

IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands.

IMPROVING ACCESS TO TEXT

Centre of Competence for mass digitisation of text material

IMPACT Project Office, 2010

Er is door mij gebruik gemaakt van dia's uit presentaties van o.a.

Anastasios Kesidis, CIL,
Athene Griekenland,

en
Asaf Tzadok, IBM Haifa
Research Lab

Paul Doorenbosch

Koninklijke Bibliotheek

paul.doorenbosch@kb.nl

Werkbijeenkomst Digitalisering
Historische Handschriften,

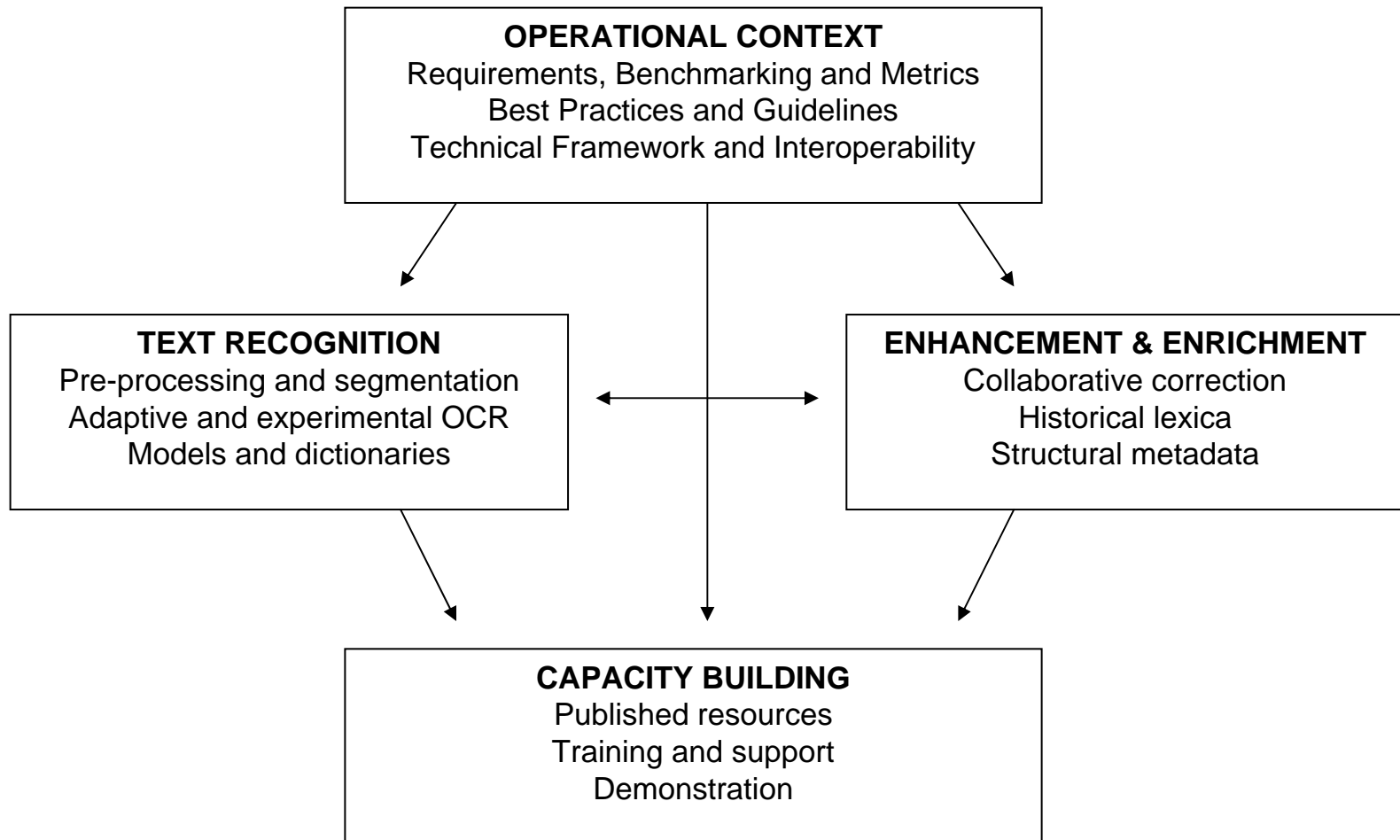
Leuven, 01-07-2010

IMPACT objectives

Significantly improve mass digitisation of historical printed text by:

- Innovating OCR software and language technology
- Sharing expertise and building capacity across Europe
- Ensuring that tools and services will be sustained after the end of the project

IMPACT Project Architecture



Tools for Text Recognition (OCR)

Technologies for the extraction of text in a digital form from the page

- **Adaptive OCR engine:** Cutting-edge software system which adapts itself to the material during OCR process, making use of the results of several other tools:
 - Image enhancement toolkit
 - Segmentation toolkit
 - Post-correction modules

- **Experimental prototypes and tools:** Inventory extraction, Typewritten OCR, Word Spotting



Tools for Enrichment (language technology)

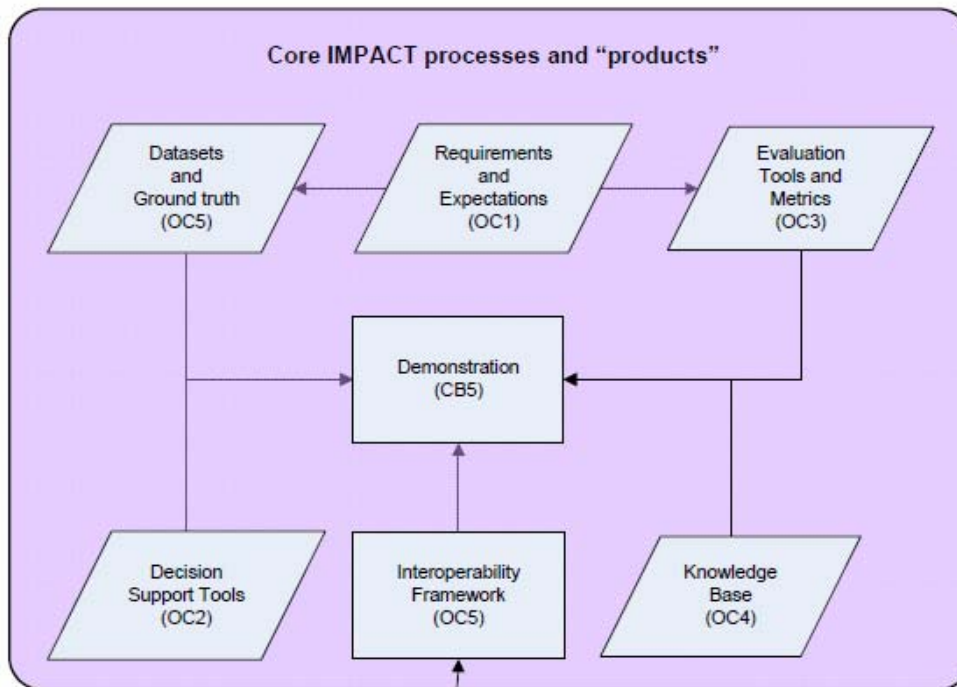
Make the OCR results more accurate and more accessible

- **Collaborative correction**
 CONCERT: Web-based platform where volunteers can validate and correct OCR results.
- **Historical lexica**
 - General and Named Entities lexica for Dutch, German, English
 - Lexica for Spanish, French, Polish, Bulgarian, Czech and Slovene
 - Toolboxes for historical lexicon building
 - Web-based workspace for named entity management
- **Structural metadata**
 Functional Extension Parser: a set of web services for automatically detecting and tagging structural metadata of scanned material

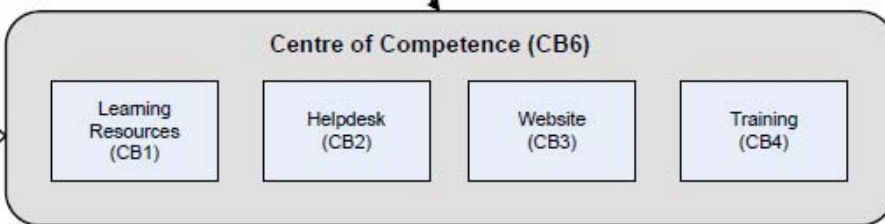
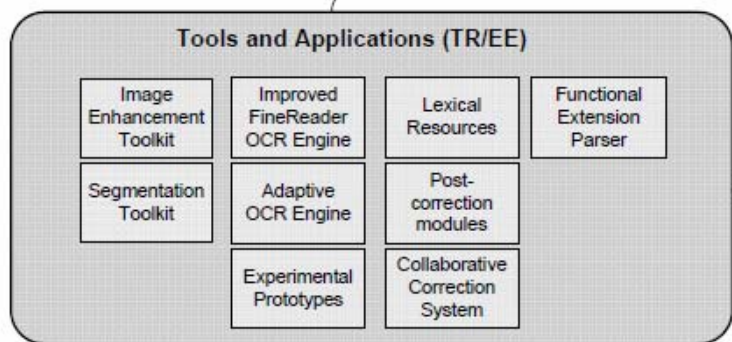
IMPACT tool: CONCERT

Collaborative Engine for Correction of Extracted Text

- Web-based platform to validate and correct OCR results
 - Volunteers identify incorrect words or characters with a single mouse click
 - Separate session for different aspects of data entry:
 - Character/symbol level verification (carpets)
 - Word level data entry (characters in context)
 - Page level
 - Results from the correction session feed back into the OCR machine that will adapt to this new information
- Enables the general public to help with mass digitisation efforts



The IMPACT "products":
 - Tools and applications, IIF
 - Datasets and ground truth
 - Guidelines and support



Experimental OCR engines: Word Spotting

IMPACT TR4

Anastasios Kesidis

*Computational Intelligence Laboratory
Institute of Informatics and Telecommunications
National Center for Scientific Research "Demokritos"
GR-153 10 Agia Paraskevi, Athens, Greece*

Uit: IMPACT: description of work TR4

A word spotting engine capable of searching words in texts which cannot be OCR processed in the traditional sense of the word.

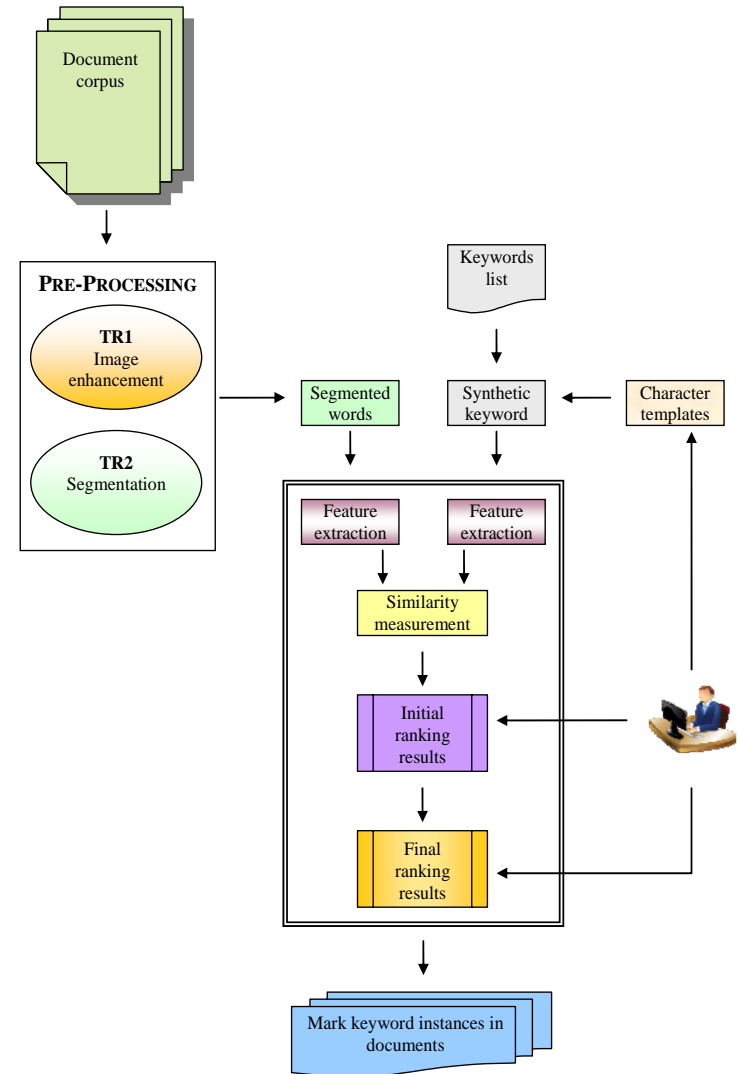
Traditional approaches in document indexing usually involve an Optical Character Recognition step which performs well in modern printed documents and documents of high quality printing. In the case of printed historical documents OCR, several factors affect the final performance like low paper quality, paper positioning variations, low print contrast, large variety of fonts and typesetting imperfections. Usually, printed OCR systems involve a character segmentation step followed by a recognition step using pattern classification algorithms. Due to document degradations, conventional OCR systems often fail to support a correct segmentation of the printed historical documents into individual characters.

We propose to develop an alternative technique for historical document indexing based on spotting words directly on document images with the help of word matching while avoiding conventional OCR procedure. A similar approach has been already applied with success to the handwritten historical documents of the Library of Congress, while an application to Old Greek books has been proposed by NCSR. Word segmentation is first applied to all document pages and then a spotting of the most interesting words (keywords) is applied directly on the document images. The keywords may be important names, places, terms etc

Architecture overview

The main operational parts of the Word Spotting engine are:

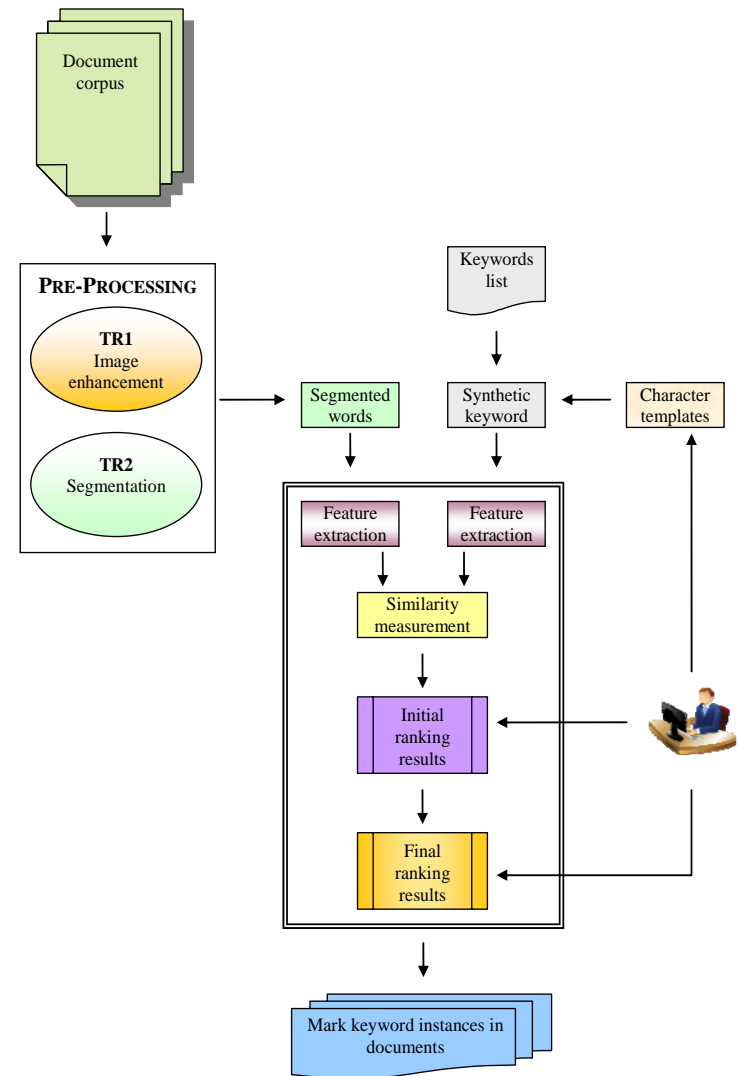
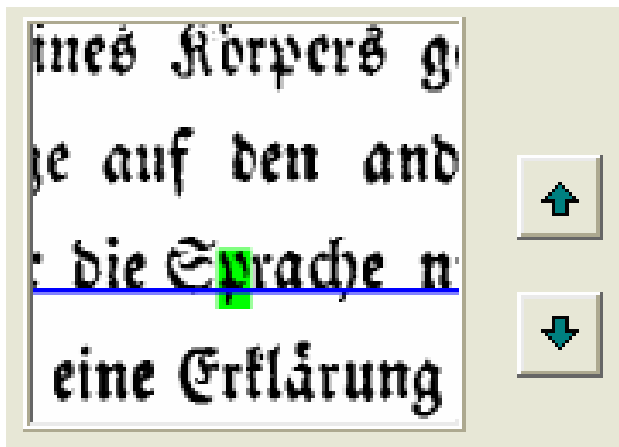
- Marking character templates
- Feature extraction & word matching
- User feedback
- Searching
- User access control



Word Spotting engine

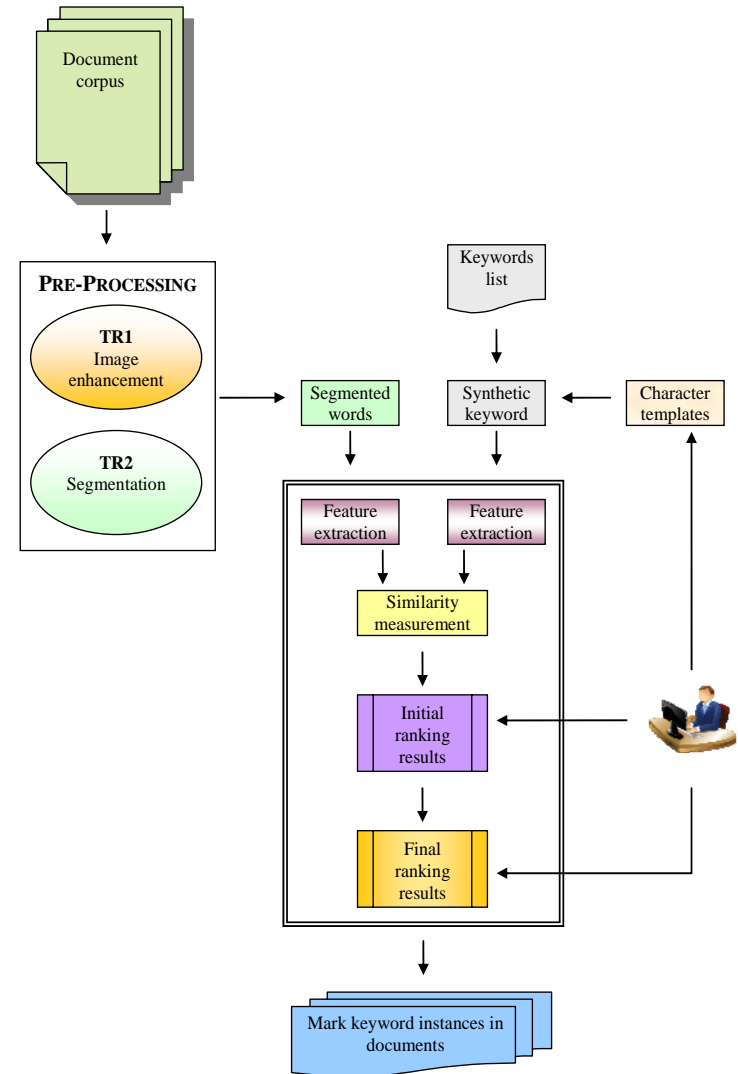
➤ Marking character templates

- Applied directly on a text image
- Character baseline adjustment
- Performed “once-for-all” and can be used for entire books or collections with similar text characteristics



Word Spotting engine

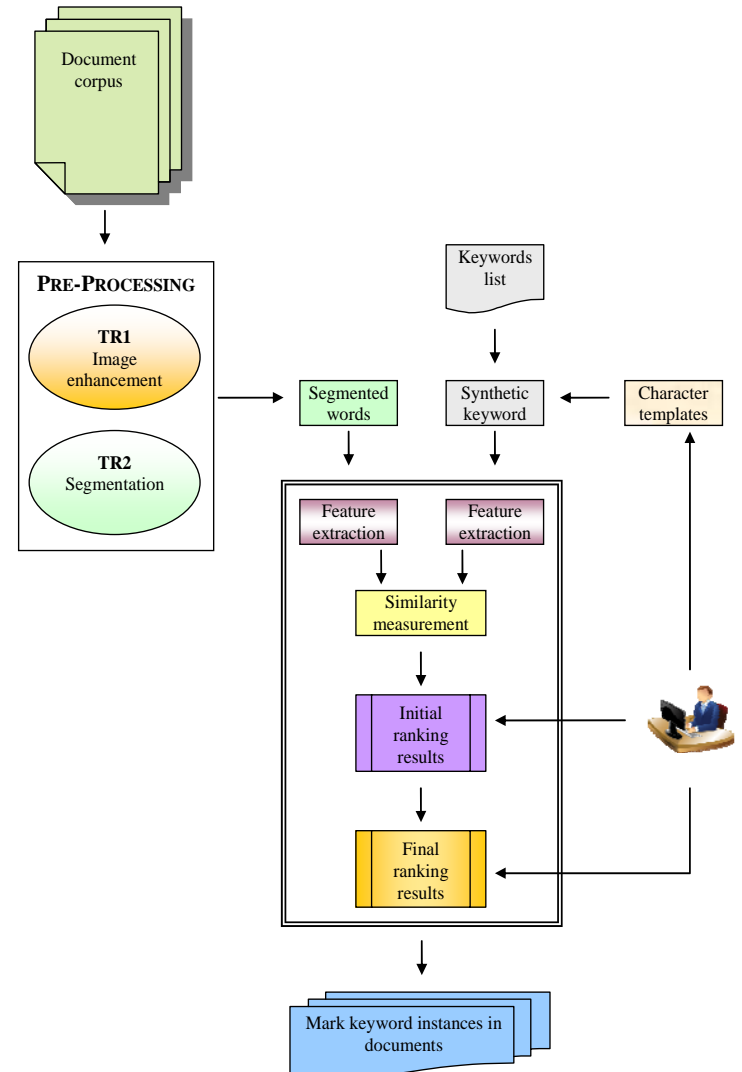
- Feature extraction & word matching
 - Describe each word (synthetic or real) by a set of features
 - Normalize
 - Match by checking similarity based on features or directly on word images



Word Spotting engine

➤ User feedback

- Produce an initial ranking list comparing the synthetic word to the segmented words
- The user selects as input query one or more correct results from the list
- A new matching process is initiated. The segmented words are ranked according to their similarity to the selected word(s) which, in this case, are not synthetic but real words of the document's corpus.



Word Spotting engine

➤ Searching

- Allow the user to search the image corpus for instances of query keywords that have already undergone the user feedback spotting process.
- The user selects one of the processed keywords and the application shows all the instances of this keyword in the images of the corpus.
- The user can navigate through the results in an instance level (showing one instance per time) or in a page level (showing all instances in a page).

27

Ablick einer Rose; so unterschieden die Menschen auch ihre Zeichen, denn die **Eprache** durch Zeichen führte erst zu der **Eprache** durch Töne.

Wäre der Mensch nicht durch die gröbere Organisation seines Körpers gehindert, die feinsten Eindrücke der Dinge auf den andern Menschen zu bemerken, so würde er die **Eprache** nicht nötig haben, die nichts anders als eine Erklärung der Eindrücke ist.

Es giebt eine engere Verbindung des menschlichen Geistes und Herzens, und Anschaulichkeit ist ihre **Eprache** —

Es giebt zwei **Eprachen**, die **Eprache** des Herzens, und die **Eprache** des Verstandes. Die **Eprache** des Herzens ist weniger dem Betrage, als die **Eprache** des Verstandes unterworfen. Die **Eprache** des Herzens schöpft ihre Bilder aus Empfindungen; die **Eprache** des Verstandes aus Worten von Empfindungen.

Die **Eprache** des Herzens hat wenig Worte, und sagt vieles; die **Eprache** des Verstandes hat viele Worte, und sagt oft wenig.

Je mehr eine **Eprache** Worte hat, desto unvollkommener ist sie, die **Eprache**; denn sie ist desto eher dem Irrthum unterworfen.

Die Hieroglyphen der Älten waren eine anschauliche **Eprache**, eine **Eprache** des Auges. Die Chineser haben so viel Buchstaben als Worte — —

Das Ohr ist der schwächste und betrüglichste Sinn,
und

Development plan

➤ D-TR4.3 Word spotting prototype (M36)

Internal deadlines:

- **Word Spotting specifications & architecture document (M18)**
- **Internal version with minimum functionality (M19)**
- **1st version of the Word Spotting Prototype (M22)**
- **Evaluation of the 1st version (M24)**
- 2nd version of the Word Spotting Prototype (M34)
- Evaluation of the 2nd version - Release of the final prototype (M36)
 (= einde 2010)



IBM Labs in Haifa



Adaptive OCR

Asaf Tzadok
IBM Haifa Research Lab



Digitization Process being addressed

- ◇ Automatic Text Recognition
- ◇ Collaborative tools for effective verification/correction of the automatic recognition results
- ◇ Use of historic lexicons in order to facilitate above processes

- ◇ Remark: scanned process addressed indirectly via the Digitization Support Centers

Technical Objectives

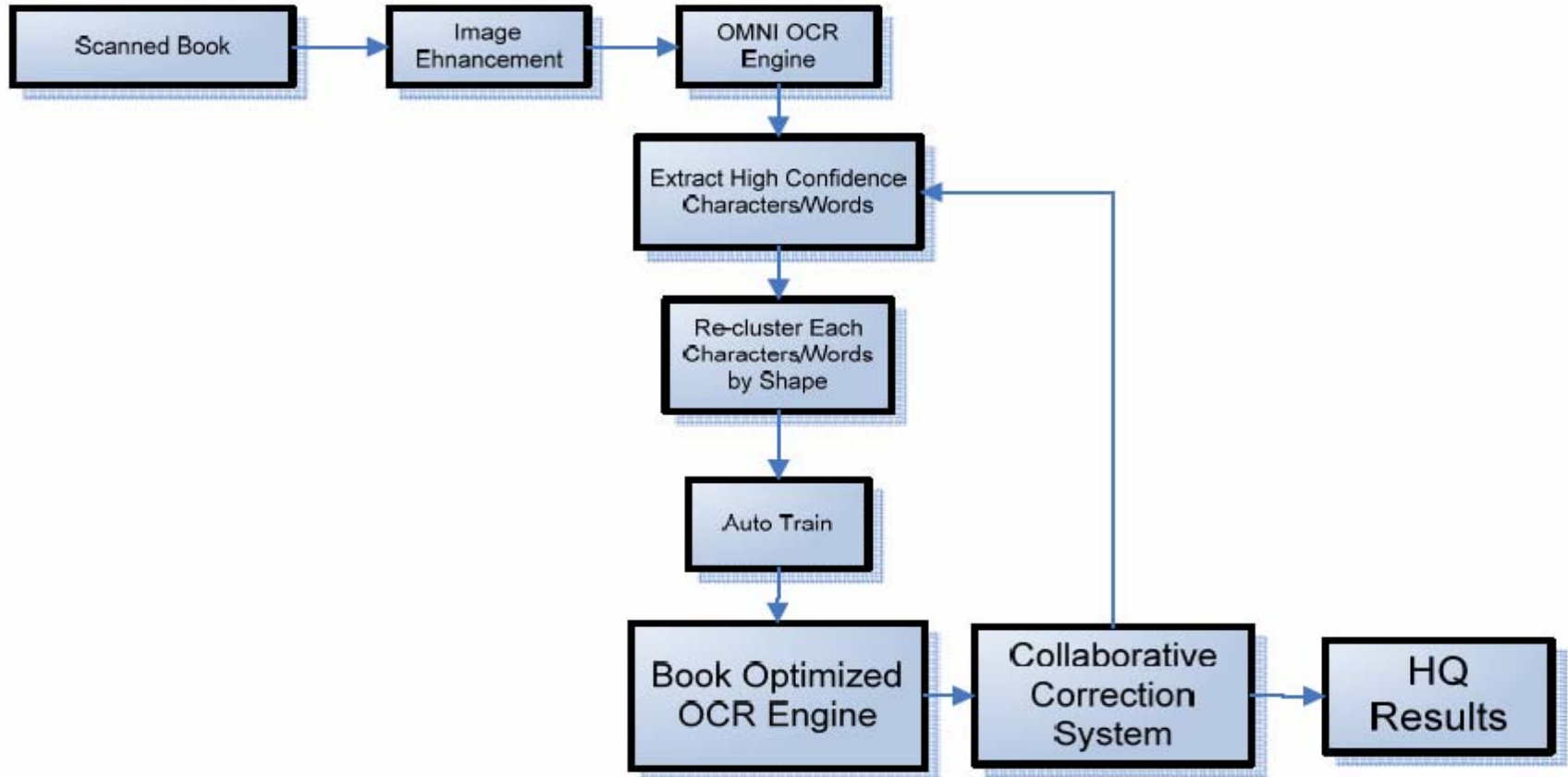
- ◆ **Improved Optical Character Recognition:**
 via Adaptive OCR engine tailored specifically to the needs of libraries, integrating several other tools (image enhancement & segmentation toolkits, post-correction modules and other OCR engines)
- ◆ **Text enhancement and enrichment tools**
 aimed at making the OCR results more accurate and more accessible including:
 - ◆ **Web based Collaborative correction system** suitable for massive volunteer participation (validates and corrects OCR results, first tool of its kind to be directly linked to an OCR engine)
 - ◆ **Lexicons and gazetteers** (General and Named Entities lexica for Dutch, German and English as well as support for lexicon development in other European languages)
 - ◆ **Structural metadata**

General Concept

- ◇ **The system works on a large bodies of homogenous material (if text contains several languages and/or several fonts adaptation would be applied in parallel to each segment)**
- ◇ **The system would work with manual correction of the OCR results (although similar approach can be applied to purely automatic applications)**
- ◇ **OCR would receive results of manual correction and use them in order to improve recognition results (error rate for the last page would be much better than that for the first page)**
- ◇ **Adaptation would be transparent to the user**



Adaptive System Architecture





IBM Labs in Haifa



Collaborative Correction

:

Asaf Tzadok
IBM Haifa Research Lab

Collaborative Correction

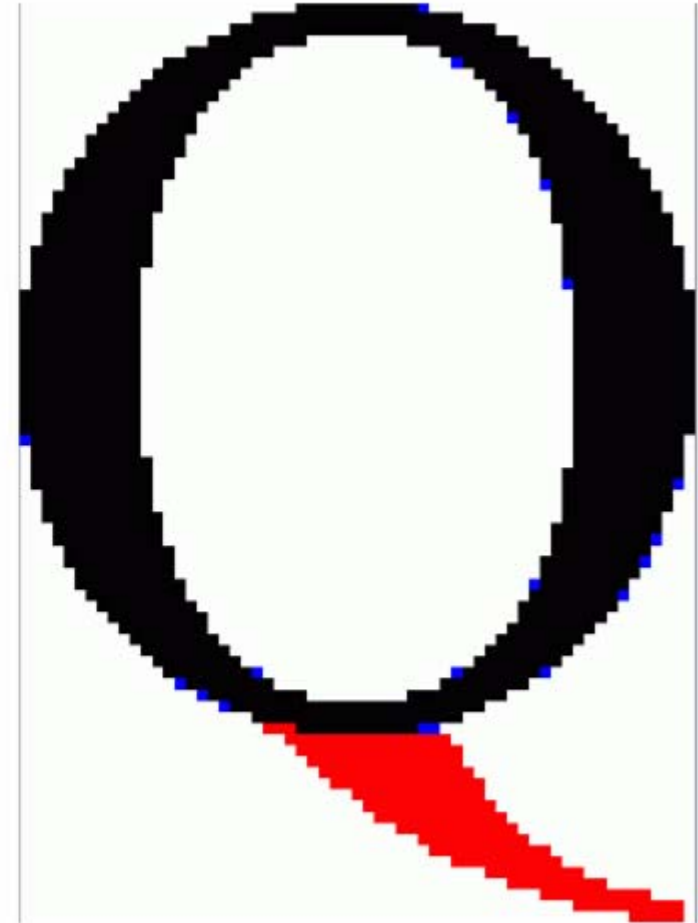
- ◇ A full Web based system
 - ◇ suitable for the wide public use
- ◇ Goal: mobilizing vast army of volunteers keen on contributing to their cultural heritage preservation
- ◇ Wikipedia type model adapted to the digitization domain

System Structure: Separate session for different aspects of data entry

- ◇ Character/symbol level verification (carpets)
- ◇ Character level data entry
- ◇ Word level data entry (characters in context)
- ◇ Page level

Carpet Session – Super Key approach

- ◇ Large-scale characters repository
- ◇ Similar shape and classification
- ◇ Minimization of the number of thumbnails for operator review
- ◇ Preserve important difference
 - ◇ Red color
- ◇ Ignore insignificant difference
 - ◇ Blue color
- ◇ Hierarchical approach



Word Session

- ◇ Suspected words will be verified by an operator
 - ◇ Speller based decision, no match in vocabulary
 - ◇ Low confidence words
- ◇ The operator will get a list of words
 - ◇ Same classification
 - ◇ Similar shape
 - ◇ A list of optional classifications
- ◇ The operator can choose
 - ◇ Accept classification
 - ◇ Choose alternative
 - ◇ Type the correct classification

Page OCR Editor/Verifier

- ◇ Page level decision
 - ◇ Avoid word segmentation error
 - ◇ Simulate existing solutions
 - ◇ Full cycle verification
 - ◇ Final compilation
 - ◇ Verify missing content
- ◇ Complex operations for character level truthing, such as:
 - ◇ Join/Split Characters (segmentation correction)
 - ◇ Insert missing content
 - ◇ Resize rectangles for segmentation inspection

Future Extension

- ◇ Super Key
 - ◇ Group similar symbols into one
- ◇ Search over OCR
 - ◇ Using unique probabilistic approach
- ◇ Auto reproduced eBook
 - ◇ Using auto-vectorized book font
 - ◇ HQ results with location for each word
- ◇ Page Distortion Estimation
 - ◇ Using words distortions