

Commercial Overview

tranSkriptorium



Presentation Outline

Introduction	2
Motivation	3
Solution	4
History	5
Team	6
Technology - General Overview	7
Company Assets	19
Conclusions	20

Introduction

- Immense collections of historical manuscripts are stored in thousands of kilometres of shelves in archives and libraries
- It is estimated that the total amount of handwritten text is still greater than the amount of mechanized text



- Digital preservation of these works shouldn't be the final goal. All efforts should go towards making the valuable information contained in them available for consumption.
- Digitalization is a necessary step, **but insufficient**

Motivation

- Is the current tendency to digitalize collections truly delivering easy access to the information?
- How is one to search through the thousands of images of a collection for the content they need?
- Can any user, without the correct context and expertise, discern the contents?
- What would be the cost in expert hours and the cost of opportunity?
- How much of this invaluable information are we ready to lose forever?
- Would you be OK with a massive binary dump of all the data in your company and no way to search or actually understand what the contents are?



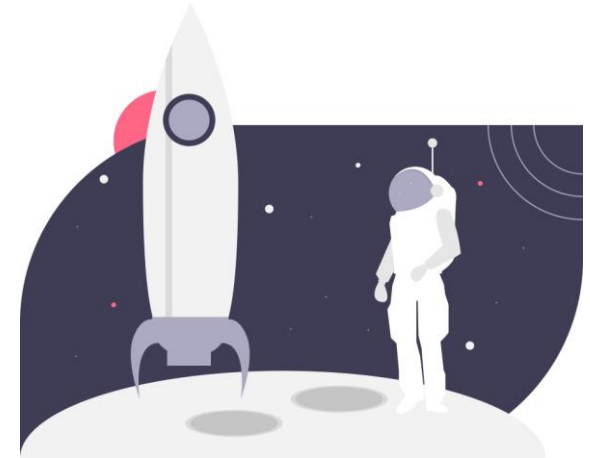
Solution



- Transcribing all these texts would facilitate access to their contents for an extraordinary number of users and researchers
- Unfortunately, manual transcription is **prohibitive** and unassisted automatic transcription **lacks the desired precision**
- Via Computer Assisted Transcription we can make precise transcriptions at affordable prices
- Even better, we can automatically Index and allow probabilistic searches without the need of transcribing
- Our probabilistic indexes allow you to perform big data analysis over the indexed documents: classification, automatic summaries,...

History

- This technology is now available as the result of the effort of a remarkable high-level research team
- It has matured over decades of research
- Product of cutting-edge national and international research projects
- Sustained by hundreds of peer reviewed articles
- The enormous international success of the developed projects guarantees the acceptance and value of this technology for an untapped market



Team



Luis Antonio
Morró González
CEO



Enrique Vidal
Researcher



Joan Andreu Sánchez
Researcher



Verónica Romero
Researcher

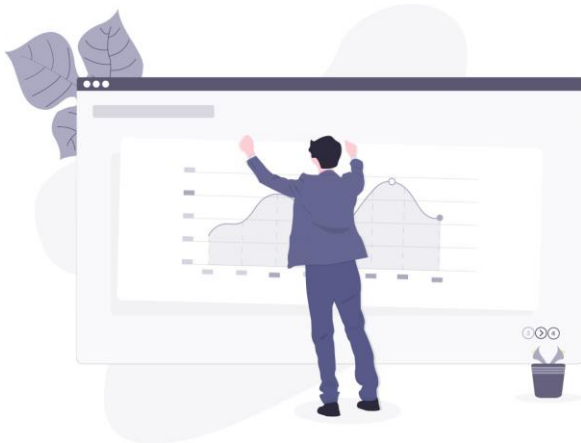


Vicente Bosch
Researcher



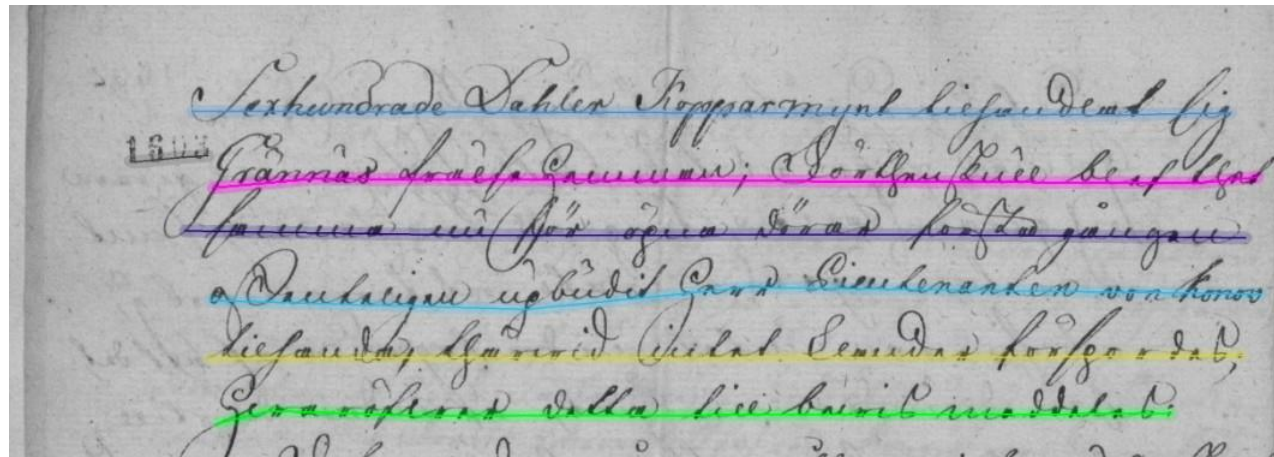
Alejandro Héctor Toselli
Researcher

Technology - General Overview



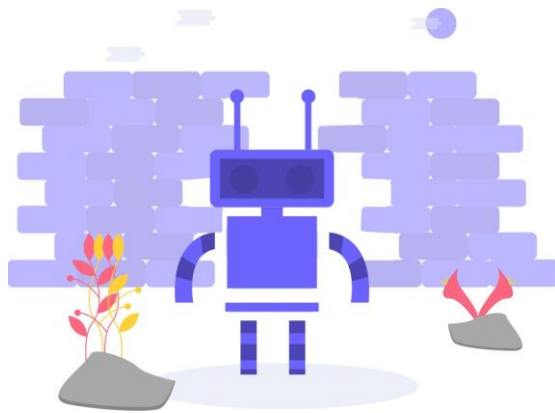
- We provide end to end solutions for transcription and indexing of digitalized documents:
 - Document Layout Analysis
 - Automatic Transcription
 - Computer Assisted Transcription
 - Entity Recognition and Linking
 - Probabilistic Indexing and Querying via our Search Engine and Web GUI
 - Big Data Analysis
- Adaptable to different types of media
- We tackle the tasks and issues no standard OCR software or company does

Document Layout Analysis (DLA)



- Developed our own DLA OSS **P2Pala** applicable to any corpus
- Based on state of the art Deep Learning U-net architecture
- Tackles both line detection and region classification
- Pre-trained model based on hundreds of thousands of text images
- Demonstrator: <http://prhlt-carabela.prhlt.upv.es/tld/>

Automatic Transcription



- Developed our own HTR OSS **PyLaia**
- Device agnostic, PyTorch based, deep learning toolkit
- Language independent. Tested in many languages: English, Spanish, Latin, Bengali, Hebrew, Arabic, Swedish, German, Italian, ...
- Relies on convolutional bi-dimensional and uni-dimensional recurrent layers
- Achieves better or equivalent results to other state of the art more expensive architectures
- Adhoc Language Model training and application that increase accuracy

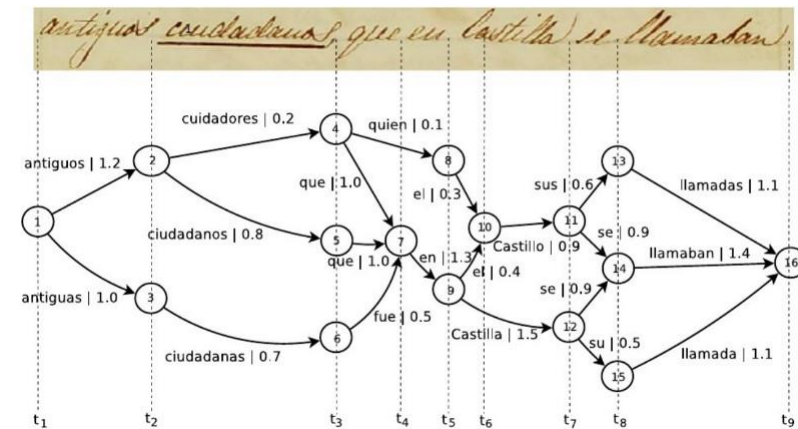
Computer Assisted Transcription

- Proprietary interactive transcription review and correction process *CATTI*

- CATTI*, measurably improves expert productivity

- In house developed web GUI and engine, actively used in many projects

- Demonstrator: <http://transkriptorium.eu/demots/htr/index.php>

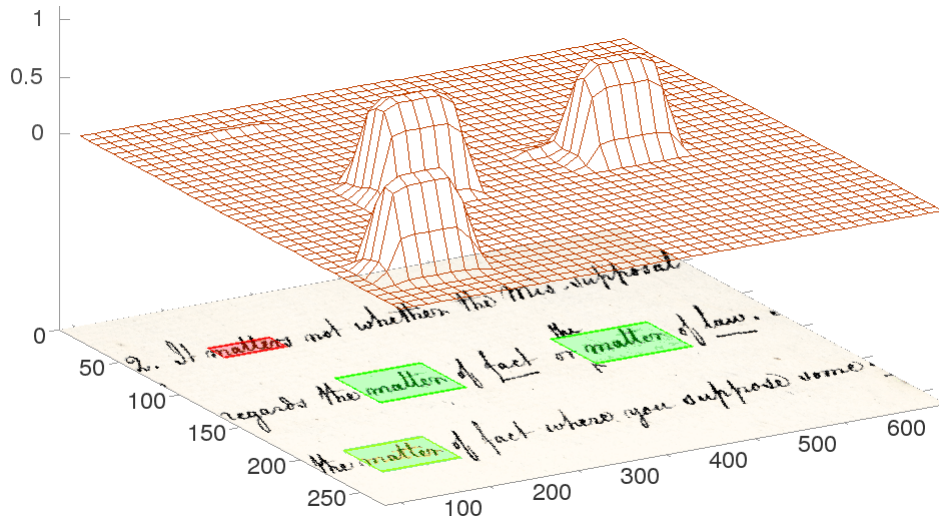


Text Image Probabilistic Indexing and Search

- Google earth meets handwritten text
- One of a kind technology
- Does not require transcription of text in the images
- Works much better than searching in automatically transcribed text
- Hardly impacted by layout analysis issues
- Proprietary non released software: index generation, index engine, search GUI



Text Image Probabilistic Indexing and Search



- A relevance probability map is computed over the whole image
- The probability and location of each detected pseudo-word is stored
- This allows to probabilistically index a word in an efficient manner

- Via a threshold the user has control over the compromise between search precision and recall (or exhaustiveness)
- This technology has been tested in very different and complex document collections

Large Scale Probabilistic Indexation is a Reality

Our team has been developing this technology during this last decade. Recently it has applied it, with great success, to five large handwritten collections making their textual contents completely available:

- *Chancery* (AN & BN, France): 83 000 pages, very abridged French & Latin, 14-15th c.
<http://prhlt-kws.prhlt.upv.es/himanis/>
- *TSO* (*Teatro del Siglo de Oro*, BN de España): 41 000 pages, Spanish, 16-17th c.
<http://prhlt-carabela.prhlt.upv.es/tso/>
- *Bentham Papers* (UCL & BL): 95 000 pages, English scrawl writing, 18-19th c.
<http://prhlt-kws.prhlt.upv.es/bentham/>
- *Carabela* (AGI + AHPC): 125 000 pages, Spanish, abstruse scripts,, 16-18th c.
<http://carabela.prhlt.upv.es/es/demonstrators>
- *FCR* (*Finnish Court Records*, NA Finland): "more than 1 000 000 pages, Swedish, 18-19th c. <http://prhlt-kws.prhlt.upv.es/fcr/>

Over 1 500 000 handwritten document images processed!

Beyond Basic Keyword Search

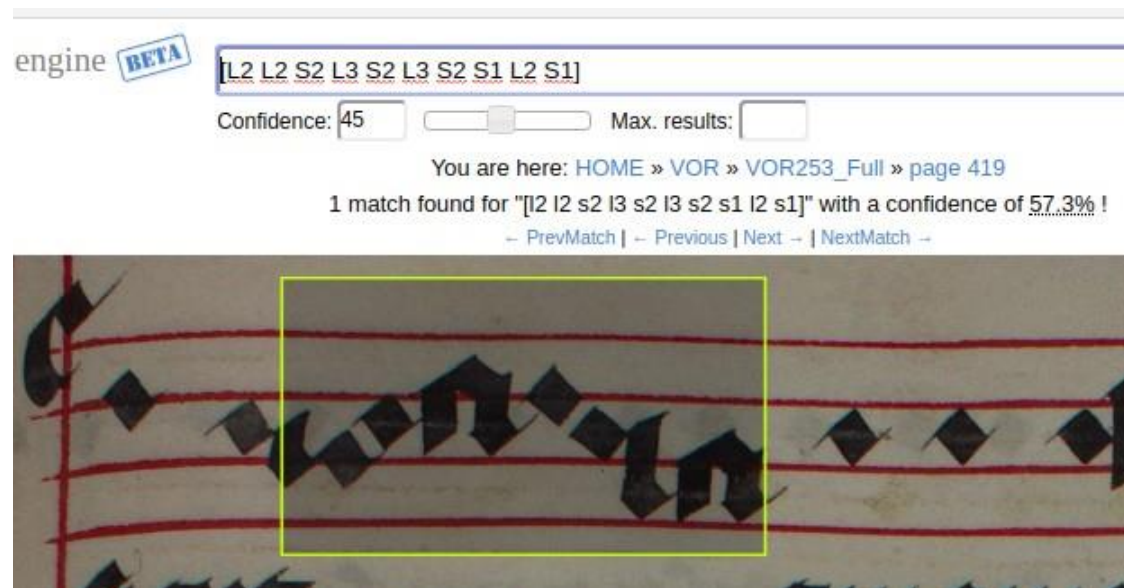
Our search engine and interface allow:

- Searches with word spelling flexibility: wild cards, approximated spelling and hyphenated words
- Boolean combination and sequence queries
- Queries taking into consideration page geometry (not allowed with other commercial software):
 - Indicating a maximum allowed distance between the searched terms
 - Allowing DB like **queries** by header and value in handwritten tables
- Semantic searching through complex queries

Beyond Text Image Probabilistic Indexing and Search

- This technology can be applied to search and retrieve any content from different media
- It can be, for example, used to spot melodic patterns in music sheet documents

<http://prhlt-carabela.prhlt.upv.es/music/>



Big data analysis: from classification to automatic summaries

- Text analytics are required to uncover insights, trends and patterns in documents
- Text features computed over digital text are required to use most big data analysis tools on documents
- Performing these types of analysis on an automatic transcription is error prone
- Fortunately these features can be accurately estimated from probabilistically indexed images:
 - Total number of running words
 - Frequency of use of a given word
 - Zipf's curves
 - Size of vocabulary



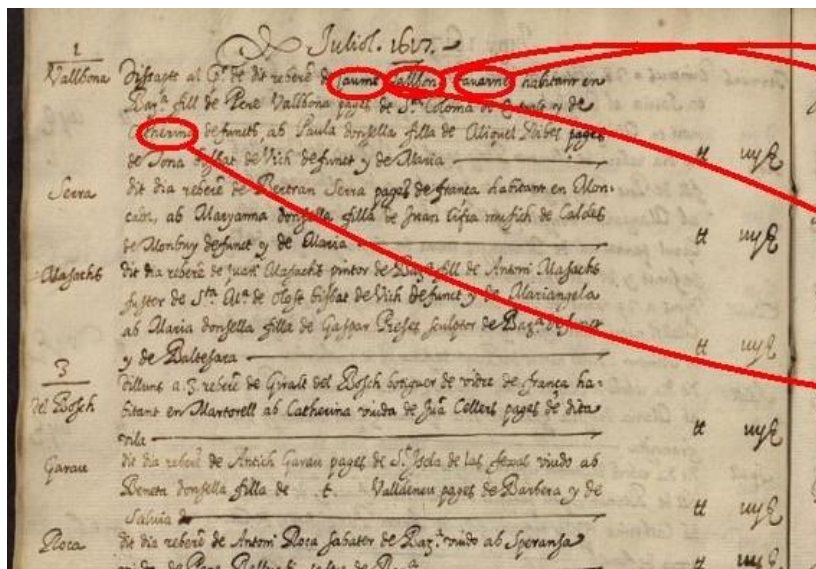
Big data analysis: from classification to automatically generated summaries



- These features enable, for example, classification of documents
- Classification by means of user provided (maybe complex) queries or via successful Machine Learning plain-text classifiers
- Applications:
 - Carabela project: classification of documents into classes of public access risk
 - TSO project: classification used to identify possible authors of currently anonymous manuscripts
 - HisClima and Passau project: retrieve data from tables for big data analysis
 - Collaboration with Universitat de Valencia: automatically process Nomenclátor

Named Entity Recognition and Information Extraction

- Effectively processing records requires the detection of semantic information contained in them
- This allows us to extract the information to a database for easy consumption
- To perform this process manually is prohibitive
- Fortunately, this information extraction process can be also carried out from probabilistically indexed images.



	A	B	C	D	
1	Groom Name	Groom Surname	Groom Occupation	Bride Name	Bride
2	Jaume	Vallbona	Javerner	Catherine	Defunc
3					
4					
5					
6					
7					

Company Assets

- A tested on-boarding method to process new corpora and deliver results
- State of the art open source and proprietary software developed and reviewed in 15 national and international research projects
- A team with over 500 peer reviewed publications
- Backed by the UPV as one of its spin-off flagships of 2020
- Part of the READ-Coop ecosystem

Conclusions

- tS presents a value-added technology, that tackles an issue not resolved by current preservation technologies
- This technology gives meaning to the effort of preserving the physical documents by solving the accessibility problem
- The market is totally neglected and the value of applying our solution is incalculable as it is not currently being exploited by other competitors
- The company's human and technological assets make up a strong team united by great motivation and determination



Team - extra



Interim CEO - Luis Antonio Morró González. Formal education as an Industrial Technical Engineer, specialized in Mechanics and Construction by the UPV. A 32 year long professional career in different areas related to business management like: CEO, General Director and Consultant in various multinational companies and in important national groups, taking on in each of them responsibility over different areas: Industrial, Services and Strategical Consultancy, Operations and Finances. Complements his capacities with training in different areas related to business management.

[back](#)

Team - extra



Enrique Vidal, Emeritus professor in the Polytechnic University of Valencia (UPV) has been co-director during decades of the Pattern Recognition and Human Language Technologies (PRHLT) research centre. Co-author of more than 250 scientific publications in the areas of Pattern Matching, Multimodal Interaction and application in the automatic processing of language, spoken and written. In these areas and applications he has lead several large projects, including various international projects and a Spanish of the Consolider 2010 Ingenio programme. Dr. Vidal is a fellow of the

International Association for Pattern Recognition (IAPR). His H-index is 49, according to Google Scholar.

[back](#)

Team - extra



Joan Andreu Sánchez, Associate Professor in the UPV and member of the PRHLT. The research areas he is interested in include Pattern Matching, Machine Learning and their application to Handwritten Text Recognition. Dr. Sánchez has participated in different European and national projects related with this theme and has led the tranScriptorium European project. He is co-author of more than 100 articles published in different magazines and proceedings of international conferences.

[back](#)

Team - extra



Verónica Romero, Doctor in Computer Science by the UPV since 2010. In 2005 she joined the PRHLT research centre. Her area of interest include pattern matching, multi-modal interaction and applications of handwritten text recognition. She obtained the award of best thesis of the year of the UPV for her work on assisted transcription of handwritten documents. In these fields she has published more than 60 articles in magazines, conferences and books of high impact. Currently, she works as a researcher in the PRHLT centre

working in the different projects related with handwritten text recognition in historical documents. Additionally, she is a lecturer in the Statistics, Operative Investigation and Quality department of the UPV.

[back](#)

Team - extra



Vicente Bosch Campos, As of January 2020 Doctor in Computer Science by the UPV. Post graduating in the UPV in 2005 he joined an international renowned consulting firm where he worked for 7 years. During his time in this firm he worked as System Architect, Development Team Lead, Financial Officer and Lyason with the off-shore team. He was involved in projects ranging from State services, passing through telcos and pharmaceuticals to Supermarket chains

and resources companies. In 2012 he decided to take a sabbatical period and rejoin the university to take a Masters in AI. Upon finishing his Masters degree he joined the PRHLT research centre where he performed his PHD while participating in various research projects. His major fields of interests are pattern recognition and document layout analysis in which he has performed 11 peer reviewed publications. He is currently working as Senior Technical Officer in the PRHLT. [back](#)

Team - extra



Alejandro Héctor Toselli, Electrical Engineer but the National University of Tucumán in Argentina (1997) and Doctor in Computer Science by the UPV since 2004. He has performed various post-doctoral stays of relevance like the one in the “Institut de Recherche en Informatique et Systèmes Alatoires” (IRISA, Rennes France, 2008), with the “Recognition and interpretation of Images and Documents” (IMADOC) research group. He has worked as a full-time researcher in the PRHLT centre participating actively in the different European projects (tranScriptorium, READ, etc.). He is currently working as an Associate Research Scientist in the North-eastern University.

[back](#)