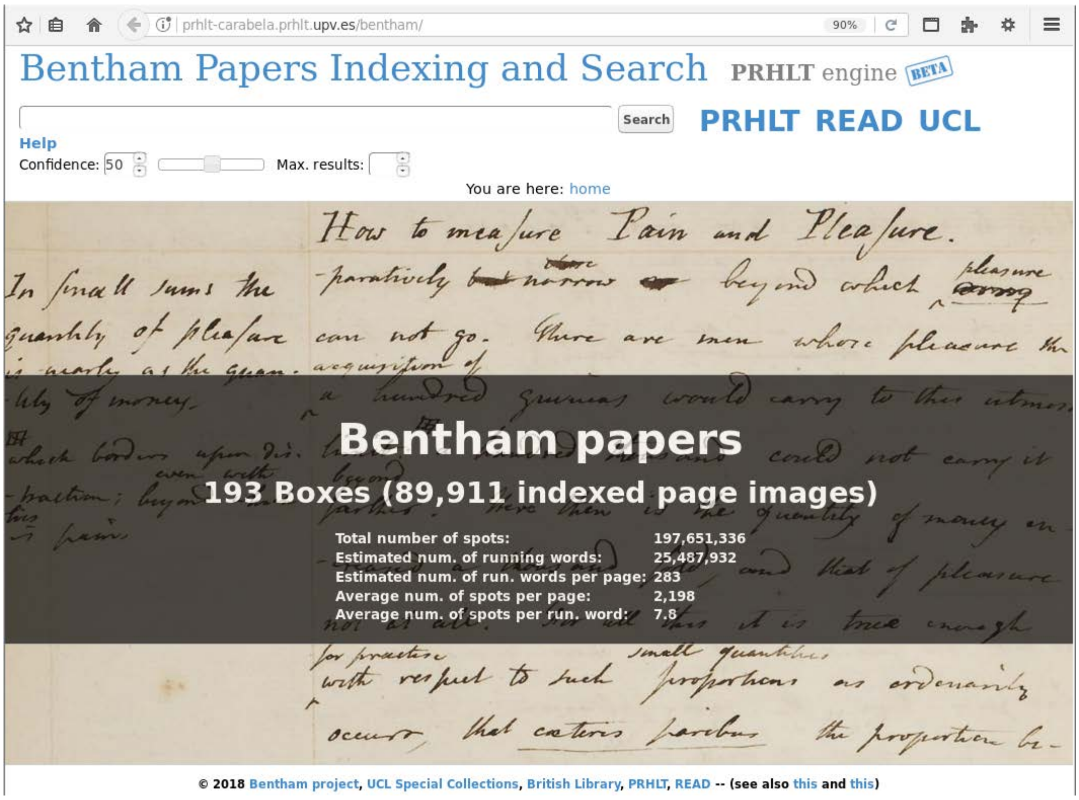
The following table summarizes the main features of the whole probabilistic word index produced for the Bentham Papers collection.

Probabilistic Index basic statistics (as of Nov-2018)

Computed:	
#Boxes	173
#Page images / *Indexed*	95,247 / *89,911*
#Spots	197,651,336
Average #Spots / Page	2,198

Estimated from index probabilities:	
Running words	25,487,932
Running words / Page	283
Average #Spots / Running word	7.8

The search and retrieval demonstrator for the 89,911 indexed page images of the full Bentham paper collection is publicly available at `http://prhlt-carabela.prhlt.upv.es/bentham`. Its front page is shown in Fig. 6



prhlt-carabela.prhlt.upv.es/bentham/

Bentham Papers Indexing and Search PRHLT engine BETA

Search

PRHLT READ UCL

Help

Confidence: 50

Max. results:

You are here: home

Bentham papers

193 Boxes (89,911 indexed page images)

```
Total number of spots:                       197,651,336
Estimated num. of running words:             25,487,932
Estimated num. of run. words per page:  283
Average num. of spots per page:              2,198
Average num. of spots per run. word:         7.8
```

© 2018 Bentham project, UCL Special Collections, British Library, PRHLT, READ -- (see also this and this)

Figure 6: Front page of the UPVLC PWI search and retrieval interface for Bentham Papers.

6.2 Spanish "Teatro del Siglo de Oro" (TSO) PWI features

All the 328 manuscripts (182 authored and 146 anonimous) provided by the BNE were successfully indexed. The process required about 4 months of multi-core computation and the resulting probabilistic index contains about 43 million entries and requires about 2 gigabytes of storage. All in all, this workflow involved handling about 50 gigabytes of data during the a time span of about six months.

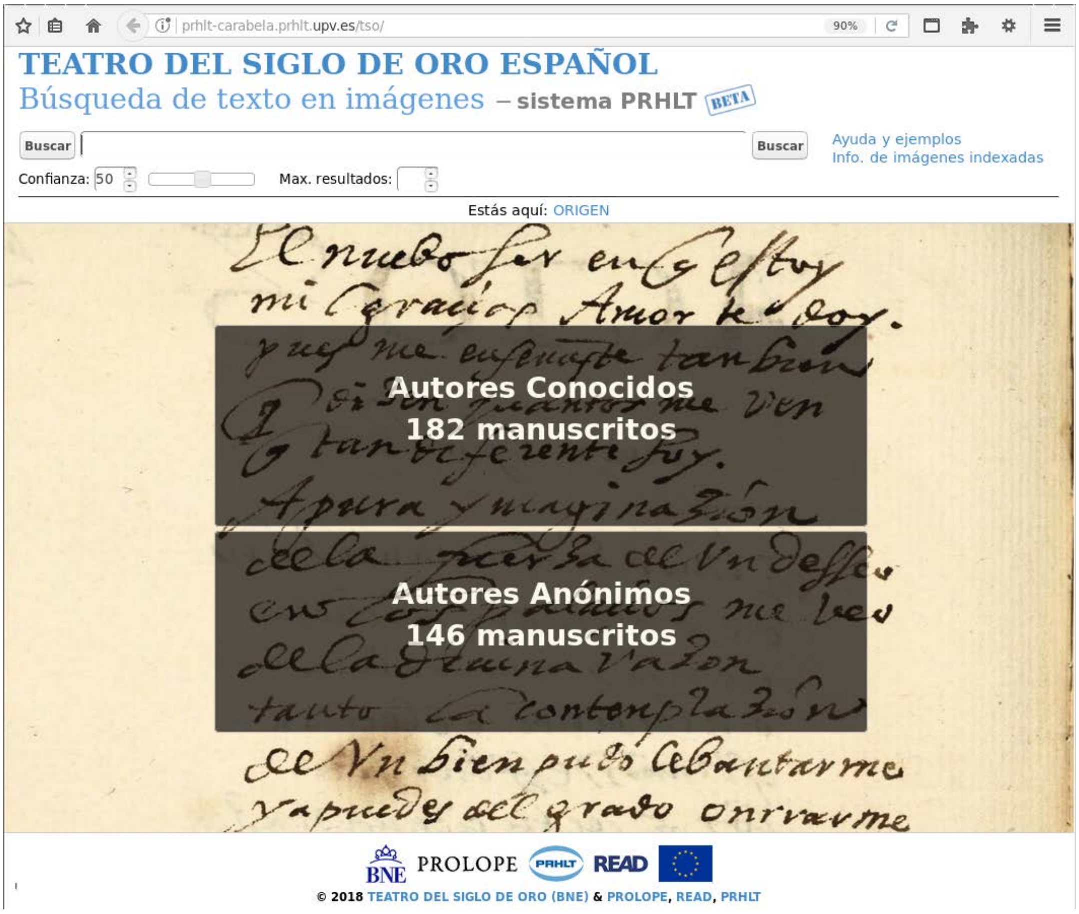
The following table summarizes the main features of the whole probabilistic word index produced for the TSO collection.

Probabilistic Index basic statistics (as of Nov-2018)

Computed:	
#Manuscripts	328
#Page images/*Indexed*	41,122 / *36,010*
#Spots	42,477,144
Average #Spots / Page	1,180

Estimated from index probabilities:	
Running words	5,396,497
Running words / Page	150
Average #Spots / Running word	7.9

The search and retrieval demonstrator for the 36,010 indexed page images of the full TSO collection is publicly available at `http://prhlt-carabela.prhlt.upv.es/tso`. Its front page is shown in Fig. 7



TEATRO DEL SIGLO DE ORO ESPAÑOL

Búsqueda de texto en imágenes – sistema PRHLT

Confianza: 50

Max. resultados:

Estás aquí: ORIGEN

Autores Conocidos
182 manuscritos

Autores Anónimos
146 manuscritos

© 2018 TEATRO DEL SIGLO DE ORO (BNE) & PROLOPE, READ, PRHLT

Figure 7: Front page of the UPVLC PWI search and retrieval interface for the Spanish TSO.

7 Usage Statistics of Large-Scale PWI Search and Retrieval Demonstartors

7.1 Spanish “Teatro del Siglo de Oro” (TSO) usage statisitcs

The number of queries honored per country are shown in Table 1. Queries from UPVLC computers are not inluded as most of them were made for debuging and testing purposes.

Table 1: TSO: Number of queries per country, 02/Nov–20/Dec, 2018

Queries	*Country*
19418	Spain
77	Austria
73	Belgium
32	United Kingdom
27	New Zealand
23	Argentina
17	France
15	USA
13	Netherlands
12	Greece
10	Switzerland
8	Hungary
8	Brazil
3	Portugal
19736	TOTAL

7.2 Bentham Papers usage statisitcs

The number of queries honored per country are shown in Table 2. Queries from UPVLC computers are not inluded as most of them were made for debuging and testing purposes.

Table 2: Bentham Papers: Number of queries per country, 08/Oct–20/Dec, 2018

Queries	*Country*
2054	Italy
1335	United Kingdom
903	Austria
511	Ireland
298	Netherlands
194	Germany
106	USA
102	Finland
86	Spain
85	Estonia
61	Argentina
41	France
40	New Zealand
36	Switzerland
21	Australia
12	China
10	Serbia
10	Greece
7	Belgium
7	Armenia
6	Japan
5	Croatia
4	Bulgaria
3	Sweden
3	Israel
3	Denmark
3	Canada
2	Russian Federation
1	Philippines
1	Chile
5950	TOTAL

References

[1] Jon Almazán, Albert Gordo, Alicia Fornés, and Ernest Valveny. Word spotting and recognition with embedded attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(12):2552–2566, 2014.

[2] Angelos P Giotis, Giorgos Sfikas, Christophoros Nikou, and Basilis Gatos. Shape-based word spotting in handwritten document images. In *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, pages 561–565. IEEE, 2015.

[3] Eva Lang, Joan Puigcerver, Alejandro Héctor Toselli, and Enrique Vidal. Probabilistic indexing and search for information extraction on handwritten german parish records. In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 44–49. IEEE, 2018.

[4] Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze, et al. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[5] Ramakrishnan Mukundan and KR Ramakrishnan. *Moment functions in image analysis theory and applications*. World Scientific, 1998.

[6] Ernesto Noya-García, Alejandro H Toselli, and Enrique Vidal. Simple and effective multi-word query spotting in handwritten text images. In *8th Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA)*, Faro, Portugal, 2017.

[7] Giorgos Sfikas, Angelos P Giotis, Georgios Louloudis, and Basilis Gatos. Using attributes for word spotting and recognition in polytonic greek documents. In *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, pages 686–690. IEEE, 2015.

[8] Alejandro H Toselli, Enrique Vidal, Joan Puigcerver, and Ernesto Noya-García. Probabilistic multi-word spotting in handwritten text images. *Pattern Analysis and Applications*, pages 1–10, 2018.

[9] Alejandro H Toselli, Enrique Vidal, Verónica Romero, and Volkmar Frinken. HMM word graph based keyword spotting in handwritten document images. *Information Sciences*, 370-371:497–518, 2016. Information Sciences 370-371 (2016) 497-518.

[10] Tijn Van der Zant, Lambert Schomaker, and Koen Haak. Handwritten-word spotting using biologically inspired features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):1945–1957, 2008.

[11] Enrique Vidal, Alejandro H Toselli, and Joan Puigcerver. High performance query-by-example keyword spotting using query-by-string techniques. In *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, pages 741–745. IEEE, 2015.

[12] Konstantinos Zagoris, Ioannis Pratikakis, and Basilis Gatos. A framework for efficient transcription of historical documents using keyword spotting. In *Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing*, pages 9–14. ACM, 2015.