

# *tranScriptorium*



## **Workshop: Automatisierte Handschriftenerkennung**

**Joan Andreu Sánchez**

**Pattern Recognition and Human Language Research group  
(Technical University of Valencia)**

**Günter Mühlberger, Sebastian Colutto, Philip Kahle  
Digitisation and Digital Preservation group  
(University of Innsbruck)**

# Agenda



- Part 1: Handwritten Text Recognition (HTR)
- Part 2: Introduction to TRANSKRIBUS  
(Transcription and Recognition Platform)
- Part 3: Introduction to the expert GUI of  
TRANSKRIBUS
- Part 4: Hands-on-Session
  - Discussion

# Interactive Handwritten Text Recognition and Indexing of Historical Documents: the tranScriptorium Project

V. Romero, A.H. Toselli, M. Villegas, J.A. Sánchez, E. Vidal  
{vromero, ahector, mvillegas, jandreu, evidal}@prhlt.upv.es

Presenter: Joan Andreu Sánchez

*Pattern Recognition and Human Language Technology  
Research Center*



*tranScriptorium*



Universitat Politècnica de València (Spain)

February 2015

DHd-Tagung 2015

# Index

- 1 Handwritten Text Recognition (HTR) and Indexing ▷ 1
- 2 The tranScriptorium Project ▷ 3
- 3 Selected Handwriting Datasets ▷ 5
- 4 HTR and Interactive-Predictive HTR ▷ 10
- 5 Interactive HTR: Transcription Demonstration ▷ 12
- 6 HTR and Interactive-Predictive HTR Results ▷ 14
- 7 Handwritten Text Images Indexing: Search Demonstration ▷ 16
- 8 Handwritten Text Images Indexing and Search Results ▷ 19
- 9 Conclusion ▷ 22

# Index

- 1 *Handwritten Text Recognition (HTR) and Indexing* ▷ 1
- 2 The tranScriptorium Project ▷ 3
- 3 Selected Handwriting Datasets ▷ 5
- 4 HTR and Interactive-Predictive HTR ▷ 10
- 5 Interactive HTR: Transcription Demonstration ▷ 12
- 6 HTR and Interactive-Predictive HTR Results ▷ 14
- 7 Handwritten Text Images Indexing: Search Demonstration ▷ 16
- 8 Handwritten Text Images Indexing and Search Results ▷ 19
- 9 Conclusion ▷ 22

# Handwritten Text Recognition (HTR) and Indexing

Huge amounts of handwritten historical documents are being published by on-line digital libraries world wide

However, for these raw digital images to be really useful, they need be annotated with informative content

*This presentation introduces efficient solutions for the indexing, search and full transcription of historical handwritten document images*

# Index

- 1 Handwritten Text Recognition (HTR) and Indexing ▷ 1
- 2 *The tranScriptorium Project* ▷ 3
- 3 Selected Handwriting Datasets ▷ 5
- 4 HTR and Interactive-Predictive HTR ▷ 10
- 5 Interactive HTR: Transcription Demonstration ▷ 12
- 6 HTR and Interactive-Predictive HTR Results ▷ 14
- 7 Handwritten Text Images Indexing: Search Demonstration ▷ 16
- 8 Handwritten Text Images Indexing and Search Results ▷ 19
- 9 Conclusion ▷ 22

# The tranScriptorium Project

<http://www.transcriptorium.eu>

- STREP of the FP7 in the ICT for Learning and Access to Cultural Resources challenge (1 January 2013 to 31 December 2015)
  - *tranScriptorium* aims to develop innovative, efficient and cost-effective solutions for the indexing, search and full transcription of historical handwritten document images, using modern, holistic Handwritten Text Recognition technology
1. **Enhancing HTR technology for efficient transcription**
  2. **Bringing the HTR technology to users**
  3. **Integrating the HTR results in public web portals**



Supported by:



EU Cultural Heritage:



# Index

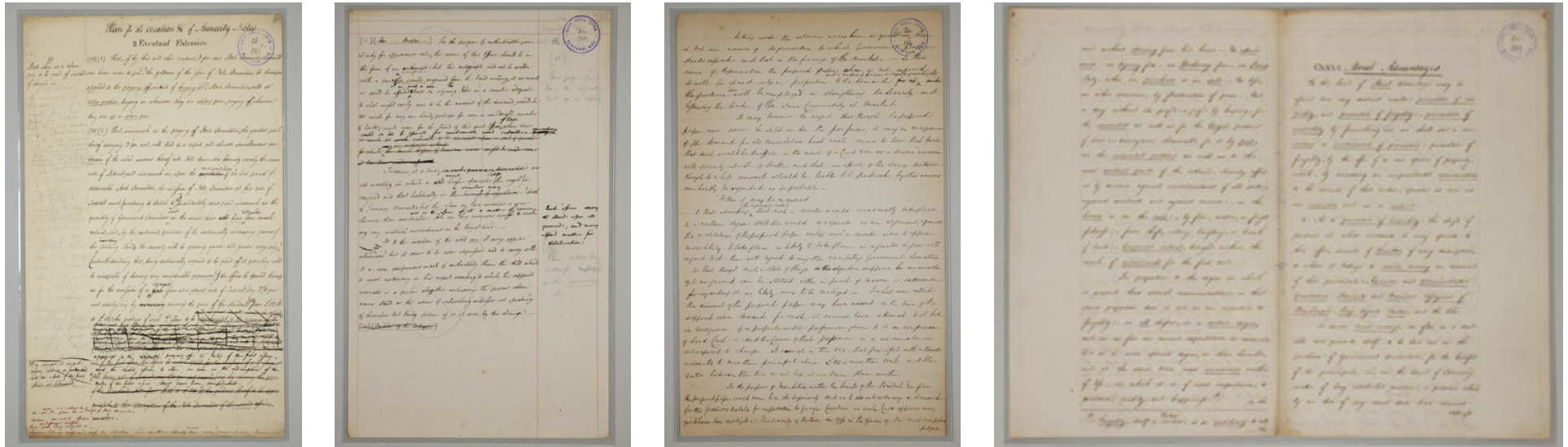
- 1 Handwritten Text Recognition (HTR) and Indexing ▷ 1
- 2 The tranScriptorium Project ▷ 3
- 3 *Selected Handwriting Datasets* ▷ 5
- 4 HTR and Interactive-Predictive HTR ▷ 10
- 5 Interactive HTR: Transcription Demonstration ▷ 12
- 6 HTR and Interactive-Predictive HTR Results ▷ 14
- 7 Handwritten Text Images Indexing: Search Demonstration ▷ 16
- 8 Handwritten Text Images Indexing and Search Results ▷ 19
- 9 Conclusion ▷ 22

## Selected Handwriting Datasets

- **BENTHAM**: XVIII/XIX centuries collection of over 4, 000 pages of drafts and notes, written by several hands in English
- **PLANTAS**: XVII century botanical specimen manuscript collection of seven volumes written by a single hand in Old Spanish – kindly provided by the BNE
- **HATTEM**: XV century Medieval Manuscript composed of 573 sheets written by a single hand in Dutch
- **ESPOSALLES**: XVII century Marriage License records written by several hands in old Catalan and other languages
- **AUSTEN**: XVIII century Juvenilia manuscripts by Jane Austen (single hand in English) – kindly provided by the BL
- **REICHSGERICHT**: early XX century manuscripts of court decisions written by a several hands in German

# “BENTHAM” Dataset

XVIII century collection of over 4,000 sheets of drafts and notes, written by several writers in English



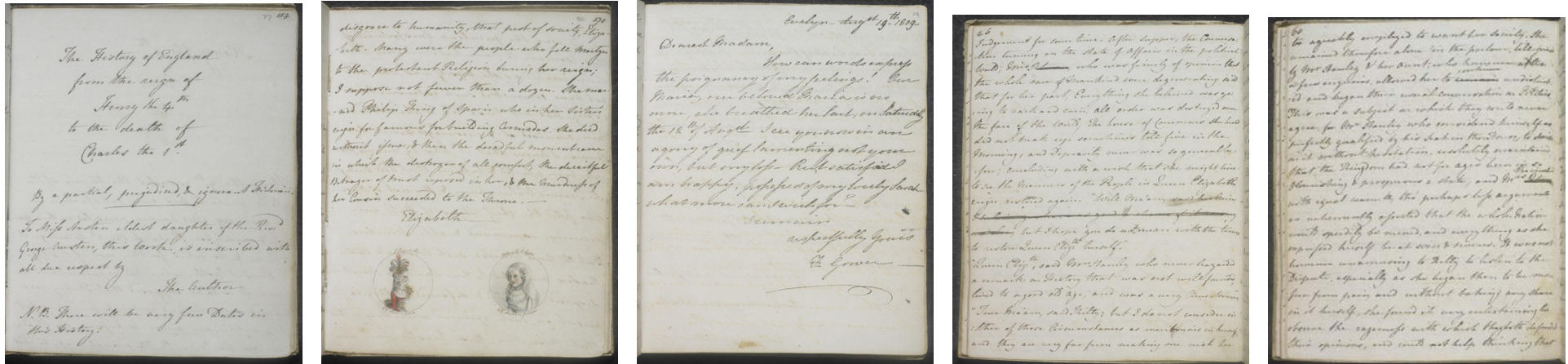
Experiments on a first batch of 433 pre-selected page images

version of the whole amount thereof into Note Annuities bearing nearly the same rate of Interest; and inasmuch as, upon the <sup>redemption</sup> completion of the last parcel of redeemable Stock Annuities, the emission of Note Annuities at this rate of Interest must precede to Article 6<sup>th</sup> immediately <sup>and</sup> <sup>and</sup> inasmuch as the quantity of Government Annuities in the mean time <sup>will</sup> <sup>will</sup> have been much reduced, and, by the continued operation of the continually increasing power of the sinking fund, the scarcity will be growing greater and greater every day, <sup>in consequence</sup> notwithstanding that, being continually exposed to be paid off at par, they will

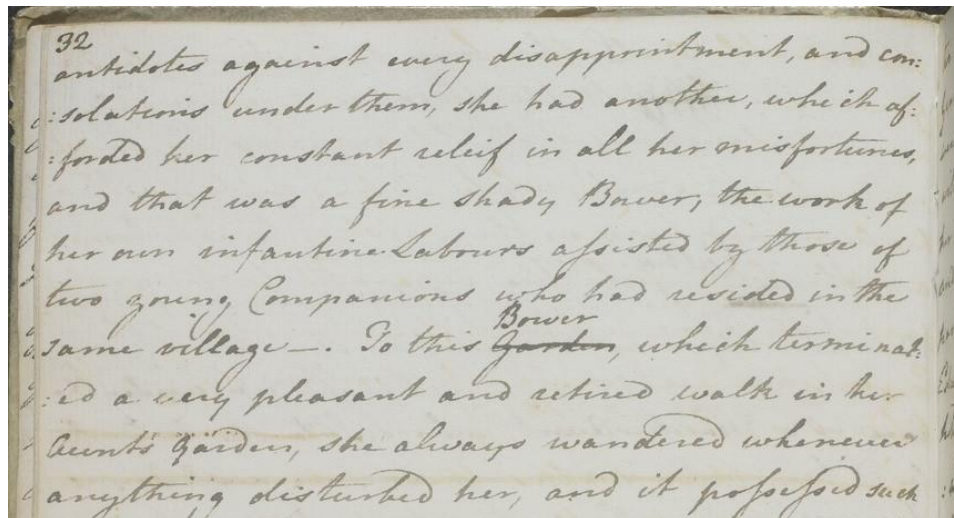
Number of:	Total
Pages	433
Lines	11 473
Running words	106 905
Lexicon size	9 717
Running characters	550 674
Character set size	86

# “AUSTEN” Dataset

## Jane Austen’s *Juvenilia*: XVIII century single hand manuscript



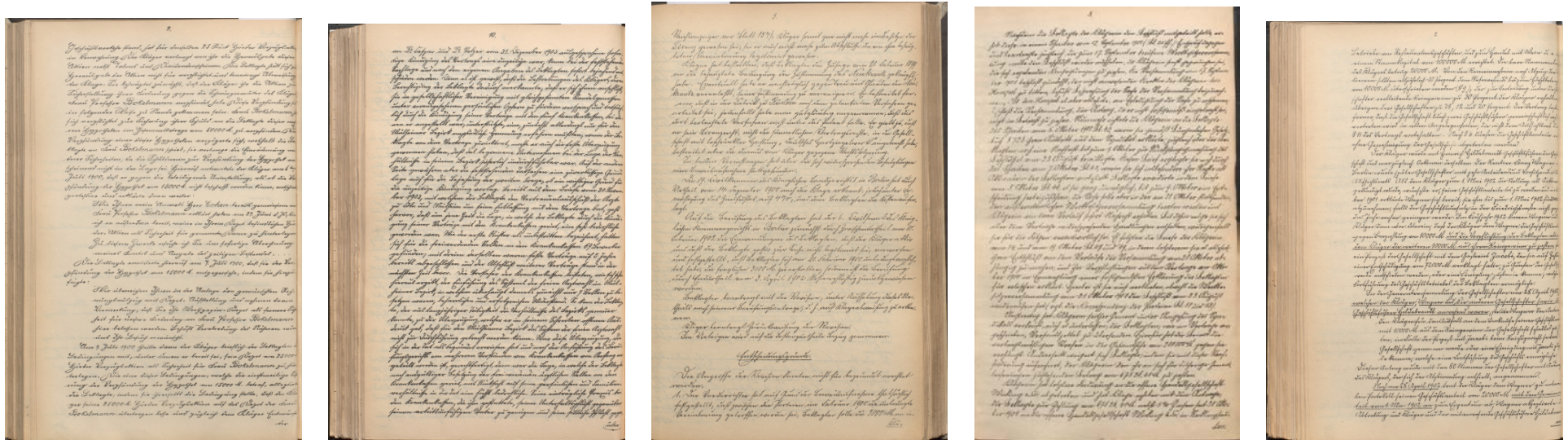
## Experiments on Volume The Third



Number of:	Total
Pages	128
Lines	2 693
Running words	25 291
Dataset lexicon	3 567
Running characters	118 881
Character set size	81

# “REICHSGERICHT” Dataset

Court decisions from the German High Court from 1900-1914.



Experiments on a first batch of 114 pre-selected page images

*Handwritten text from a court decision, showing dense cursive script.*

Number of:	Total
Pages	114
Lines	4 046
Running words	40 566
Dataset lexicon	6 098
Running characters	251 813
Character set size	92

# Index

- 1 Handwritten Text Recognition (HTR) and Indexing ▷ 1
- 2 The tranScriptorium Project ▷ 3
- 3 Selected Handwriting Datasets ▷ 5
- 4 *HTR and Interactive-Predictive HTR* ▷ 10
- 5 Interactive HTR: Transcription Demonstration ▷ 12
- 6 HTR and Interactive-Predictive HTR Results ▷ 14
- 7 Handwritten Text Images Indexing: Search Demonstration ▷ 16
- 8 Handwritten Text Images Indexing and Search Results ▷ 19
- 9 Conclusion ▷ 22

# HTR and Interactive-Predictive HTR

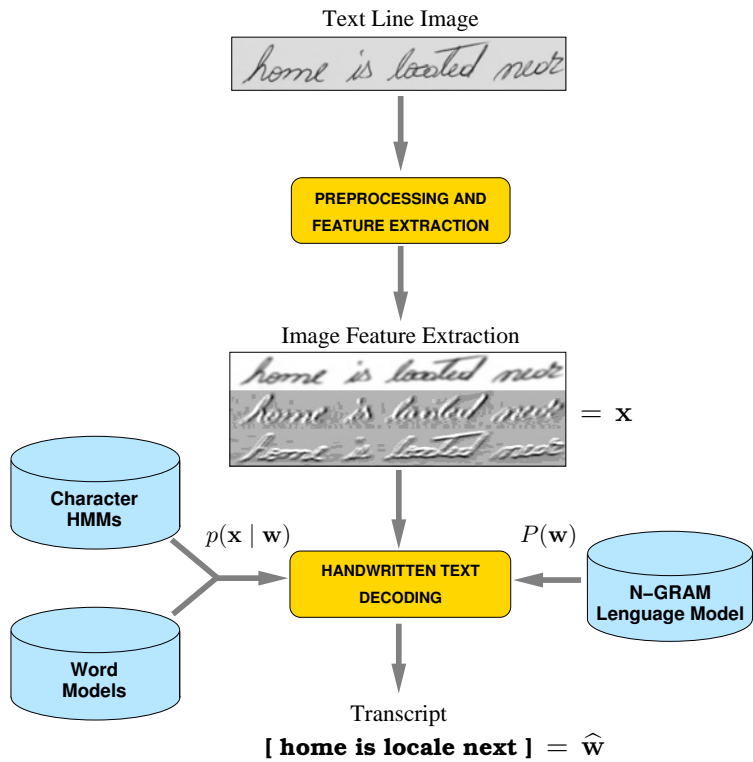
*HTR current state-of-the-art:*

- Segmentation-free approach: no explicit segmentation of text images into words or characters is required
- The basic input unit is a handwritten text line image
- Statistical modeling at different perception levels:
  - Optical (character shape), using Hidden Markov Models (HMMs)
  - Lexical, by means of finite-state character representation of words
  - Syntactical, based on statistical language models, such as  $N$ -grams

*Interactive-predictive* framework: rather than full transcription automation, the system *assists* the human transcriber

- Combines HTR efficiency with the accuracy of human experts, leading to cost-effective perfect transcripts

# HTR Architecture



- **Preprocessing**

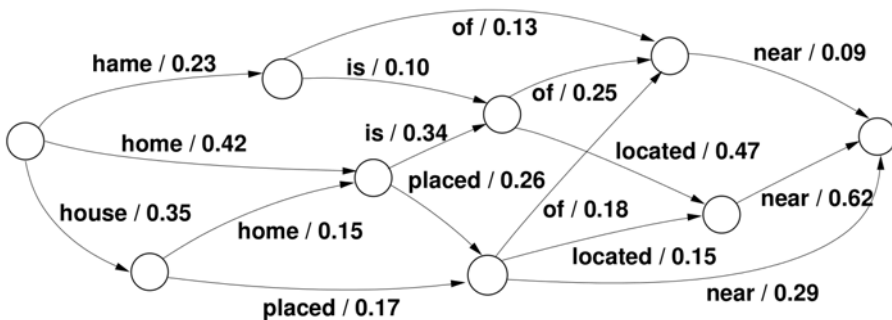
- **Feature Extraction**

$$\mathbf{x} = \vec{x}_1, \vec{x}_2, \dots, \vec{x}_n, \vec{x}_i \in \mathbb{R}^D$$

- **Decoding**

$$\begin{aligned} \hat{\mathbf{w}} &= \arg \max_{\mathbf{w}} p(\mathbf{w} | \mathbf{x}) \\ &= \arg \max_{\mathbf{w}} p(\mathbf{x} | \mathbf{w}) \cdot P(\mathbf{w}) \end{aligned}$$

## Word Graph



Huge sets of  $N$ -best hypotheses can be arranged into a *Word Graph* or *Lattice*.

# Index

- 1 Handwritten Text Recognition (HTR) and Indexing ▷ 1
- 2 The tranScriptorium Project ▷ 3
- 3 Selected Handwriting Datasets ▷ 5
- 4 HTR and Interactive-Predictive HTR ▷ 10
- 5 *Interactive HTR: Transcription Demonstration* ▷ 12
- 6 HTR and Interactive-Predictive HTR Results ▷ 14
- 7 Handwritten Text Images Indexing: Search Demonstration ▷ 16
- 8 Handwritten Text Images Indexing and Search Results ▷ 19
- 9 Conclusion ▷ 22

# Interactive HTR: Transcription Demonstration

- It is just a “*demo*” ! not intended for real operation (other systems do that)
- Everything is *real*. No tricks to make demo look better than real
- Web client-server architecture: Web browser front-end, back-end server providing off-line HTR-CATTI
- Off-line HTR-CATTI decoder based on word graphs
- Three tasks:
  - BENTHAM: 78K words open vocabulary
  - AUSTEN: 78K words external, open vocabulary from Bentham texts  
20K words external, open vocabulary from Austen texts
  - REICHSGERICHT:  
6K words open vocabulary

# Index

- 1 Handwritten Text Recognition (HTR) and Indexing ▷ 1
- 2 The tranScriptorium Project ▷ 3
- 3 Selected Handwriting Datasets ▷ 5
- 4 HTR and Interactive-Predictive HTR ▷ 10
- 5 Interactive HTR: Transcription Demonstration ▷ 12
- 6 *HTR and Interactive-Predictive HTR Results* ▷ 14
- 7 Handwritten Text Images Indexing: Search Demonstration ▷ 16
- 8 Handwritten Text Images Indexing and Search Results ▷ 19
- 9 Conclusion ▷ 22

## HTR and Interactive-Predictive HTR Results

- BENTHAM: Training: OMs with 400 pages, LM Lex. 78K words. Test: 33 pages.

WER = 22.0 %    WSR = 17.2 %    EFR: 21.5 % wrt post-edit  
CER = 9.9 %

- AUSTEN: *No training*; just using Bentham models

WER = 45.0 %    WSR: 27.5 %    EFR: 38.9 % wrt post-editing  
CER = 25.5 %

AUSTEN: Training: OMs with 50 pages, LM Lexicon 20K words. Test: 78 pages

WER = 32.2 %    WSR = 21.4 %    EFR: 33.5 % wrt post-editing  
CER = 15.9 %

- REICHSGERITICH: Training: OMs with 88 pages, LM Lex. 6K words. Test: 26 pag.

WER = 33.3 %    WSR: 25.1 %    EFR: 24.6 % wrt post-editing  
CER = 14.5 %

WER/CER: percentage of mis-recognized words/characters.

Experiments with *open-vocabulary* lexica and bi-gram LMs.

WSR = Percentage of word-level corrections to achieve ground truth transcripts.

EFR = “*Estimated Effort Reduction*”.

# Index

- 1 Handwritten Text Recognition (HTR) and Indexing ▷ 1
- 2 The tranScriptorium Project ▷ 3
- 3 Selected Handwriting Datasets ▷ 5
- 4 HTR and Interactive-Predictive HTR ▷ 10
- 5 Interactive HTR: Transcription Demonstration ▷ 12
- 6 HTR and Interactive-Predictive HTR Results ▷ 14
- 7 *Handwritten Text Images Indexing: Search Demonstration* ▷ 16
- 8 Handwritten Text Images Indexing and Search Results ▷ 19
- 9 Conclusion ▷ 22

# Handwritten Text Images Indexing and Search

- There are massive text image collections out there, but their textual content remains practically inaccessible
- If perfect or sufficiently accurate text image transcripts were available, image textual context could be straightforwardly indexed for plaintext textual access.
- But fully automatic transcription results lack the level of accuracy needed for useful text indexing and search purposes
- And manual or even interactive-predictive assisted transcription is entirely prohibitive to deal with massive image collections
- Good news: indexing and search can be directly implemented on the images themselves, *without explicitly resorting to any image transcripts*, as we will see now.

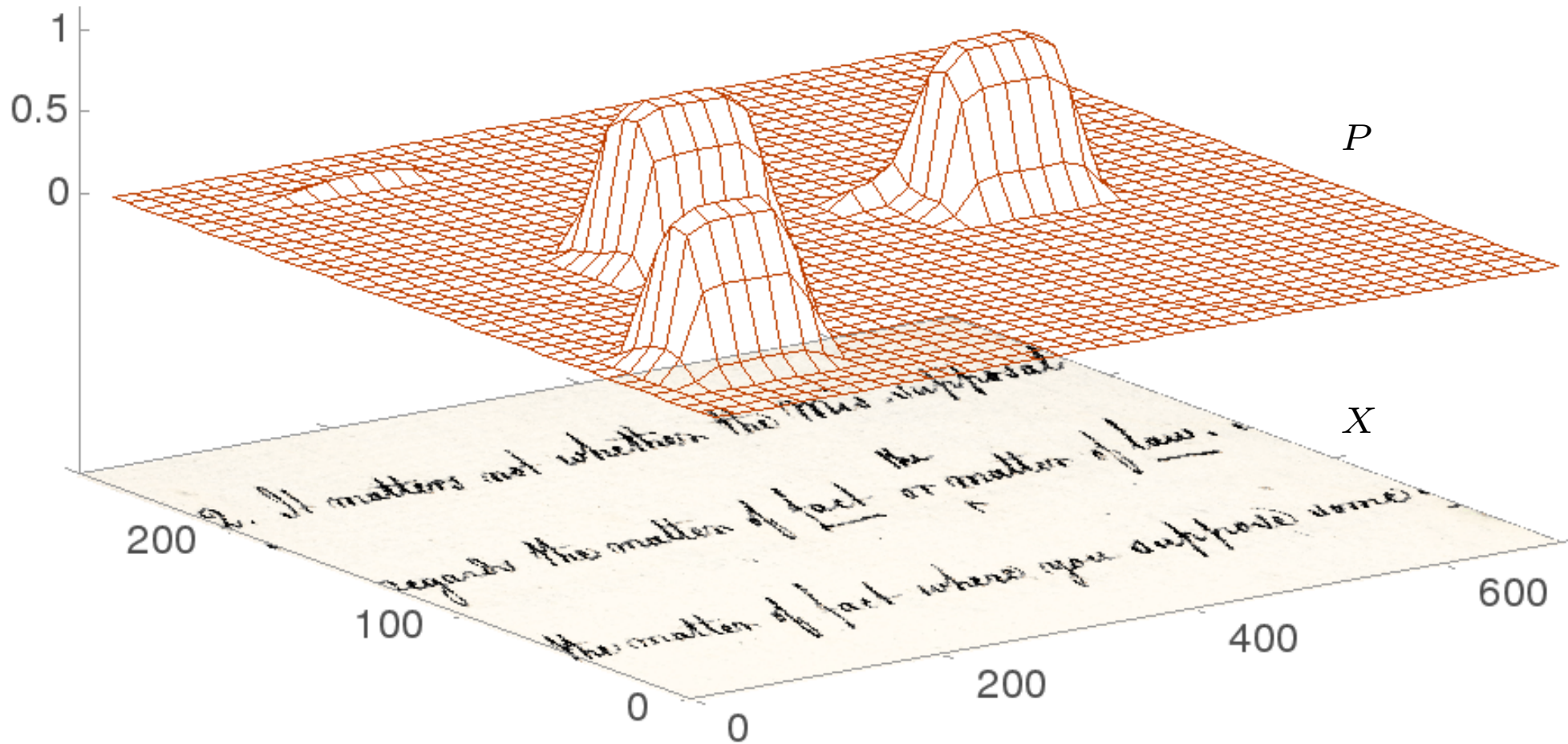
# Handwritten Text Images Indexing and Search: Demonstration

- It is just a “*demo*”! not (yet) intended for real operation. But everything is *real* – no tricks to make demo look better than real
- Line-level indexing according to the *precision-recall trade-off model*:  
Rather than exact searching, search is carried out with a *confidence threshold*, specified by the user as part of the query in order to meet the required *precision-recall trade-off*
- Word confidence scores are based on pixel-level probabilities and computed for *line-shaped regions*. Spotted word positions are marked only approximately
- Two tasks:
  - AUSTEN: Trained on Austen (50p), 20K words open vocabulary. Demo on the whole “Juvenile volume The Third” (128 pages)
  - PLANTAS: Trained on Plantas (224p), 21K words open vocabulary. Demo on Volume I (about 1 000 pages)

# Index

- 1 Handwritten Text Recognition (HTR) and Indexing ▷ 1
- 2 The tranScriptorium Project ▷ 3
- 3 Selected Handwriting Datasets ▷ 5
- 4 HTR and Interactive-Predictive HTR ▷ 10
- 5 Interactive HTR: Transcription Demonstration ▷ 12
- 6 HTR and Interactive-Predictive HTR Results ▷ 14
- 7 Handwritten Text Images Indexing: Search Demonstration ▷ 16
- 8 *Handwritten Text Images Indexing and Search Results* ▷ 19
- 9 Conclusion ▷ 22

# Indexing and Search for Handwritten Text Images: Pixel-level Posteriorgram

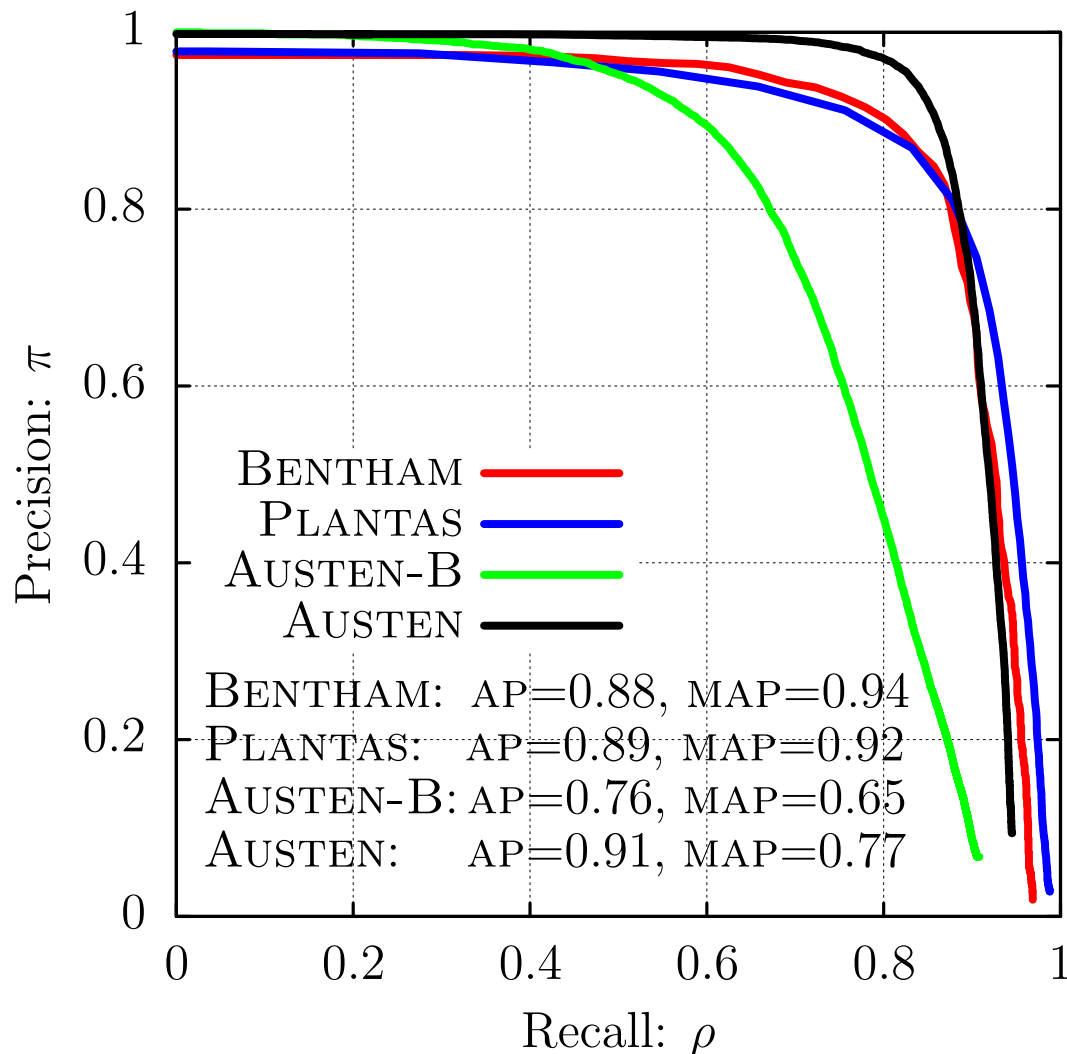


Pixel-level posterior probabilities  $P$  for a text image  $X$  and word  $v = \text{"matter"}$ .

An accurate, contextual ( $n$ -gram based) *word classifier* was used to compute  $P$ . This helped to achieve very low posteriors in a region of  $X$  around  $(i = 100, j = 200)$ , where a very similar word, **"matters"**, is written.

# Results on tranScriptorium Data Sets

*Average Precision (AP)  
Mean Average Precision (MAP)  
and Recall-Precision curves*



## *Datasets training and test details*

- **BENTHAM**: *Multi-hand. Training*: 400 pg. from Bentham, 87 char. HMMs, 2-gram LM trained on Bentham texts; Lexicon 9 341 tokens.  
*Test*: 33 pages; query set: 6 962 keywords
- **AUSTEN-B**: *Single hand. No training*; using Bentham char. HMMs, lexicon and LM.  
*Test*: 78 pages; query set: 9 000 keywords
- **AUSTEN**: *Single hand. Training*: 50 Austen pages, 81 char. HMMs, 2-gram LM trained on Austen texts; Lexicon 20K tokens.  
*Test*: 78 pages; query set: 9 000 keywords

# Index

- 1 Handwritten Text Recognition (HTR) and Indexing ▷ 1
- 2 The tranScriptorium Project ▷ 3
- 3 Selected Handwriting Datasets ▷ 5
- 4 HTR and Interactive-Predictive HTR ▷ 10
- 5 Interactive HTR: Transcription Demonstration ▷ 12
- 6 HTR and Interactive-Predictive HTR Results ▷ 14
- 7 Handwritten Text Images Indexing: Search Demonstration ▷ 16
- 8 Handwritten Text Images Indexing and Search Results ▷ 19
- 9 *Conclusion* ▷ 22

# Conclusions and Future Work

## *Conclusions*

- Automatic or assisted handwritten text transcription and fully automatic indexing is now becoming perfectly feasible
- Models trained for a given collection can provide quite useful performance on images from other similar collections, without need of (re-training)
- Several demonstrators have been implemented and made publicly available for first-hand experience in real use; see: <http://transcriptorium.eu/demonstrations>

## *On-going and future work*

- Research to overcome the line-detection bottleneck
- Indexing and search experiments with massive handwriting document collections (thousands to millions page images)



## Part 2

# Introduction to the Transcription and Recognition Platform - TRANSKRIBUS



*TRANSKRIBUS enables collaboration among humanities scholars, computer scientists, archives and volunteers with the ultimate goal to revolutionise recognition, transcription and access to historical handwritten documents.*



ARCHIVES & LIBRARIES

HUMANITIES  
SCHOLARS

COMPUTER  
SCIENTISTS

PUBLIC USERS



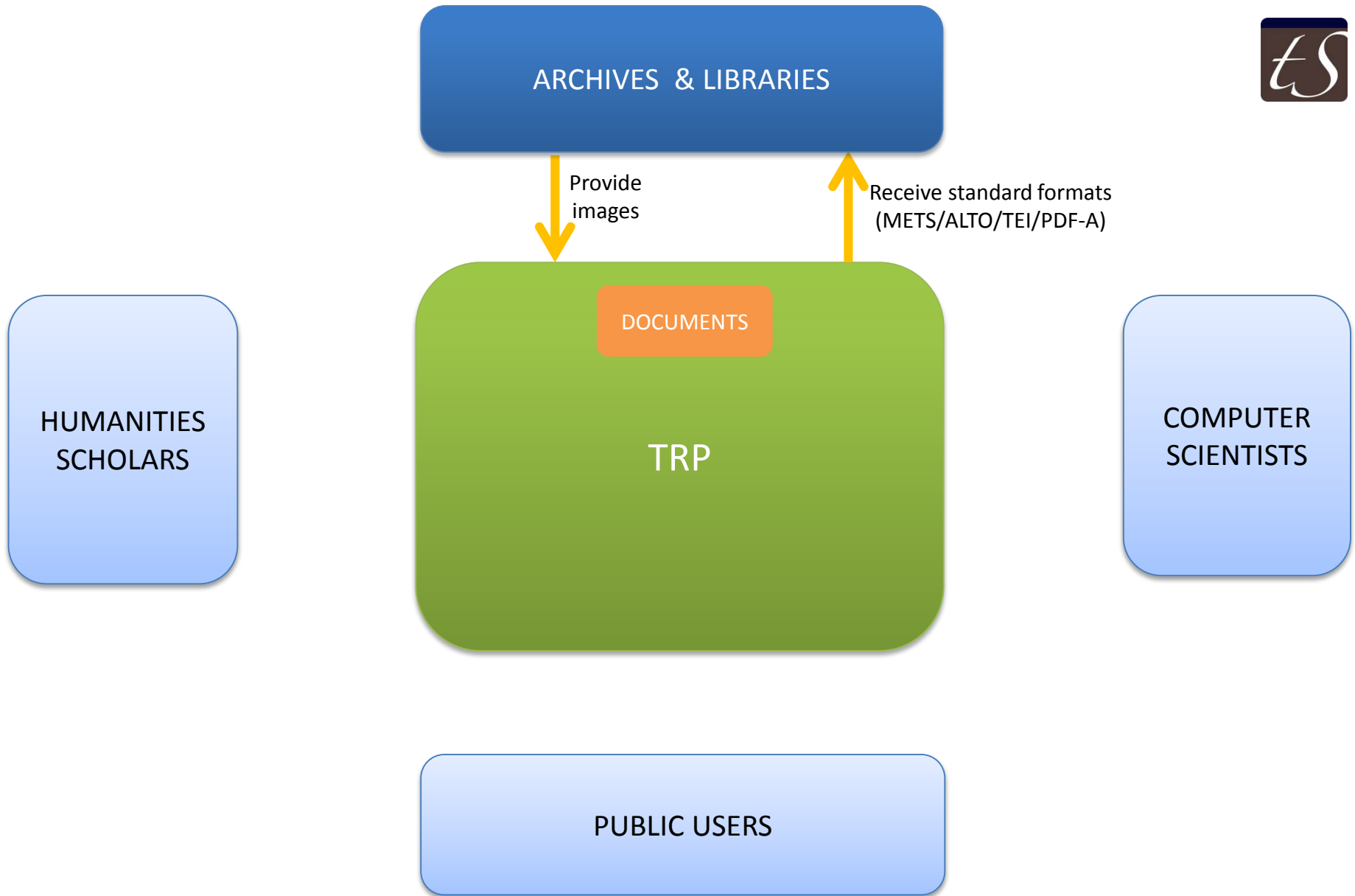
ARCHIVES & LIBRARIES

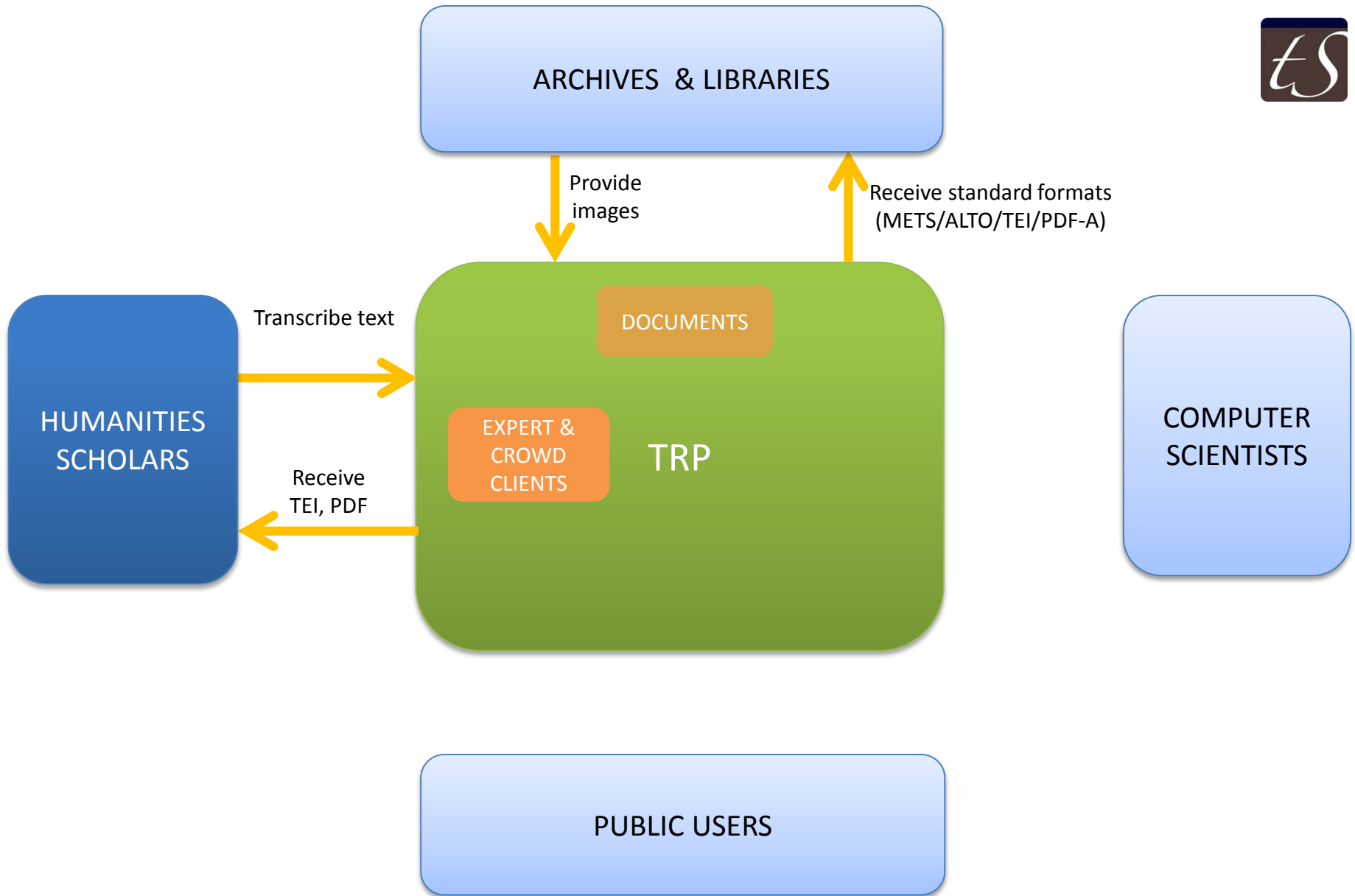
HUMANITIES  
SCHOLARS

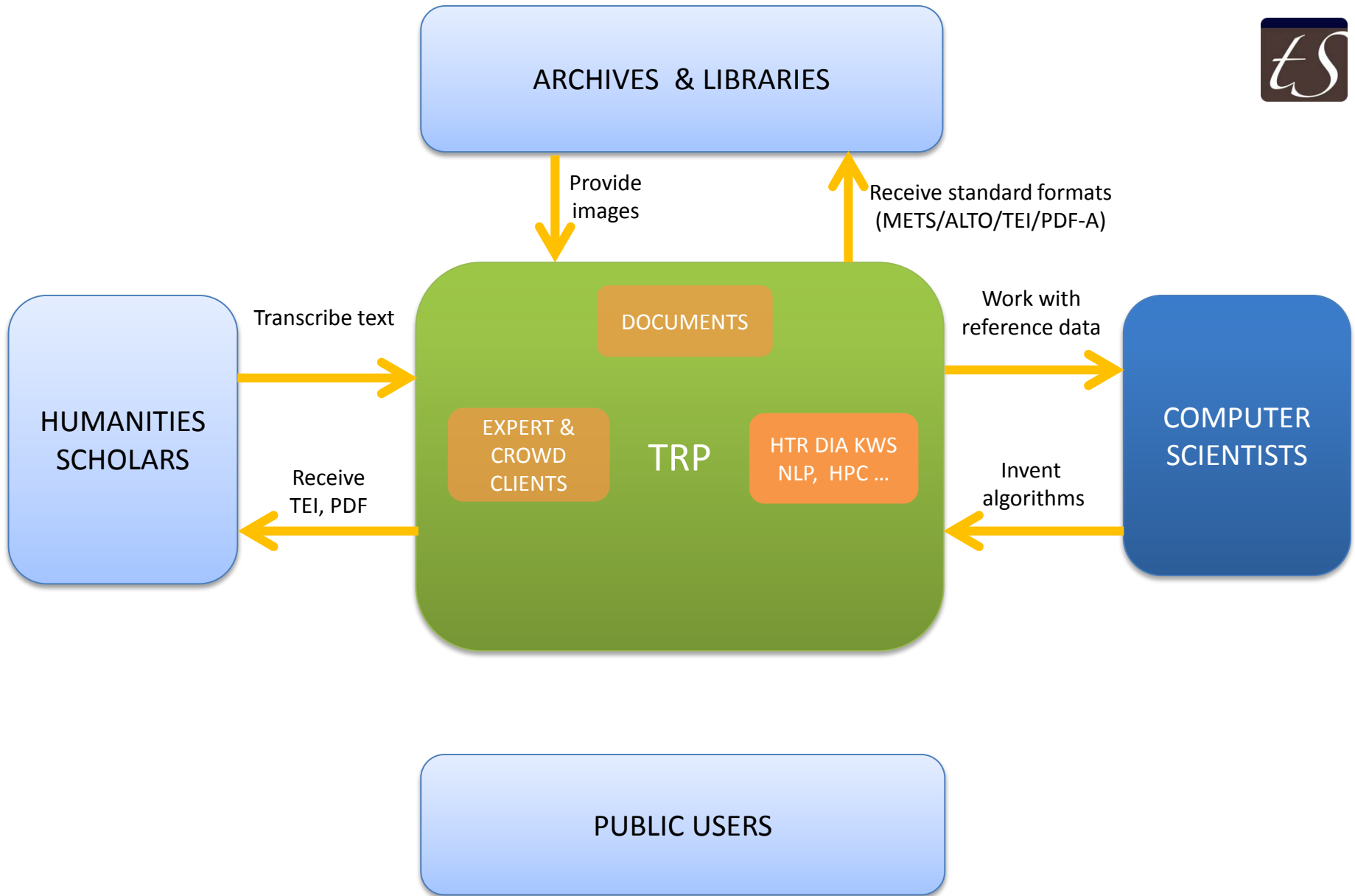
TRANSCRIPTION &  
RECOGNITION  
PLATFORM

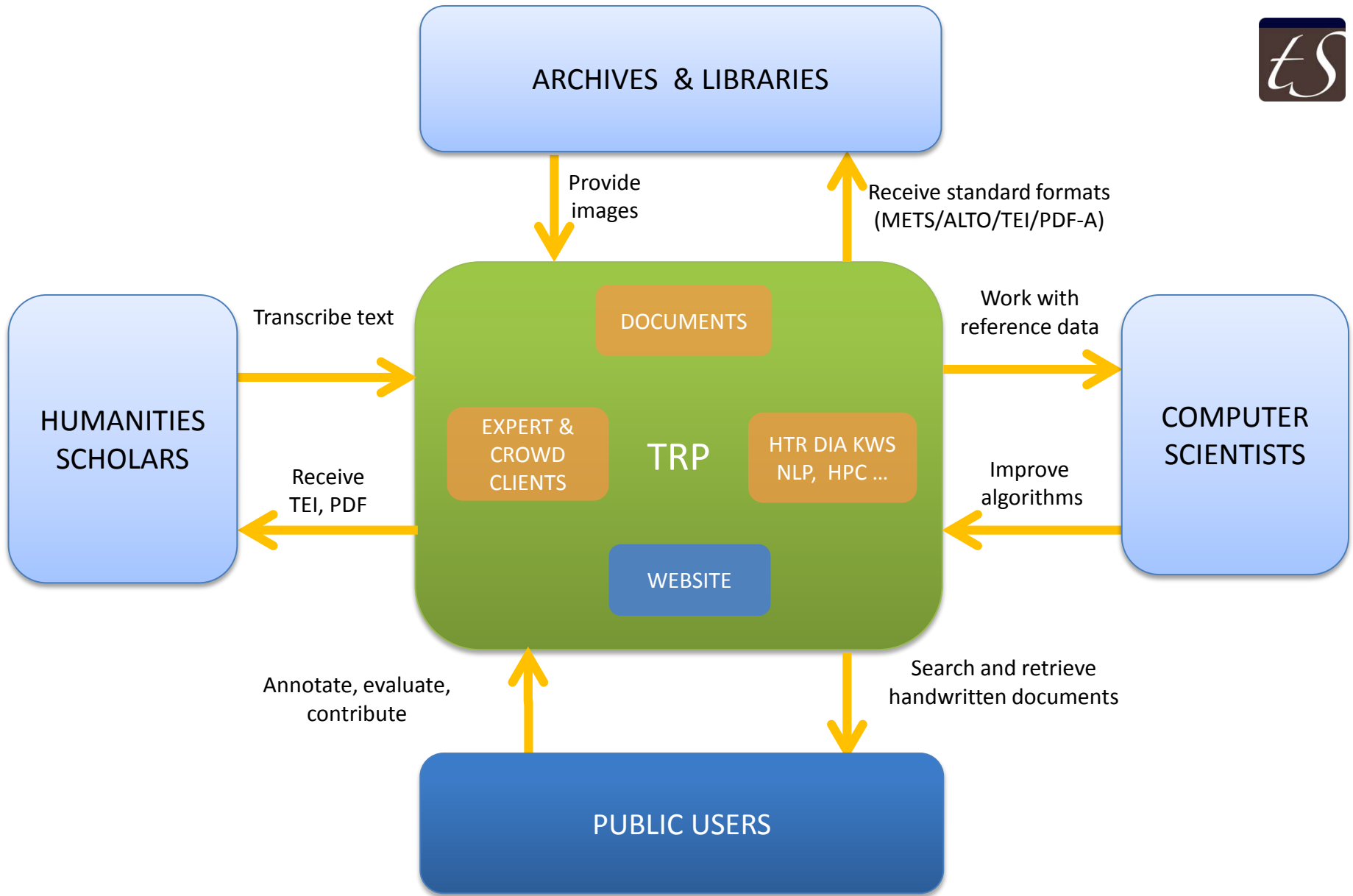
COMPUTER  
SCIENTISTS

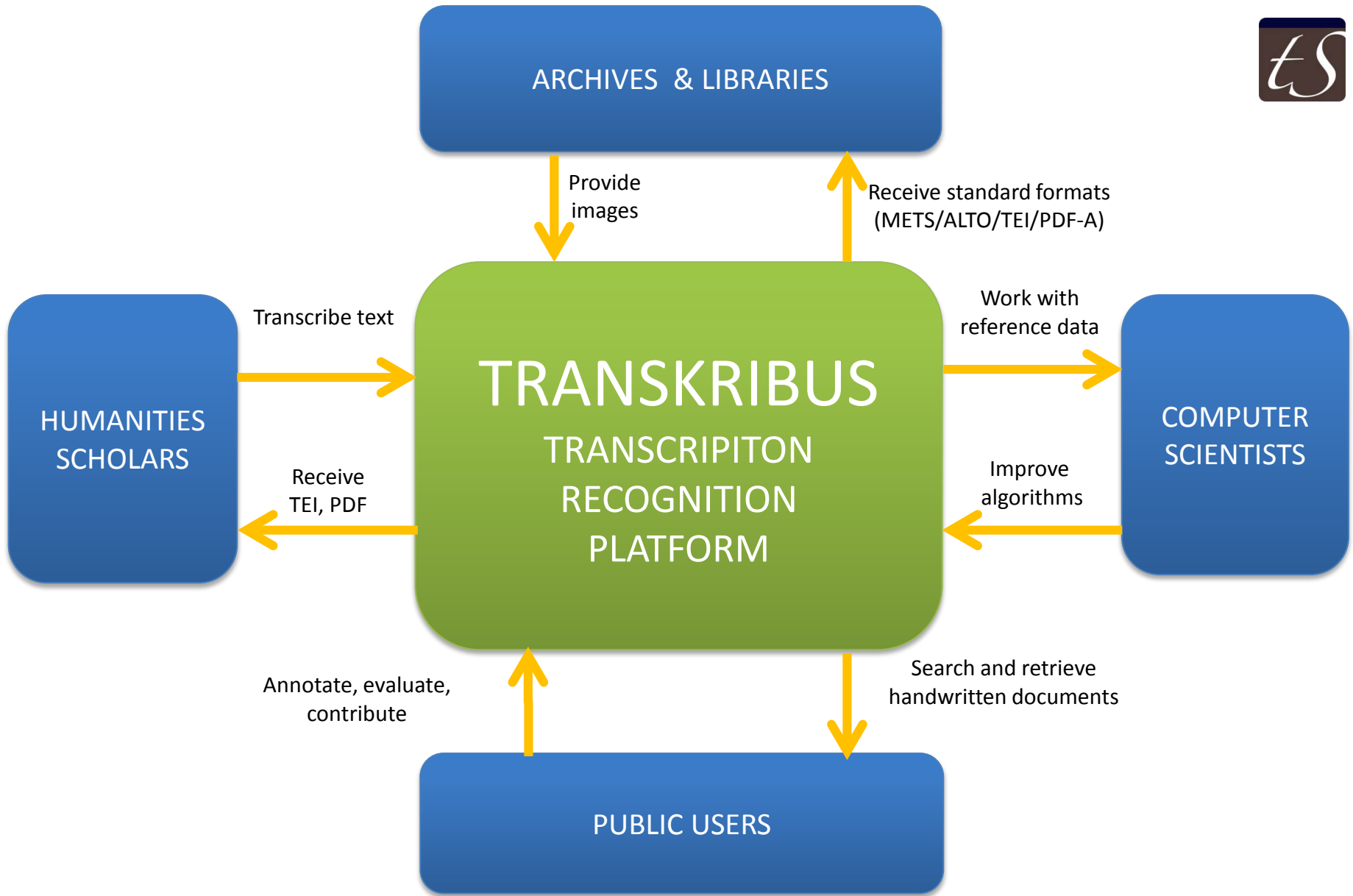
PUBLIC USERS



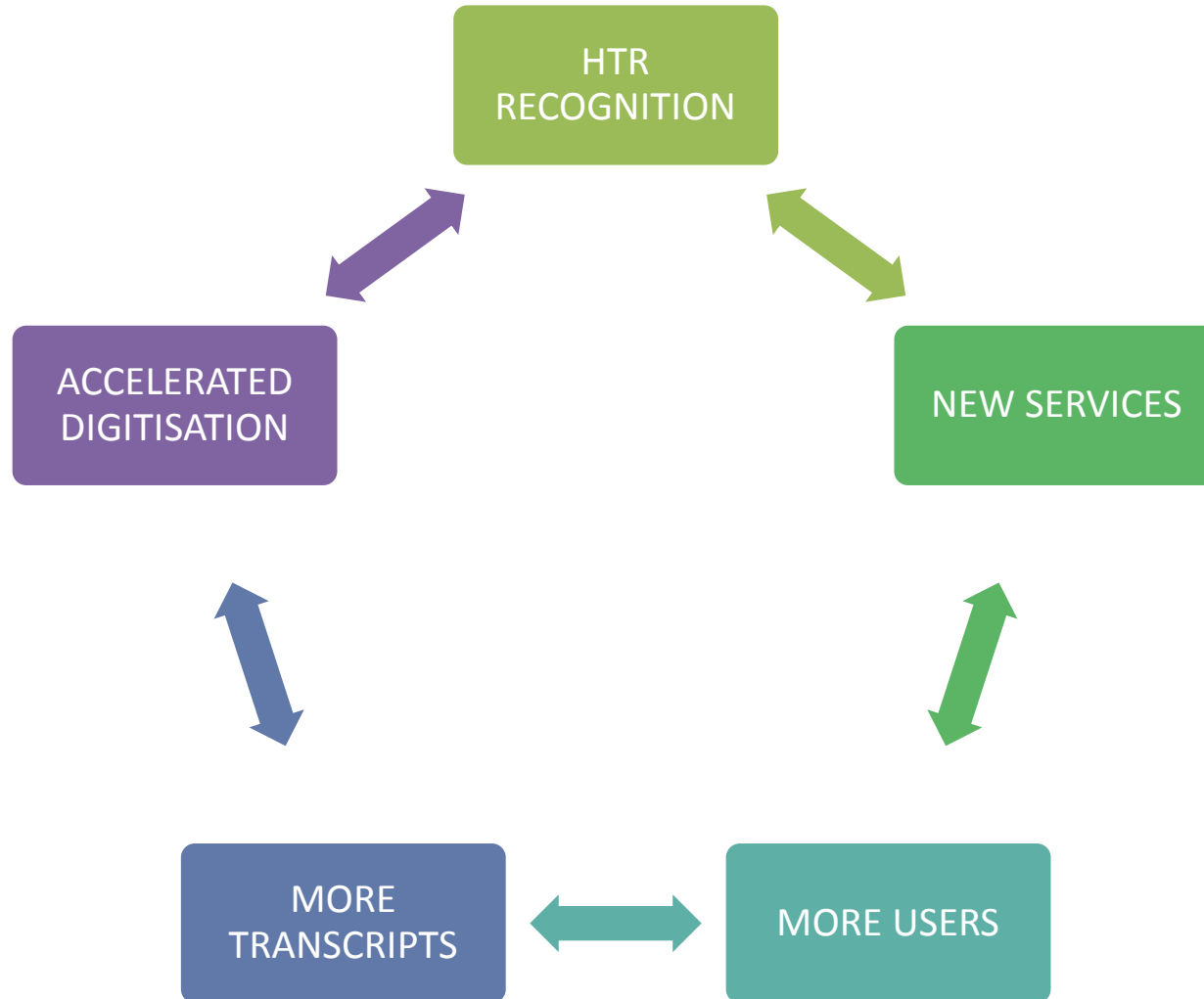




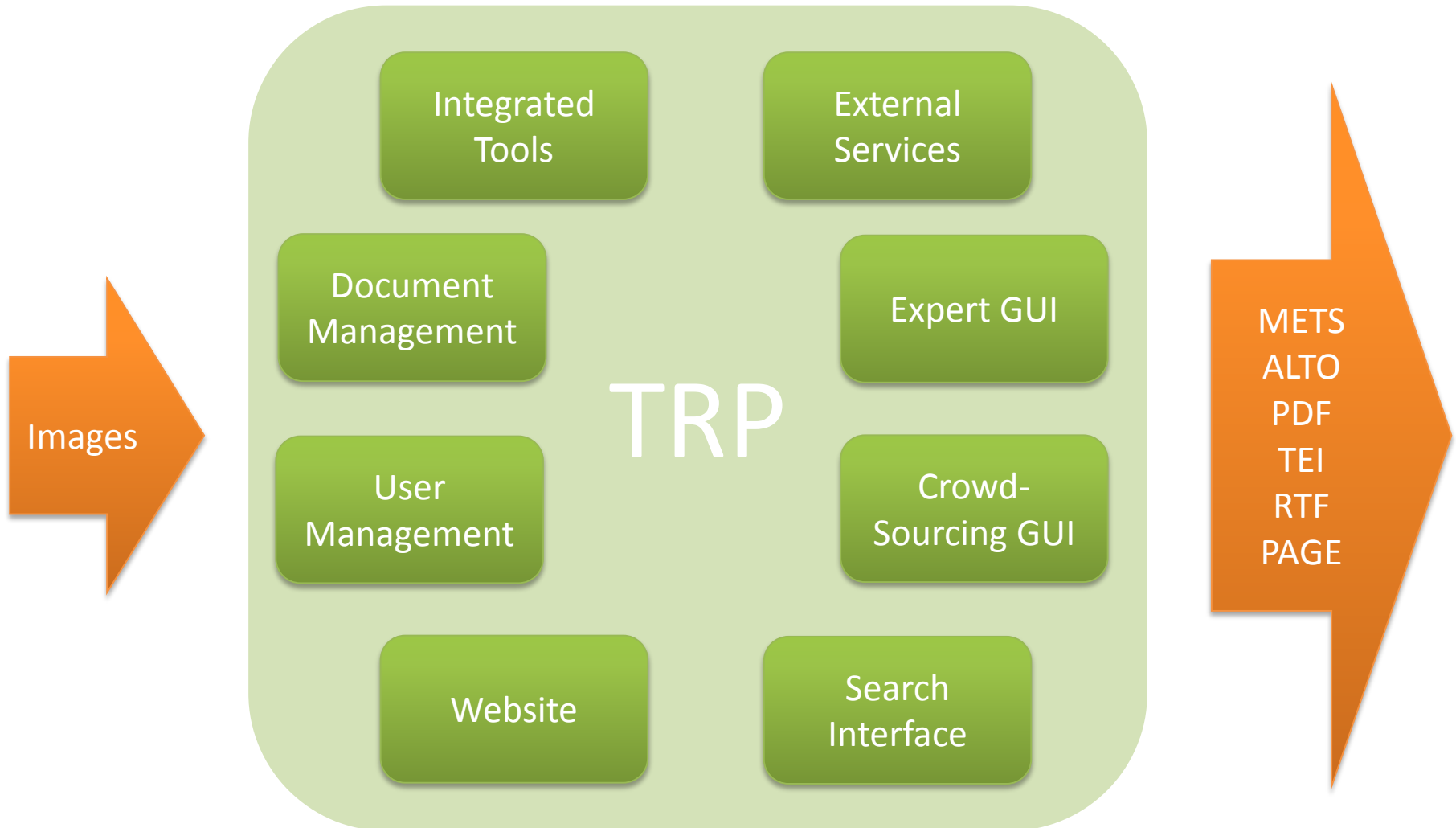




# Cycle of growth



# Architecture



# Archives/collection holders are enabled to



- Manage the transcription process of large and heterogeneous document collections in a standardized and effective way
- Expose their digitised documents to their own employees, humanities scholars, volunteers and the crowd
- Involve users according to their background
- Support humanities scholars in their research
- Outsource transcription (or the training of an HTR model) to service providers (e.g. off-shore)
- Make large amounts of handwritten documents full-text searchable (if a trained HTR model is available)
- Import transcribed documents in machine readable formats (METS/ALTO)
- Provide standardized requirements to researchers and technology providers dealing with “issues of interest” (e.g. manage tendering)

# Humanities scholars are enabled to (1)



- Use TRANSKRIBUS for free (registration)
- Upload as many single documents as they want
- Transcribe their documents semi-automated in a way that image and text are linked to each other (improved transcription quality!)
- Enrich their documents with Named Entities (person and geo names, dates) and personal tags
- Normalize their transcription in a transparent way (abbreviations, character sets, editorial declaration)
- Export their document in various formats
  - TEI: transcribed text with machine and human readable text (“scientific style”)
  - PDF-Text: transcribed text with some formatting (“book style”)
  - PDF-Image: transcribed text in the background, image in the foreground (“Scanned PDF style”)
  - RTF: transcribed text for human readability (“working style”)
  - METS/ALTO: full machine readable package (“digital library style”)
  - METS/PAGE: full machine readable package (“computer science style”)

# Humanities scholars are enabled to (2)



- Manage their documents in a private collection (no one else has access!)
- Invite other users to collaborate (e.g. students, colleagues) and to work on the same document
- See different versions of their document and compare them to each other
- Make their documents available to every registered user (=crowd-sourcing scenario)
- Expose their documents also in a simplified web-based transcription GUI (TSX)
- Transcribe documents with the support of HTR if a model is available (needs offline training in beforehand)
- Exploit language resource database in the background (e.g. words, variants, frequencies, etc.)
- Search within a handwritten text collection
- Benefit from other **TRANSKRIBUS** users in an indirect way. The following resources will be available to every user of the platform:
  - Trained HTR models
  - Normalized editorial declarations
  - Normalized Named Entities
  - Language resources

# Computer scientists are enabled to



- Benefit from ongoing transcription work in the platform
- Access large amounts of highly valuable data (transcribed text with segmentation) in standardized format (PAGE)
- Download documents (if they are public domain, if the document owner has agreed, or if just small random sets are used)
- Investigate new methods, algorithms and tools on the basis of “real world data”
- Have a tool available with which humanities scholars and archives can express their requirements in a highly standardized way
- Use several web-services for integrating their tools into the platform
- Expose their research results/tools/methods within the platform and increase their reputation among the several communities (humanities, public, archives)
- Compare their results with other research groups (competitions) on the basis of a common document set
- Manage the generation of ground truth (reference data) in an effective way
- Collect and gain feedback from different user groups

# Volunteers and crowd-users are enabled to



- Work with the same tools and in the same way as “professional researchers”
- Make a valuable contribution by enriching archival collections (crowd-sourcing)
- Take part in “citizen science” projects (e.g. produce “Ground Truth”)
- Benefit from easy-to-use web-based interfaces for low-barrier participation (TSX)
- Upload their own family documents
- Benefit from state-of-the-art technology in Computer Vision, Document Layout Analysis, HTR, OCR, etc.
- Expose their documents to experts, service providers, etc.
- Search in the full-text of handwritten documents

# Business model



- Free usage
  - Single documents (who ever wants to work with the platform)
  - E.g. humanities scholars, volunteers, genealogists, archivists,...
- Service Level Agreement
  - Document collections
  - E.g. transcription projects (grants), archives, libraries (OCR collections)
  - Yearly fee depends on the amount of documents and services
- Public-Public-Partnership
  - Subsidiary model, not only for TRANSKRIBUS itself, but also for participants (e.g. archives, collection holders)
  - E.g. DARIAH, direct support via e-infrastructure funds, research agencies (may have interest on standardized formats and workflows, etc.)
- Vision
  - Hundreds of humanities scholars, thousands of volunteers, dozens of archives,...

# Next steps



- **Collect feedback**
  - Platform and GUIs are in a “usable” form, but many important features are still not included
  - First real user tests start with this workshop
- **Spread the idea**
  - Several workshops are planned in UK, Austria, Germany, Netherlands, etc.
  - Attract as many humanities scholars and volunteers as possible to try out the platform and to integrate it into their research
  - Address archives and collection holders and convince them to enter service level (pre-)agreement with University of Innsbruck / TRANSKRIBUS
- **How to contribute?**
  - Try out the tools and give us feedback!
  - Upload your documents and transcribe 100 pages for training the HTR
  - Invite us for workshops, presentations, webinars...
  - Spread the idea!



Thank you for your attention!