

# Historical Documents in Focus:

Visual and Computational Analysis from Papyri to Inscriptions

Giuseppe De Gregorio: [giuseppe.deguglio@unibas.ch](mailto:giuseppe.deguglio@unibas.ch)



# Program of the Tutorial

- **Introduction to Ancient Document Analysis**

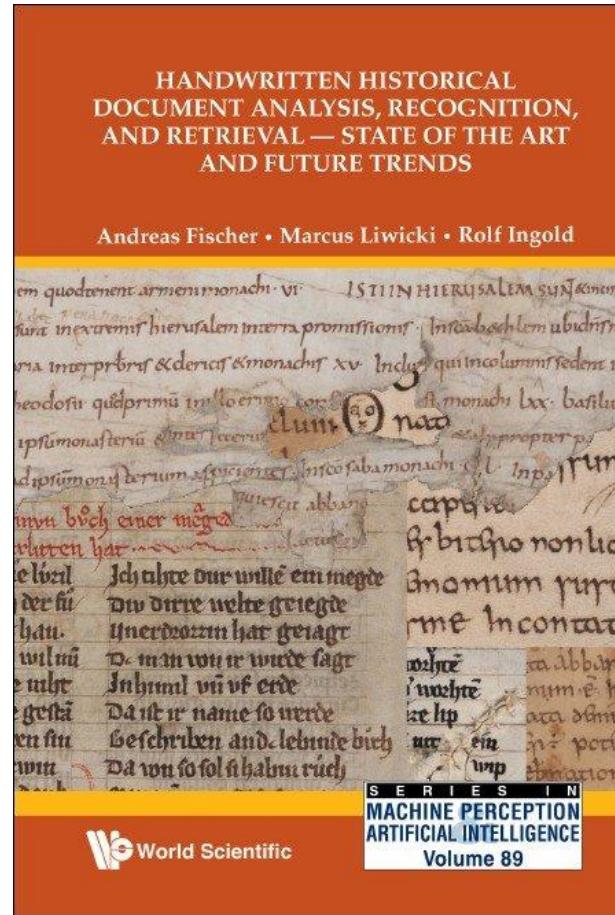
A brief introduction to what historical document analysis is and what the challenges and opportunities it offers.

- **Break**

- **Example of Application**

We will program in Python a model for the Detection and Recognition of Ancient Greek characters on papyrus from Egypt.

# An Interesting Read



**Handwritten Historical Document Analysis, Recognition, and Retrieval — State of the Art and Future Trends**  
Ed. Andreas Fischer, Marcus Liwicki and Rolf Ingold

<https://doi.org/10.1142/11353>

# What are Historical Documents?

# Early Modern Manuscript

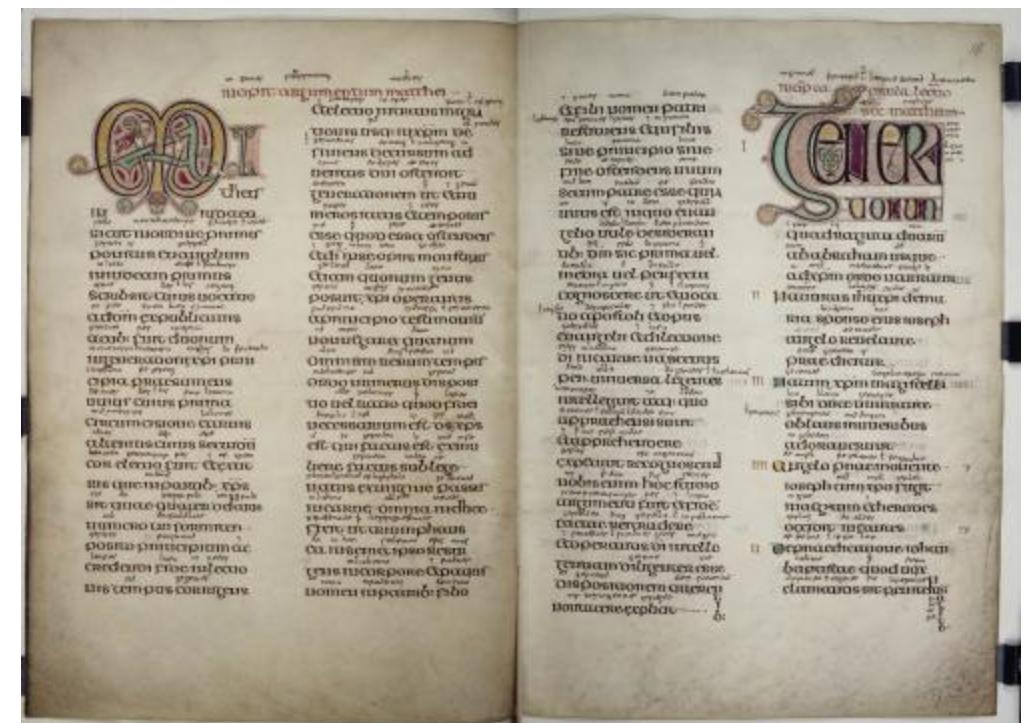
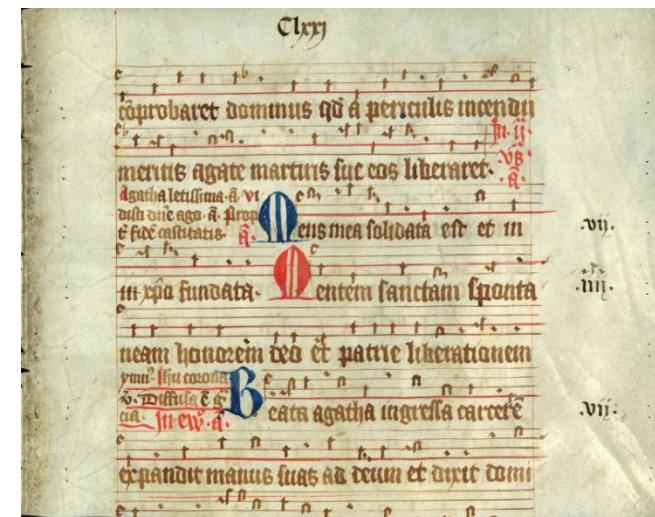
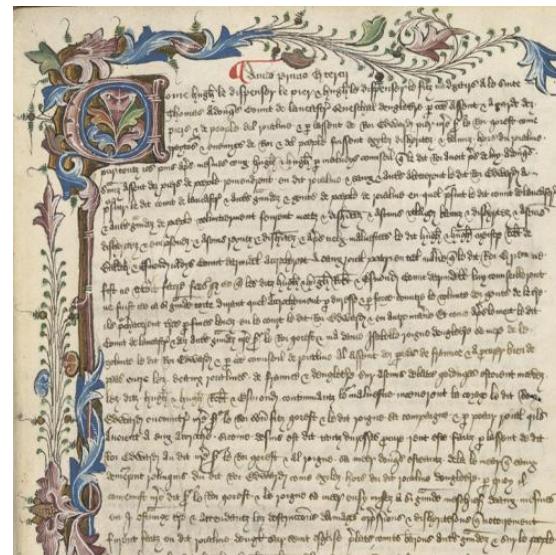
In considering with myself how it could have happened that a plow which pointed itself to me as a wrong, and future the robbery ruined & destroyed would not in this have been detected and carried into effect - the want of such a system of checks as might be adequate to the removal of this & other difficulties than may have occurred pointed itself as a natural course - as an as-  
surance &



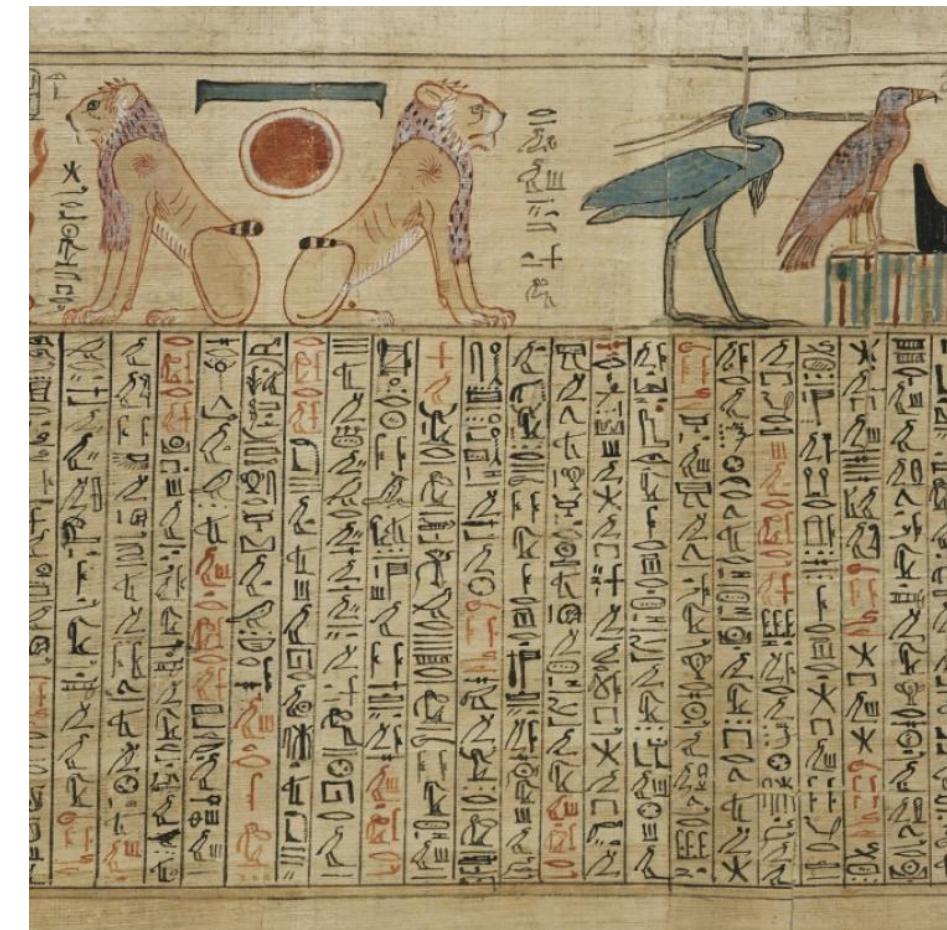
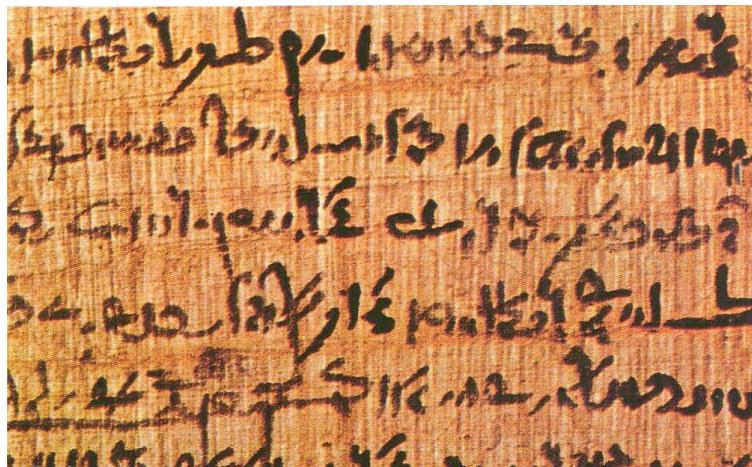
	N	P	O
A. at Lignac	1	Eys at Gienet	Eys at P. Lignac
B. at Rouen	2	Cord at V. Port	Cord at P. Lignac
Cognac at Alletta	3		
D. at V. Ares	4		
Hartland at Tinten	5	Eys at V. Loring	
Kalest at Berle	6	Plang at V. Langat	
Huge at Lapey	7	Ligiles at P.	
Narre at N.	8	Hornet at Guentane	
Narre at N.	9	Long at H.	
	10	Long at P. Lignac	
	11	Palomery at P.	
	12	Lois at P. Lignac	
	13	Portellat at Cernier	
	14	Maur. at N.	
	15	Perigat at Lata	
	16	Palen at Esse	
	17	Portlade at l. Robina	

My dear Doctor 2.1.11 26 May  
1836  
at night  
  
I hope you will be better  
and have rest Sunday - feel us  
the only day I have distinguished  
all the day strength  
  
You will oblige me by your answer  
as early as möglich tomorrow  
Yours ever  
Jeremy Bentham  
Dr. Southwood Lane

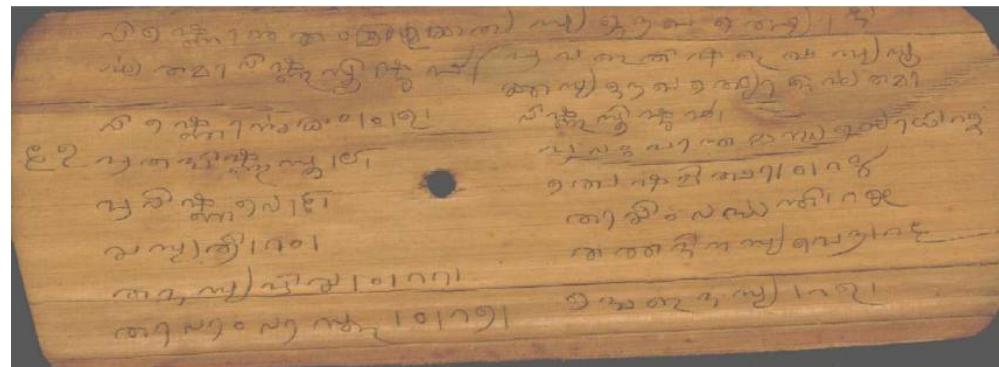
# Medieval Manuscripts



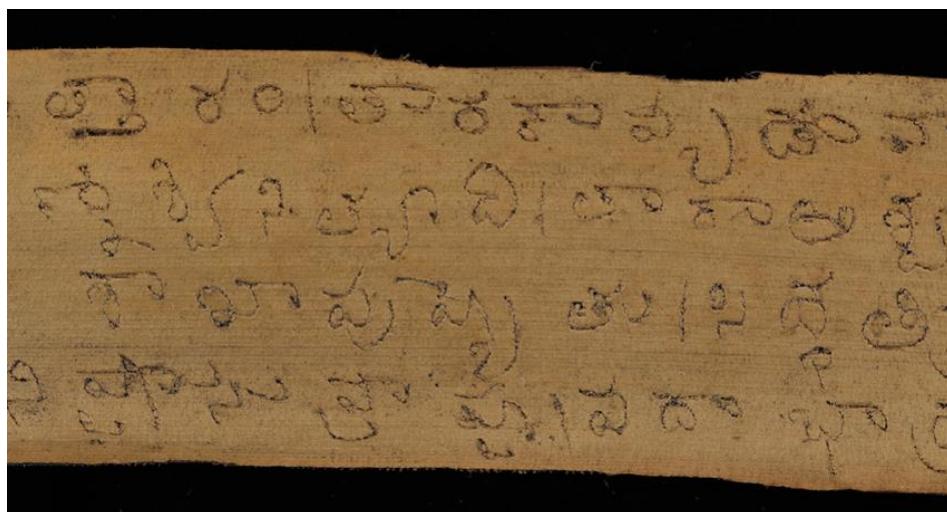
# Ancient Papyri



# Ancient Leaf Document



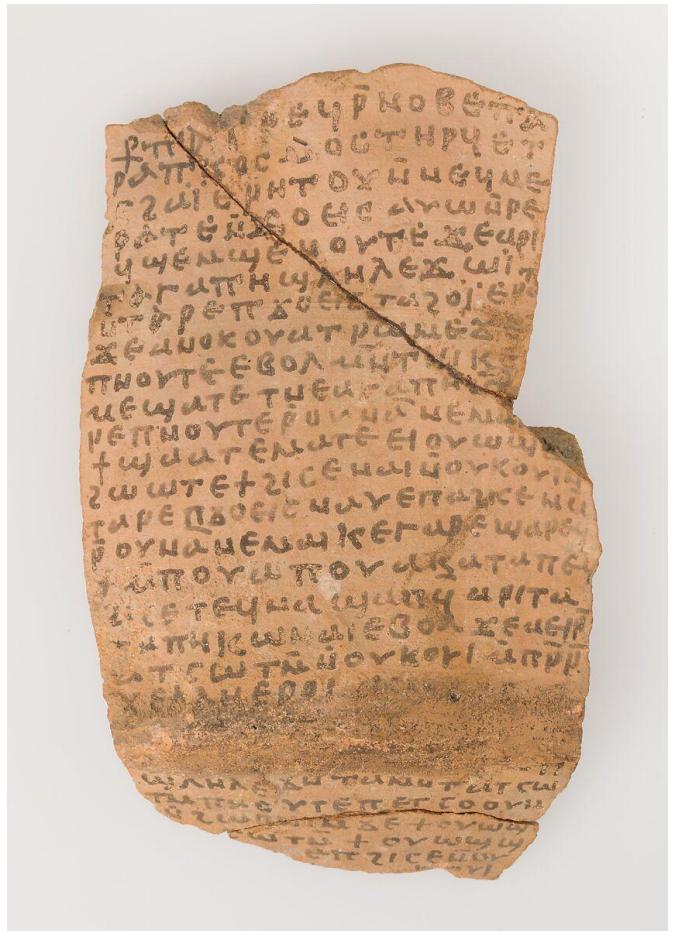
परमादित्यन् चक्रवृद्धकमग्नेश्वरसिंह  
देवतापूज्यामाना प्रयाणीयदित्यसुरय  
कर्त्तव्यद्वयामानामाना वरमध्येष्ठित  
मातृमरुष्टिर्जिनश्चावलिम्बुरुष्वरय  
मधुलोलिम्बुरुष्वरयद्वाविश्वरुष्वरन्म  
ठार्हित्यमनिनामयुवयित्ति एति द्विष्टुरुष्वार  
उविस्तुरुष्वाराग्नेयानवप्रदेवप्रमद्यते



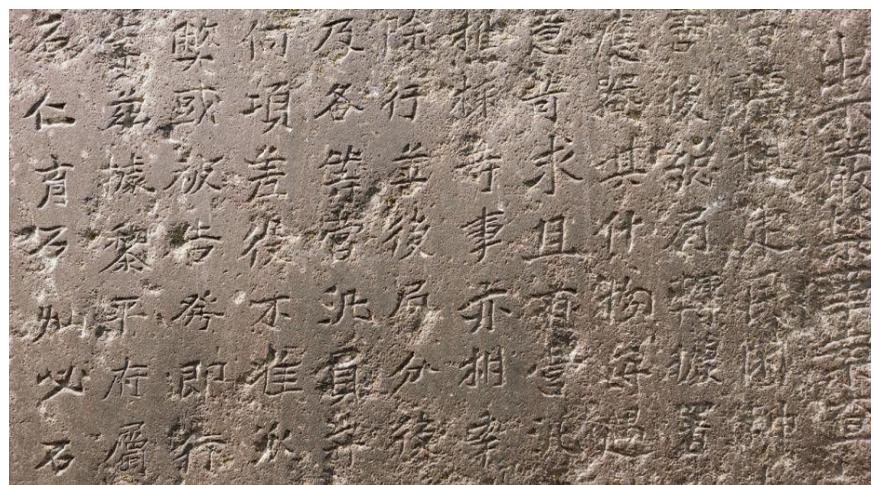
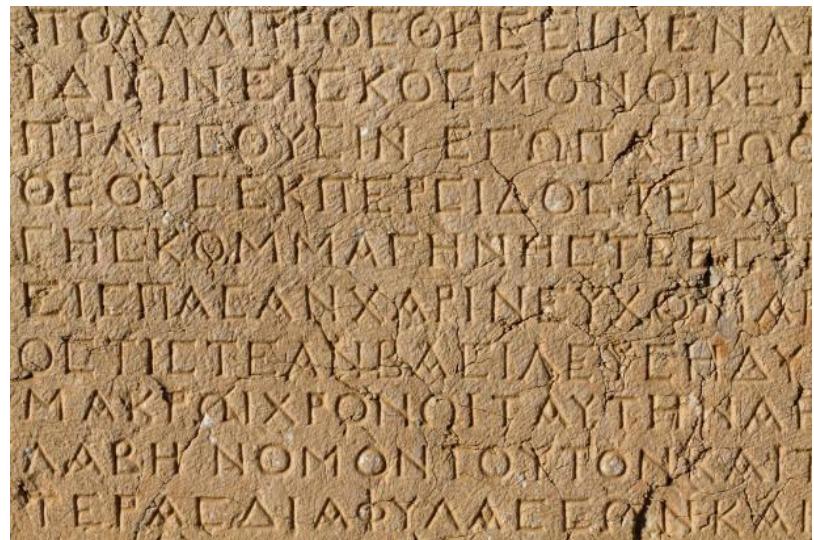
त्रिषुकारायन्त्रमृष्टेशुठम्याद्यत्पानात्  
यामां च क्षणात्तदुरुप्रसंगम्यत्पादिः  
लाभामावै सुवैश्वरिवा प्राप्तिवृन्दायया  
उद्भव्योवान्मासेवर्णनाट्यमानायत्पि  
श्यागापात्रवृग्निमध्यात्त्वायन्त्रित्यन्त्रुमः  
निदृष्टं विशेषान्त्रित्यन्त्रित्यन्त्रुमः  
तिमादितः ग्रन्तवृक्षाश्चाश्चालयन्त्रित्यन्त्रुमः ॥ असः  
ग्रामत्वा महिस्त्रिमात्रित्यन्त्रित्यन्त्रुमः ॥ क्षणा

लां श्राविष्णवज्ञेत्यन्तिर्गत्यस्तुः ३ चम्पदश्वरुद्रापिष्ठुराक्षः ५ कालकम्भनस्त्रावरुद्रतेलयद्वा  
यस्यनेत्रान्तिर्गत्येव द्वयायामयात्तिर्गत्येव तद्वदवृत्त्य यस्यस्तु तमस्याय वस्त्रद्वामानमाङ्गोचरपूर्वाद्य  
कर्त्तव्याद्वाग्निमण्डलायवनियोगमाना ५ मध्यायवास पितृत्वं आसान्त्वाक्षात्कृष्णमयोग्यत्वात्पृथुन  
यद्यर्थस्त्रिवाचलये १० एतदत्यायित्यद्यावत्वेन ३ यद्यत्वावेत्रान्तिर्गते यामांश्वर्गनमग्रममा गतिवाद् लं  
वित्तस्त्रिवाचलये अभ्यग्राहितः ५ रुद्रनवयवाद्वित्तु अत्यनुग्रहेत्वा यस्त्रद्वाविवृत्ताश्वात्पृथु स्त्रामावेन  
यस्त्रात्पृथिव्या कामादेवायाकामाद्वया ५ नामायपृथिव्यिवृत्तिरुद्धिः त्रिप्रलनानामायाद्विवृत्ताश्वात्पृथु  
त्तेषुक्षमविद्युत्त्वात्पृथिव्यामविवृत्तिः यमविवृत्तिः ५ यमसयमविवृत्तिरुद्धिः त्रिप्रलनानामायाद्विवृत्ताश्वात्पृथु  
त्तिरुद्धिः यमविवृत्तिरुद्धिः ५ यमसयमविवृत्तिरुद्धिः त्रिप्रलनानामायाद्विवृत्ताश्वात्पृथु

# Ostraca



# Inscriptions



# What are Historical Documents?

A historical document is not just “*an old book*” but **any medium** through which human beings have expressed meaning and recorded information, often using **very different materials and techniques**.

# Applications of Artificial Intelligence to Historical Documents

# Pipeline of Historical Document Processing

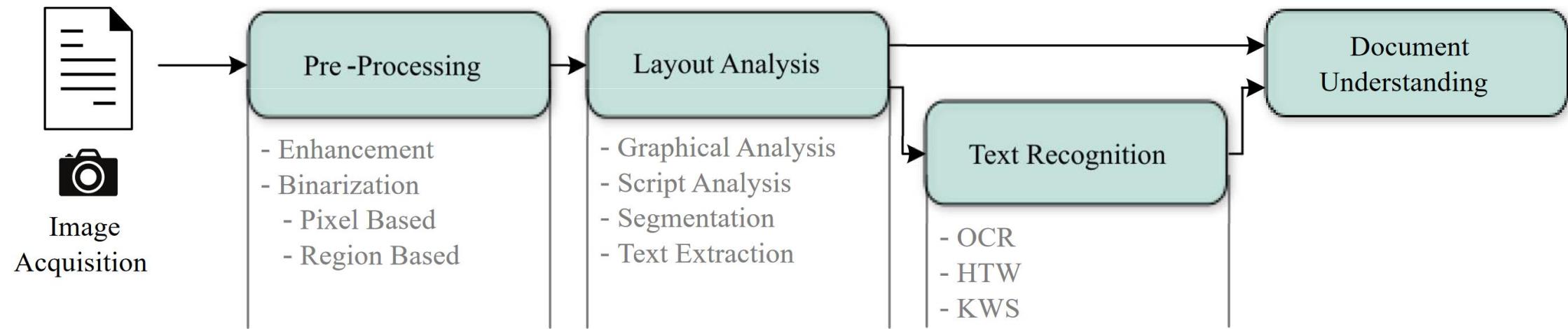


Figure 2.5: Typical Historical Document Processing Workflow.

# Pre-Processing - Binarization

lacrimas fundens p̄ gaudiū. dō & ei' beas.  
sime matrī laudes non modicas & multi-  
plies grārū actiones rependit. hec est  
laudabilis & uniuersalis impatiē ange-  
loꝝ & hominū uirgo sc̄issima. que licet  
a creatorē omniū creata sic. in ipsa crea-  
tione. l. 1. c. 1.

lacrimas fundens p̄ gaudiū. dō & ei' beas.  
sime matrī laudes non modicas & multi-  
plies grārū actiones rependit. hec est  
laudabilis & uniuersalis impatiē ange-  
loꝝ & hominū uirgo sc̄issima. que licet  
a creatorē omniū creata sic. in ipsa crea-  
tione. l. 1. c. 1.

lacrimas fundens p̄ gaudiū. dō & ei' beas.  
sime matrī laudes non modicas & multi-  
plies grārū actiones rependit. hec est  
laudabilis & uniuersalis impatiē ange-  
loꝝ & hominū uirgo sc̄issima. que licet  
a creatorē omniū creata sic. in ipsa crea-  
tione. l. 1. c. 1.

John Leedy  
J. Leedy  
Officer of  
Bowles

John Leedy  
J. Leedy  
Officer of  
Bowles

of government, is to do for  
whatever they need to have  
at all, or can not, so easily  
in their separate, answer-

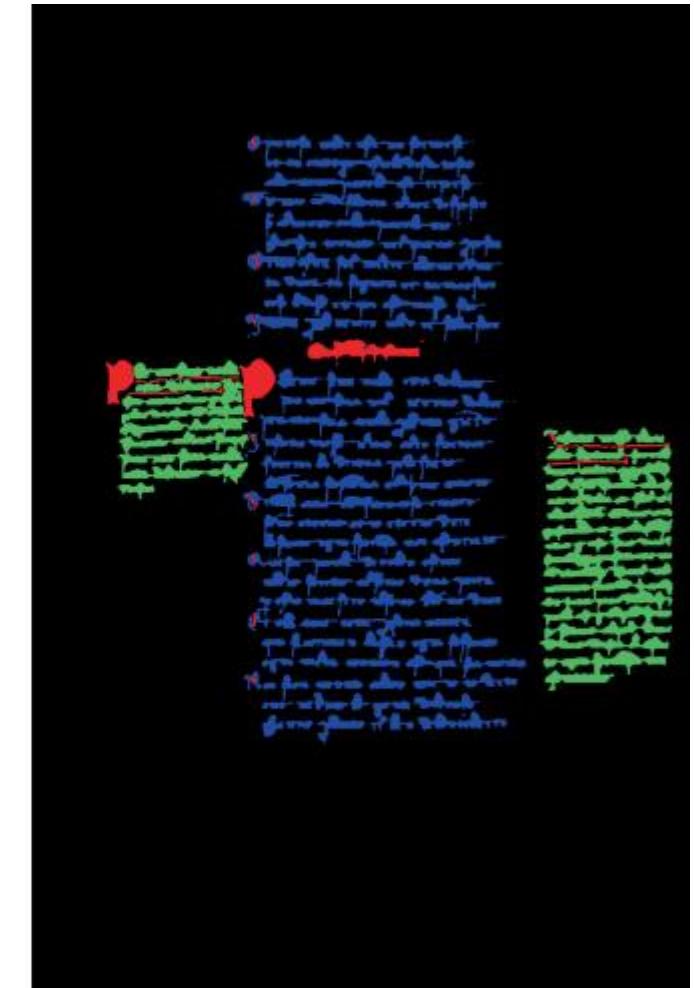
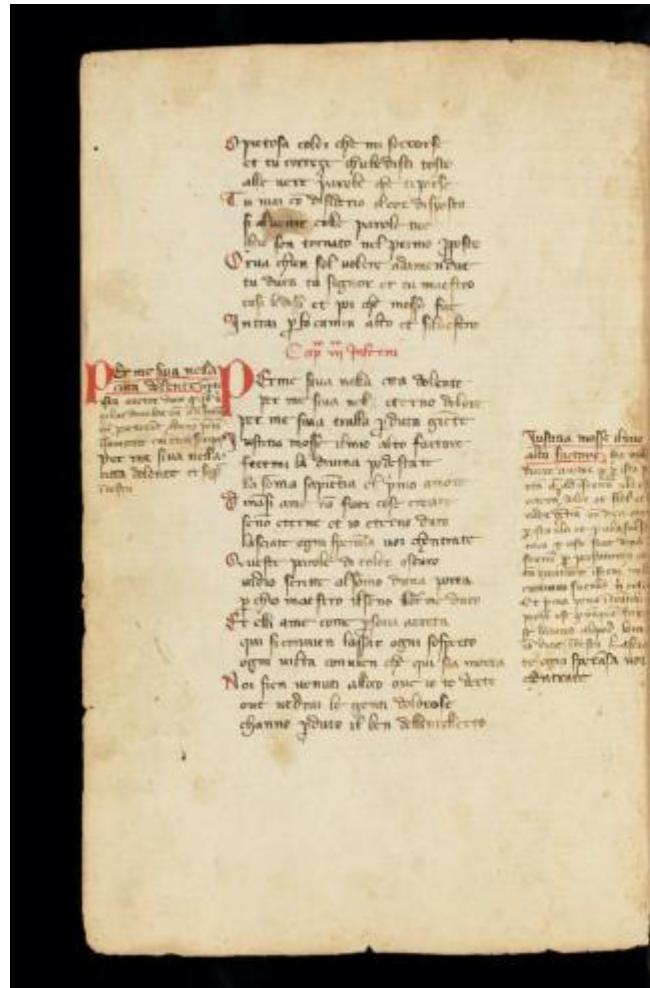
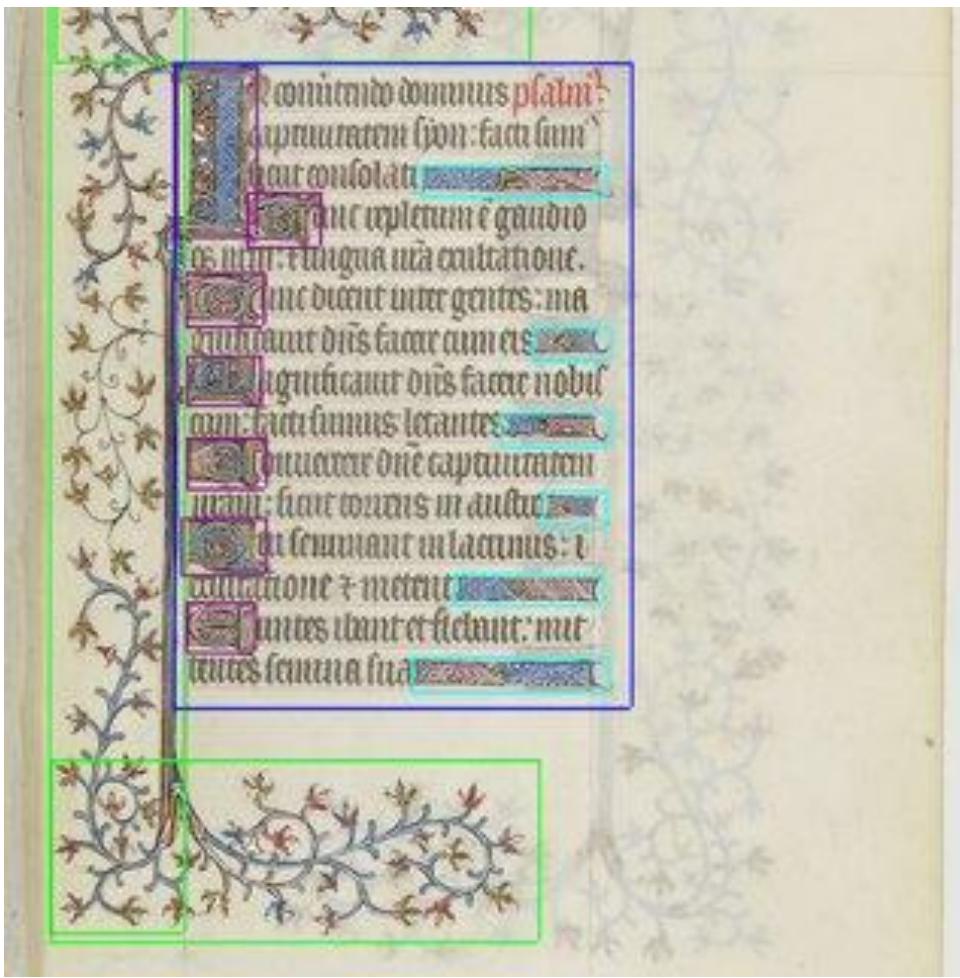
of government,  
whatever they  
at all, or can  
in their separate, answer-

# Pre-Processing - Binarization

## Deep Learning-Based Approaches (End-to-End)

- **U-Net** and variants: Encoder–decoder convolutional networks that learn to distinguish text from background. Widely used because they can capture both fine details and global context.
- **GANs** (Generative Adversarial Networks): Trained to "translate" the degraded image into a cleaner binarized version, reducing noise, speckles, and bleed-through.
- **Transformer-based** models: Some recent work explores Vision Transformers or Hybrid CNN–Transformer architectures to learn robust representations even with small datasets.
- **Self-supervised learning**: Useful when the ground truth is scarce (a typical problem in ancient documents). The networks first learn with pretext tasks (restoration, denoising) and then adapt to binarization.

# Layout Analysis – Page Segmentation



# Layout Analysis – Page Segmentation

## 1. CNN-based semantic segmentation

- **U-Net variants (U-Net++, Attention U-Net)** → widely used to distinguish text, decorations, margins and background.
- **Fully Convolutional Networks (FCN)** → treat the page as a pixel classification problem.

## 2. Object Detection & Instance Segmentation

- **Faster R-CNN, Mask R-CNN, YOLOv5/YOLOv8** → used to segment blocks of text, figure boxes, lines, or even individual words.

## 3. Transformer-based models

- **DETR (Detection Transformer)** → They exploit global attention to locate textual and graphic regions in historical documents.

## 4. Graph-based approaches

- **Graph Neural Networks (GNN)** to model the logical structure of the document: nodes = regions, arcs = spatial relationships.

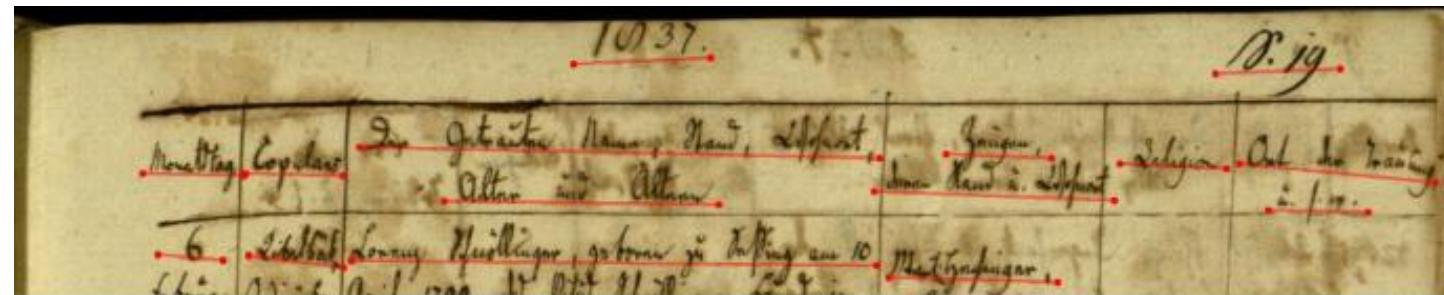
## 5. Self-supervised approaches

- **Self-supervised pretraining** (autoencoder, contrastive learning) → useful for small datasets.

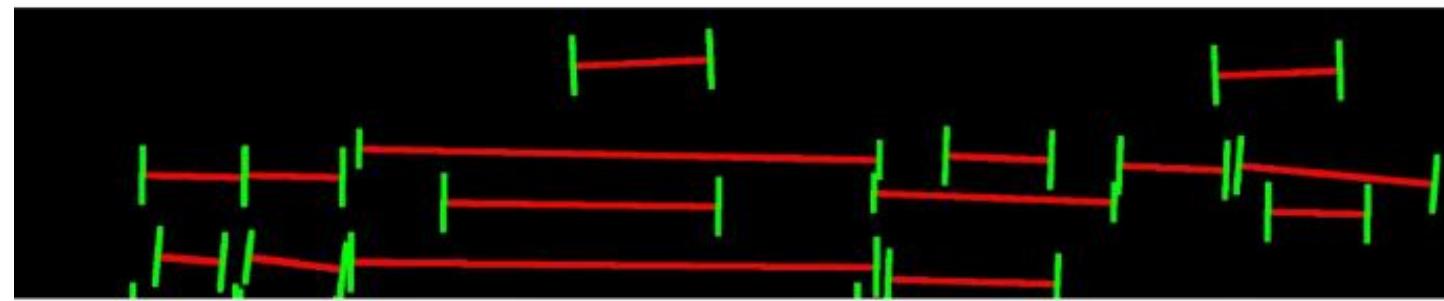
# Layout Analysis – Line Segmentation

A piece of the ruined letter  
house, you have got here.  
Your nearest bank expect  
you, until we hear from  
G. Carter has just called  
now, it is too bad. Ottoman  
lettis by express did see  
thank you again for  
it and now, the total  
loss, there are so many o,

# Layout Analysis – Baseline Detection



**(a) Baseline ground truth** – The baselines are defined by the red dots. Dots of the same baseline are connected.



**(b) Pixel ground truth** – Green encodes the separator class, red the baseline class and black the "other" class. See supplements for an algorithm to automatically generate such GT.

# Layout Analysis – Line Segmentation

## 1. CNN-based semantic segmentation

- **FCN / U-Net segmentation:** semantic segmentation networks classifying pixels as “baseline” vs. “background.”.

## 2. Object Detection & Instance Segmentation

- **Object Detection models (YOLO, Faster R-CNN, Mask R-CNN)** → detect bounding boxes or masks for text lines.

## 3. Transformer-based models

- **DETR (Detection Transformer) adapted for lines** → directly detects baselines as segments..

## 4. Skeletonization & contour-based methods

- Use CNNs or morphological algorithms to extract the “skeleton” of the writing → baseline curves can be extracted even in degraded manuscripts.

# Text Recognition and Transcription

Transkribus

The image shows a screenshot of the Transkribus software interface. On the left, a vertical scroll bar indicates the document is long. The main area displays a handwritten poem in blue ink on aged paper. The text is transcribed in a clean, modern font on the right. Several words and phrases are highlighted with colored boxes: 'Adress to dear Isabella on the Authors recovery' (green), 'How often did I think of you' (purple), 'Good care Im sure was of me taken' (blue), 'At last I daily strenght did gain' (green), 'Stay in the Parlour till the night' (blue), 'This was wrote by Marjory Fleming' (green), 'on Sunday the 15<sup>th</sup> of December 1811' (blue), and 'died on Thursday the 19<sup>th</sup> December 1811' (blue). The background of the transcription area has horizontal stripes corresponding to the highlighted regions in the original text.

Adress to dear Isabella on the  
Authors recovery

Oh Isa pain did visit me  
I was at the last extremity  
How often did I think of you  
I wished your graceful form to view  
To clasp you in my weak embrace  
Indeed I thought I'd run my race  
Good care Im sure was of me taken  
But indeed I was much shaken  
At last I daily strenght did gain  
And then diminish did my pain.  
At last the Doctor thought I might  
Stay in the Parlour till the night  
This was wrote by Marjory Fleming  
on Sunday the 15<sup>th</sup> of December 1811  
and she intended to finish it the next  
day when she was taken ill and  
died on Thursday the 19<sup>th</sup> December 1811.

# Text Recognition and Transcription

## 1. Historical OCR

- Based on the possibility of subdividing words into characters and the following classification for their recognition.

## 2. Handwritten Text Recognition (HTR)

- CNN–LSTM hybrids trained with CTC loss
- Transformer-based models

## 3. Visual Encoding – Textual Decoding

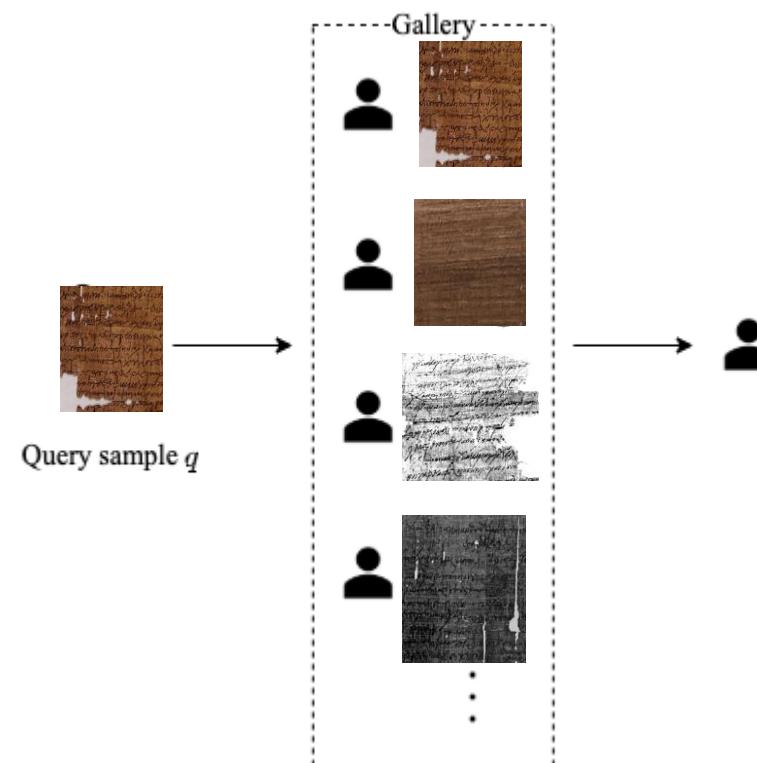
- Transformer-based OCR. Combines a Vision Transformer (ViT) encoder with a Transformer-based text decoder, trained end-to-end in a sequence-to-sequence fashion.

## 4. Vision Language Models

- such as **Donut** or **BLIP-2**, are increasingly adapted to historical scripts, enabling in-context or few-shot transcription without large amounts of training data

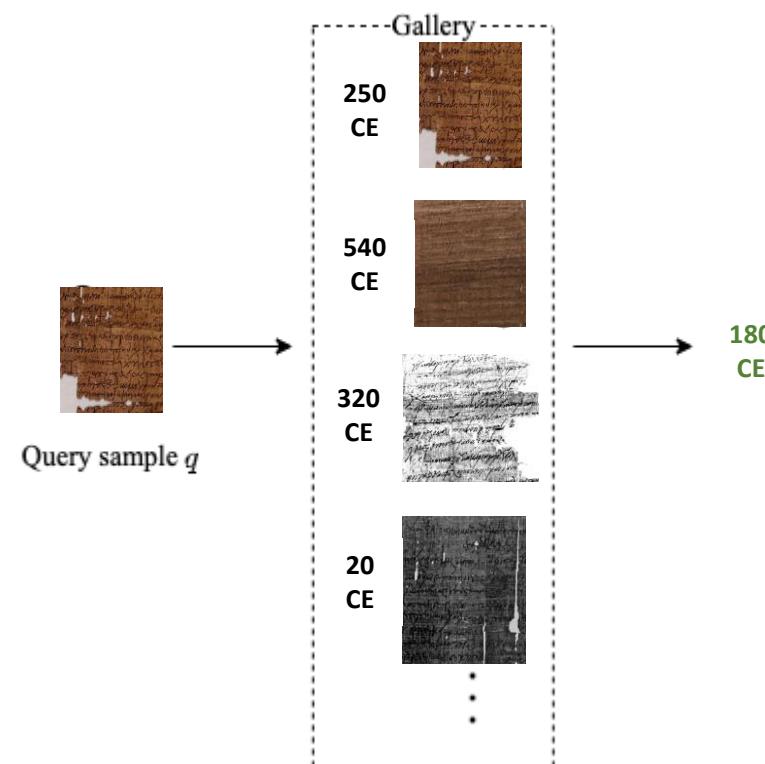
# Classification and Attribution – Writer Identification

- Identify the hand of a particular scribe

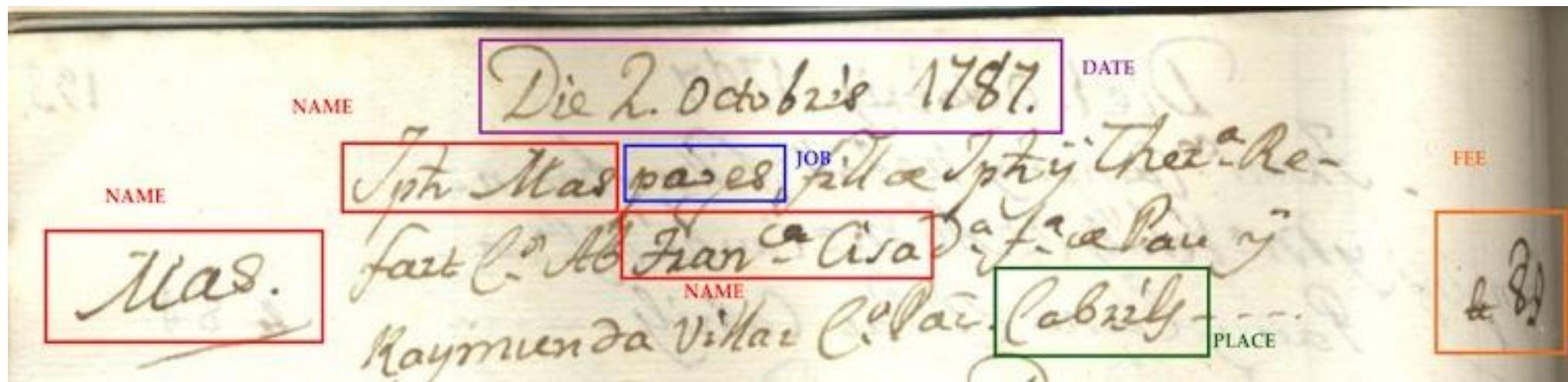


# Classification and Attribution – Dating

- Estimate the creation date of a document



# Linguistic and Semantic Analysis



# Linguistic and Semantic Analysis

## 1. Named Entity Recognition (NER)

- Detecting **people, places, institutions, dates** in historical texts.
- Often trained with **domain-adapted models** (e.g., spaCy/BERT fine-tuned on historical corpora).

## 2. Information Extraction

- **Relation extraction:** linking names, events, and places (e.g., in charters or contracts).
- **Event detection:** identifying historical events from chronicles or administrative records.

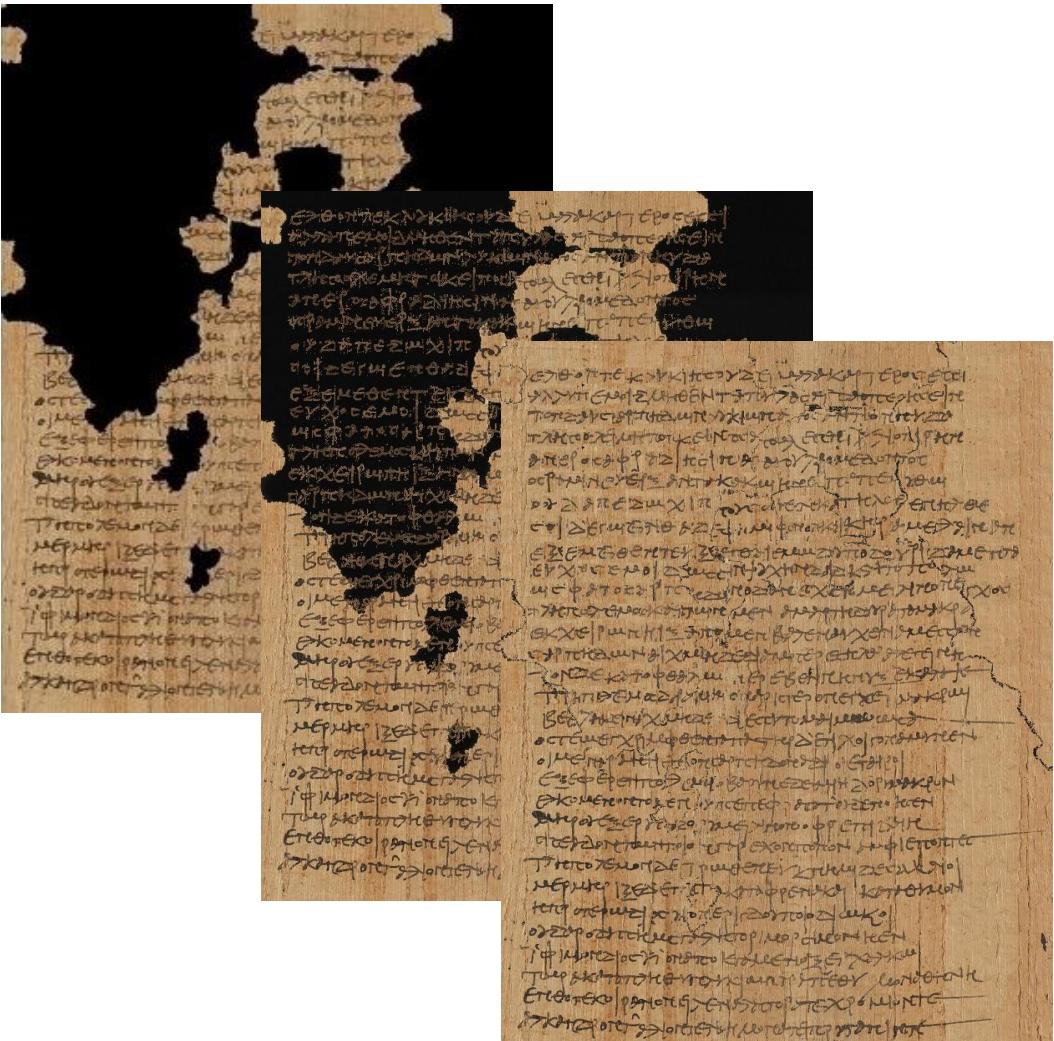
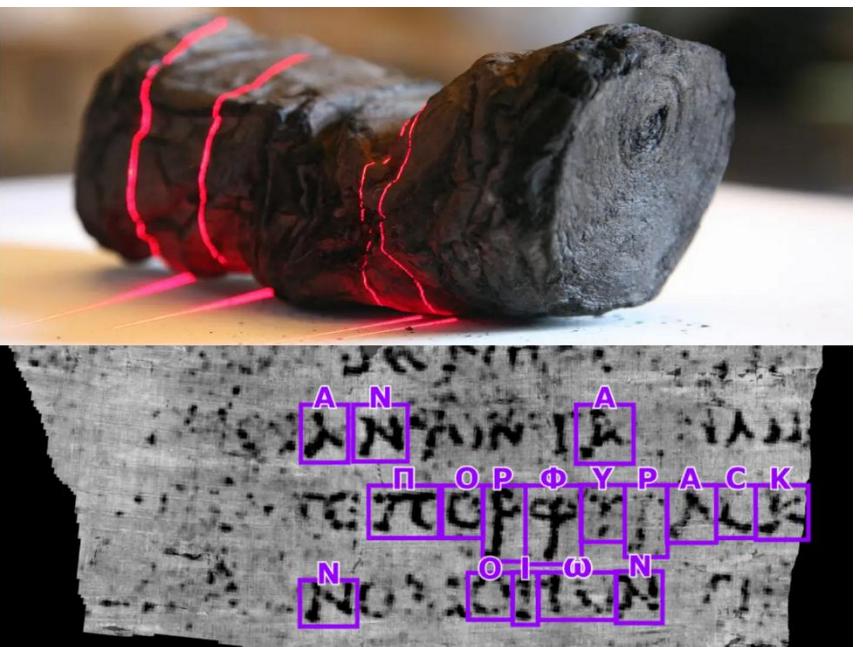
## 3. Cross-lingual & Diachronic Analysis

- **Cross-lingual embeddings** → align different languages (e.g., Latin and Greek corpora).
- **Diachronic embeddings** → track semantic change across time.

## 4. Knowledge Graphs & Linked Data

- Converting extracted entities into **knowledge graphs** (e.g., people, places, events connected).
- Enables integration with other historical datasets (archaeology, prosopography, epigraphy).

# Restoration



# Restoration

## 1. Inpainting & Hole Filling

- **GAN-based inpainting:** learns to reconstruct plausible text patterns or parchment texture.
- Used for **fragmented or damaged manuscripts**.

## 2. 3D Imaging & Tomographic Reconstruction

- **X-ray phase-contrast tomography** or micro-CT scans → used to “virtually unroll” scrolls (e.g., carbonized papyri from Herculaneum – Vesuvius Challenge).

## 3. Deep Learning Restoration Models

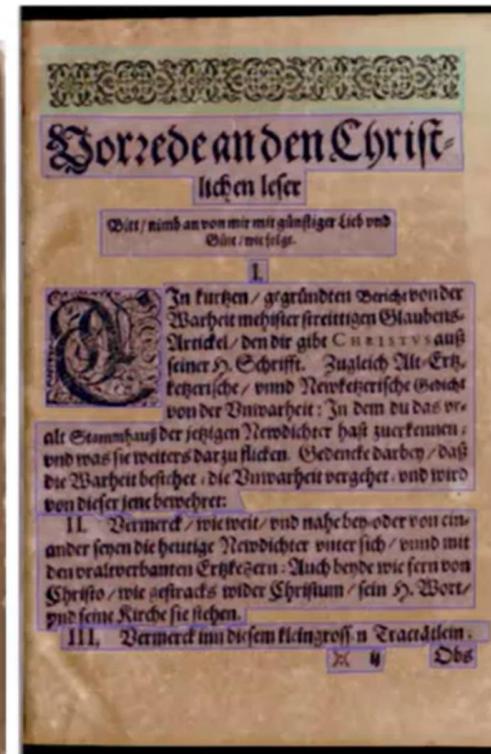
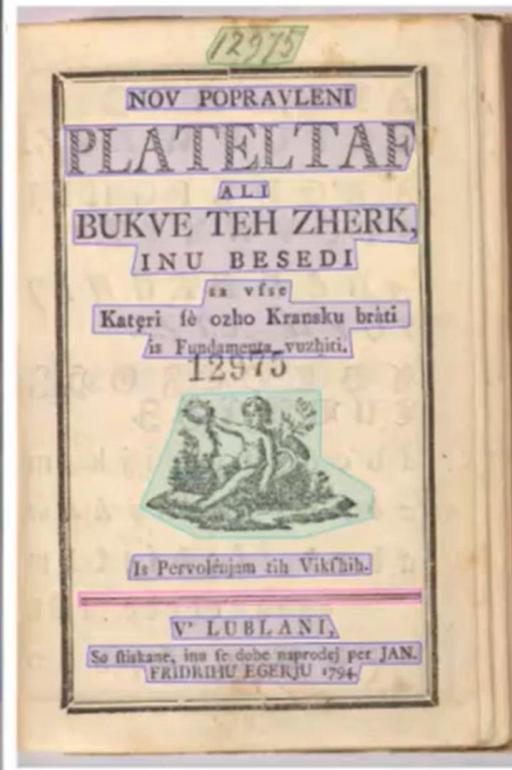
- **Autoencoders** for denoising and reconstruction of missing text.
- **GANs (pix2pix, CycleGAN)** → learn to translate degraded images into restored ones.
- **Diffusion models** (emerging trend) for generative restoration with uncertainty estimation.

# Some Datasets

# IMPACT

Early Modern to 20<sup>th</sup> Century - Layout Analysis - Different European languages

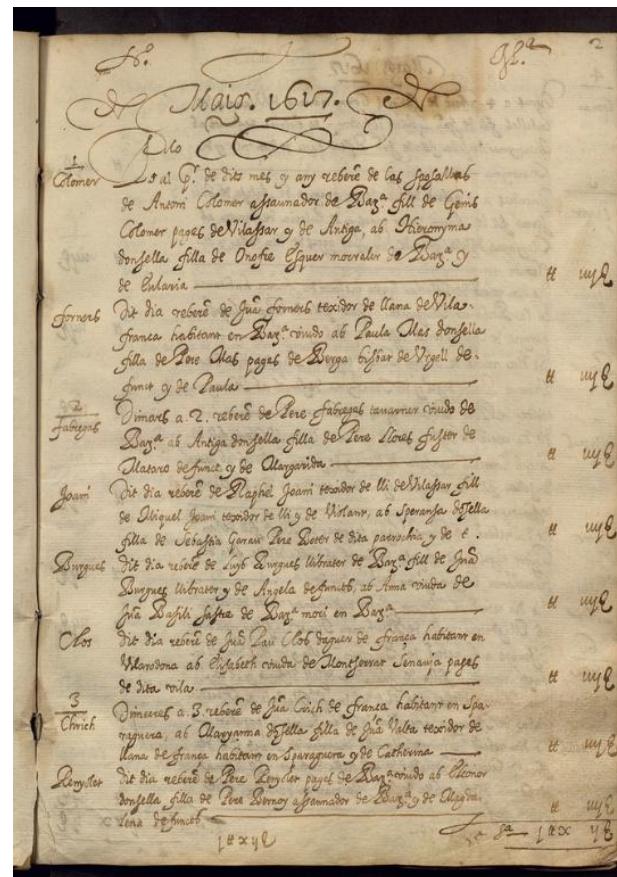
<https://www.digitisation.eu/impact-dataset/>



# Esposalles

14<sup>th</sup> to 20<sup>th</sup> Century - Semantic Analysis - Spanish

<https://dag.cvc.uab.es/dataset/the-esposalles-database/>



Llibre d'Esposalles	
1491 - 1493	
A.	
Albarell ab V. Pages	2
Armenyol ab Cornell	3
Altet ab Llob	4
Arboix ab N.	5
Anger ab Vivent	6
Albran ab V. Llunel	7
Arnau ab V. Borrell	8
Atelles ab N.	9
Albarell ab Zapenes	10
Arabay ab N.	11
Annat ab V. Oliva	12
Albarrell ab V. Folch	13
Araxari ab N.	14
Arrosset ab Bonel	15
Alcoba ab V. Galer	16
Amat ab Clerver	17
Alberguer ab Barba	18
Avinyó ab V. Salduori	19
Ariñera ab Basaldu	20
Argenyol ab Bonca	21
Artan ab N.	22
Agusti ab Reig	23
Acetamir ab Janer	24
Almenara ab Cartora	25
Aradiol ab Burquet	26
Amat ab Santit	27
Alfonso ab Duran	28
Armant ab Boygas	29
Agostí ab N.	30
Amella ab Pardols	31
Ameda ab V. Bosch	32
Arriach ab Gual	33
Affibanyana ab Varat	34
Arch ab Sala	35
Alegre ab Pastaller	36
Alegria ab V. Messara	37
Alemany ab Bosch	38

Die 2. October 1787. DATE

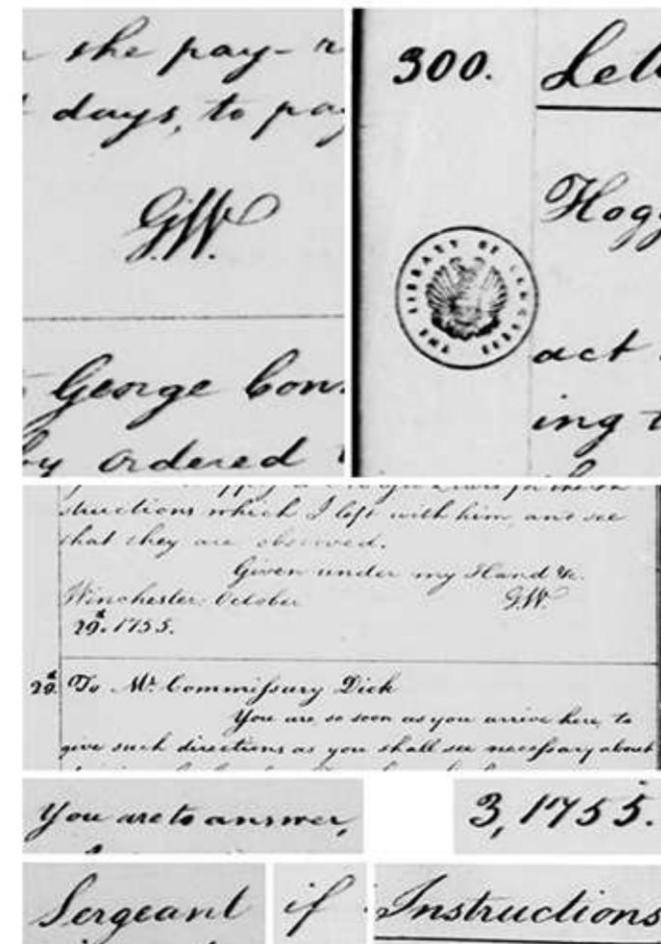
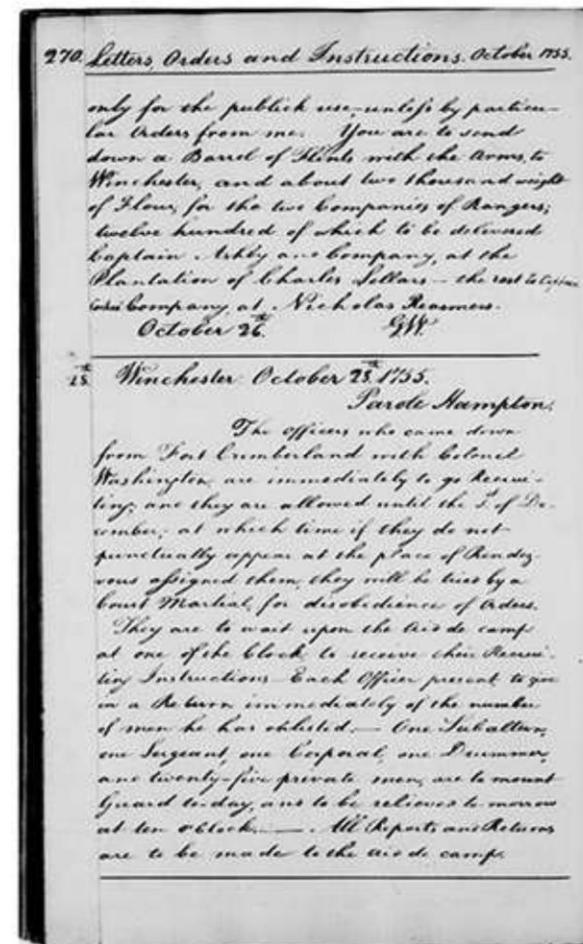
Sp. Mas	pares	JOB
Sp. Mas	Françisa	PLACE

Raymunda Villar Cabral Cabril

# Washington

18<sup>th</sup> Century - Keyword Spotting / HTR - English

<https://fki.tic.heia-fr.ch/databases/washington-database>



# Bentham Collection

18<sup>th</sup> Century - HTR - English

<https://zenodo.org/records/44519>



# HisIR19

Medieval (9<sup>th</sup> to 15<sup>th</sup>) - Image Retrieval - Different European languages

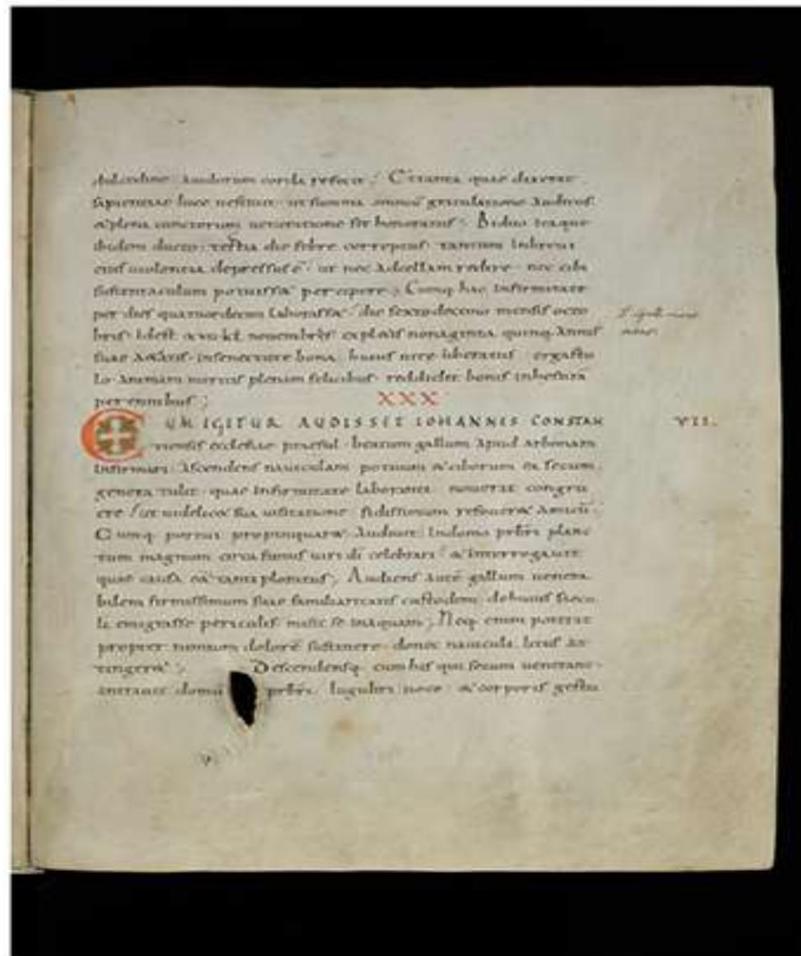
[https://tc11.cvc.uab.es/datasets/HisIR19\\_1](https://tc11.cvc.uab.es/datasets/HisIR19_1)



# Saint Gall

9<sup>th</sup> century - HTR / Line segmentation - Latin

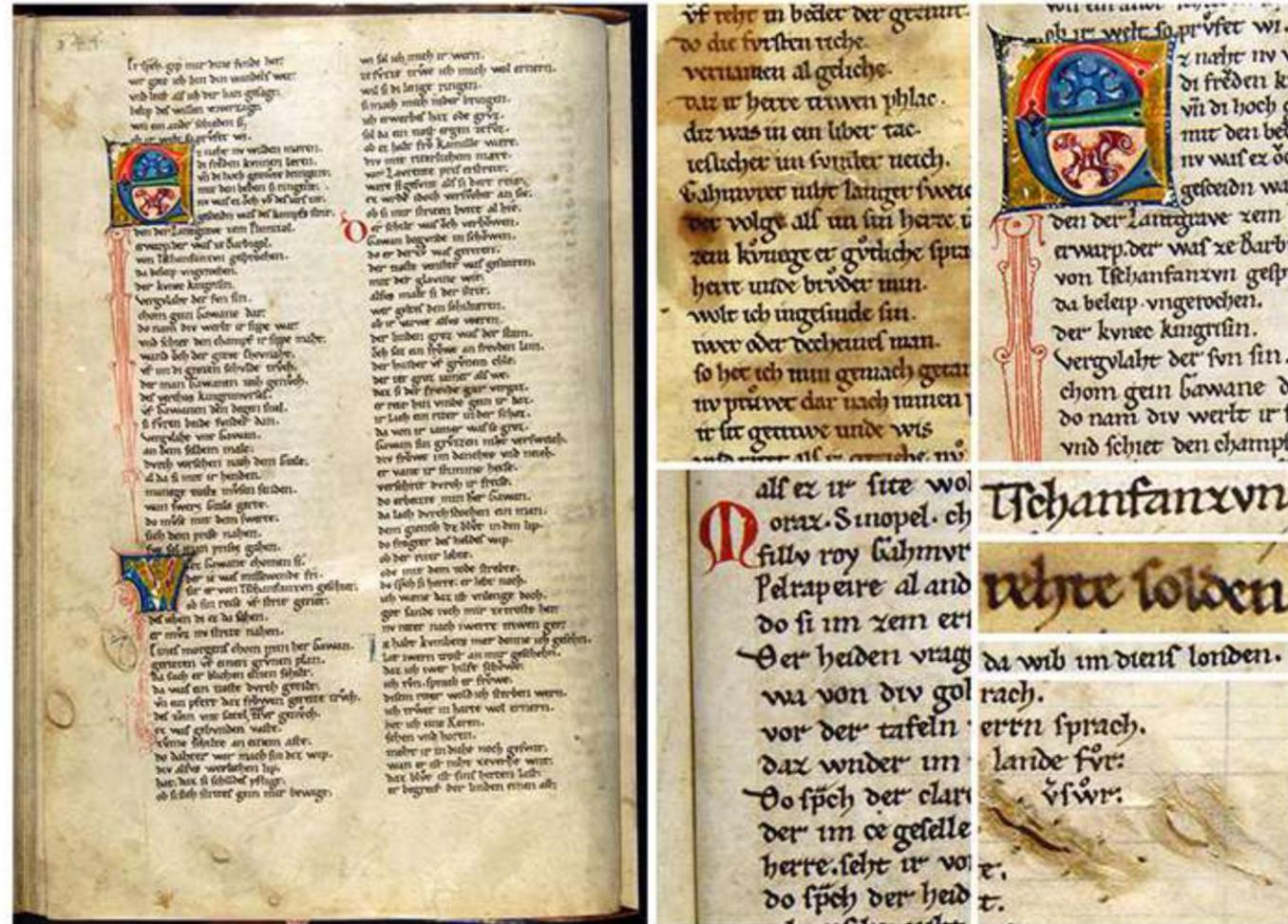
<https://fki.tic.heia-fr.ch/databases/saint-gall-database>



# Parzival

## 13<sup>th</sup> to 15<sup>th</sup> century - HTR - Old German

<https://fki.tic.heia-fr.ch/databases/parzival-database>



# M5HisDoc

13<sup>th</sup> to 15<sup>th</sup> century - HTR - Chinese

<https://github.com/HCIILAB/M5HisDoc>



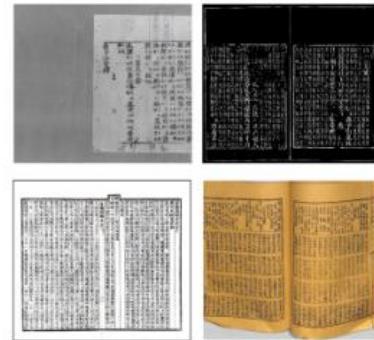
(a)



(c)



(b)



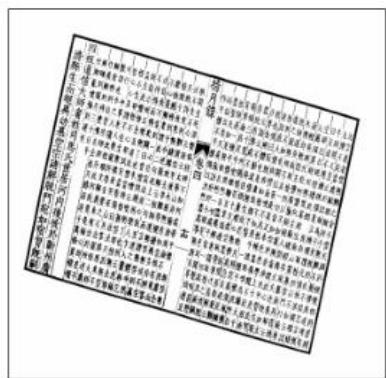
(d)



1



(e)

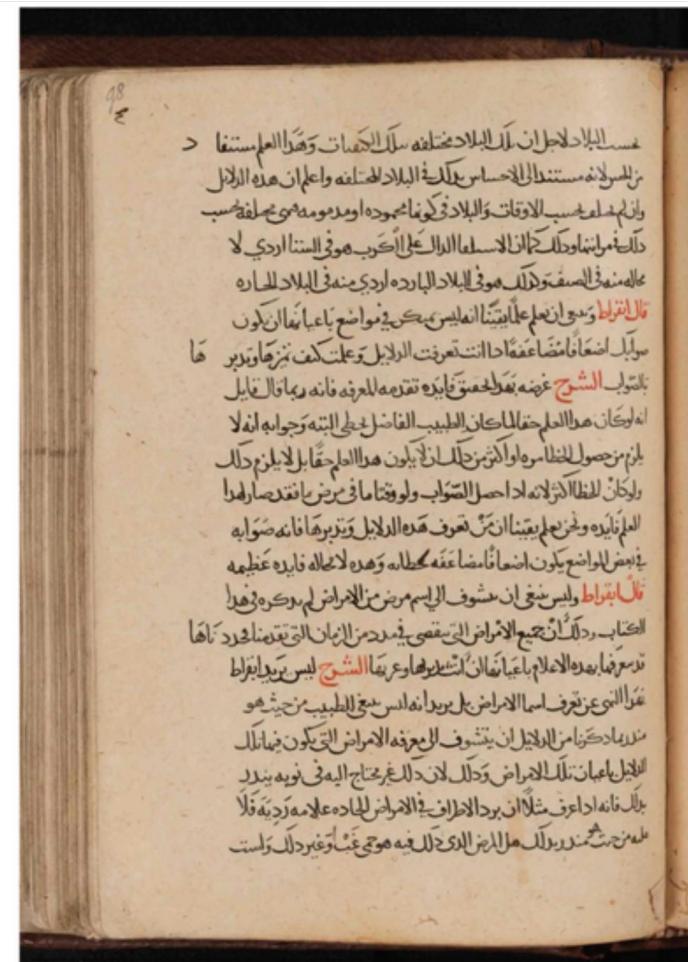


論百卷

# KERTAS

1<sup>st</sup> to 14<sup>th</sup> century - HTR / Dating - Arab

<https://qspace.qu.edu.qa/handle/10576/12024>



# GRK-Papyri

6<sup>th</sup> century - Writer Identification - Ancient Greek

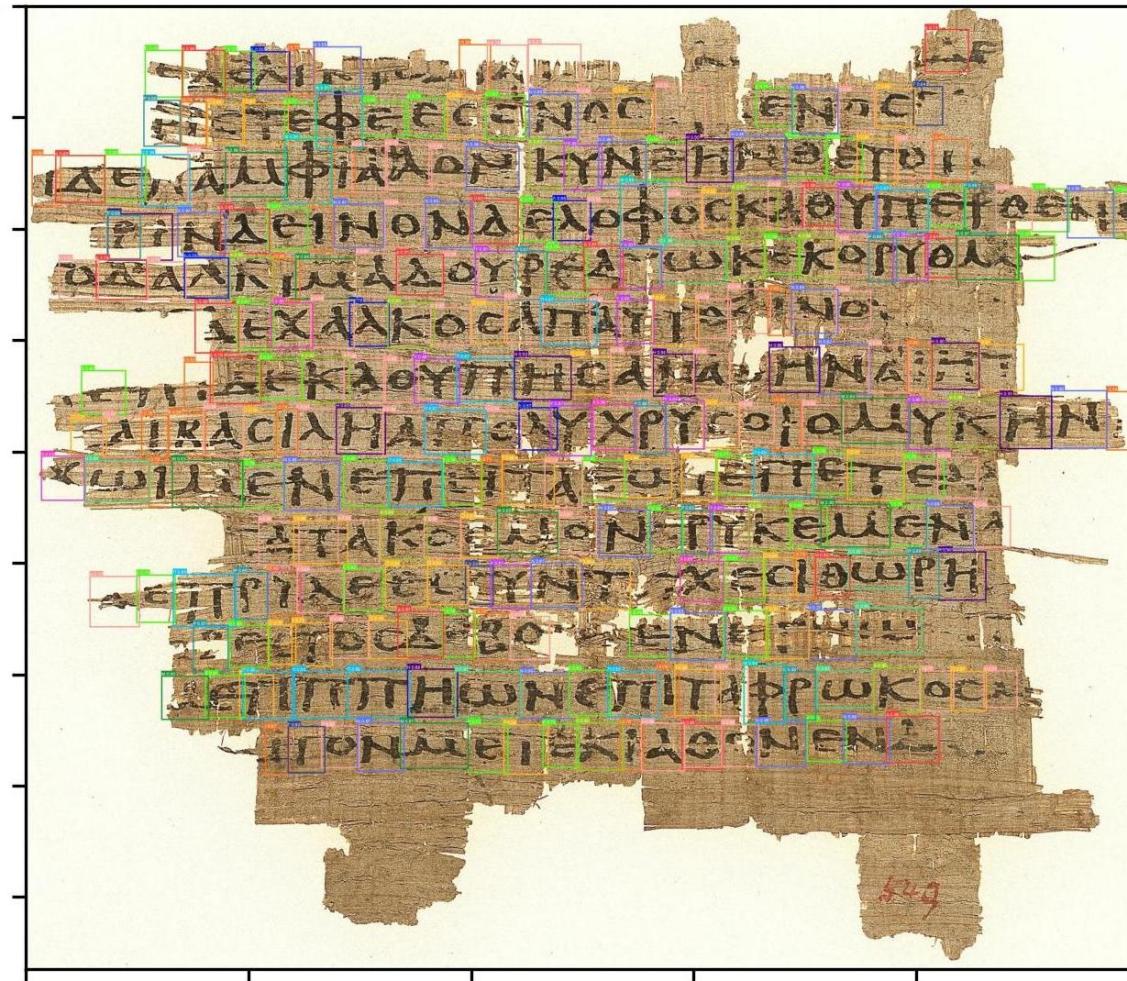
<https://d-scribes.philhist.unibas.ch/en/gkr-papyri/>



# DROGLOP

1<sup>st</sup> to 4<sup>th</sup> century - Character Detection - Ancient Greek

<https://qspace.qu.edu.qa/handle/10576/12024>



# HellDate

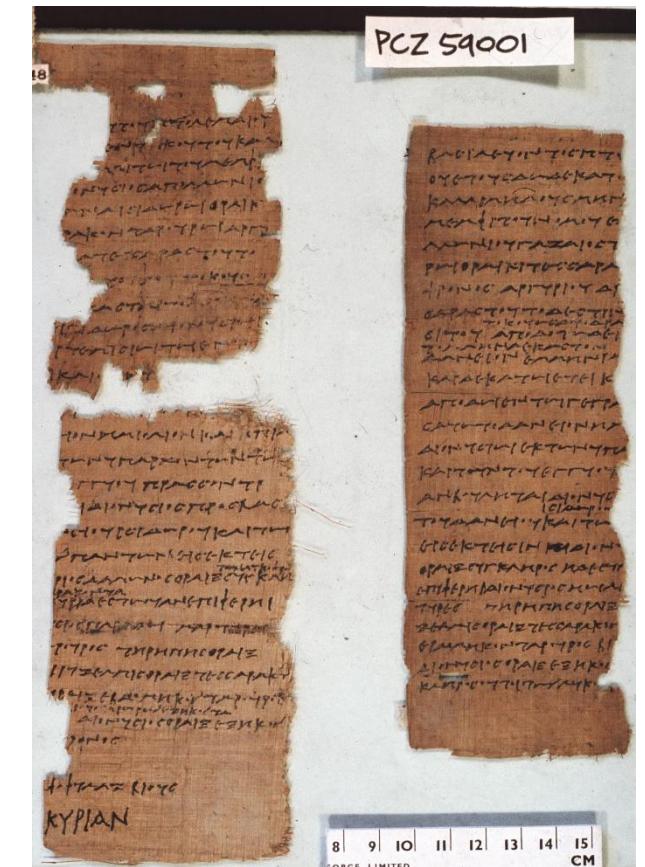
3<sup>rd</sup> to 1<sup>st</sup> century BCE - Dating - Ancient Greek

<https://zenodo.org/records/15830980>



<http://berlpap.smb.museum/02466/>

gefördert von der Deutschen Forschungsgemeinschaft DFG



# Other Datasets

- New datasets are constantly being published.
- ICDAR competitions are often organized about historical documents.
- Online archives of historical documents are available.

# Online Archives - Be careful

<https://papyri.info/>

## Papyri.info

[sign in](#)

Browse: [DDbDP](#) [HGV](#) [APIS](#) [DCLP](#) [Authors](#) [TM Number](#) or Search: [Data](#) [Bibliography](#)

**Papyri.info** has two primary components. The **Papyrological Navigator** (PN) supports searching, browsing, and aggregation of ancient papyrological documents and related materials; the **Papyrological Editor** (PE) enables multi-author, version controlled, peer reviewed scholarly curation of papyrological texts, translations, commentary, scholarly metadata, institutional catalog records, bibliography, and images.

Papyri.info aggregates material from the Advanced Papyrological Information System (APIS), [Duke Databank of Documentary Papyri \(DDbDP\)](#), [Heidelberger Gesamtverzeichnis der griechischen Papyrusurkunden Ägyptens \(HGV\)](#), [Bibliographie Papyrologique \(BP\)](#), and depends on close collaboration with [Trismegistos](#), for rigorous maintenance of relationship mapping and unique identifiers.

Search the navigator

### Partners

[APIS](#)[DDbDP](#)[HGV](#)[BP](#)[Trismegistos](#)[APD](#)

Contribute content

### More information

[Checklist of Editions](#)[papyrological resources](#)[send feedback](#)

# Online Archives - Be careful

<https://papyri.info/ddbdp/p.col;8;215>

## DDbDP transcription: p.col.8.215 [xml]

ca. AD 100 ?  
 [Reprinted from: [sb.5.7660](#)] SB5,7660

r,ctr

Απλονοῦς Θεο[μουθ]ᾶτι  
 τῇ μητρὶ πλεῖσ[τ]α χα[ίρε]ιν·  
 πρὸ μὲν πάντων εὐχόμεθά σε  
 3,ms αρα  
 ὑγενιν(\*) σὺν Απλοναρίῳ. Θέλω[σ]ε γι-  
 5 νώσκιν(\*) ὅτι ἡκουσα παρὰ τῶν {ο} ή-  
 κώτων(\*) μοι ὅτι ἡσθένηκος(\*),  
 ἔχαρην δὲ ἀκούσασ[α] ὅτι κωμ-  
 σῶς(\*) ἐσχηκος(\*). ἐρωτῶ σε μεγά-  
 λως καὶ παρακαλῶ, ἐπιμέλου  
 10 ἔατης(\*) ἄμα καὶ τῆς μικρᾶς ὡς  
 παρέλθητε τὸν χιμῶνα(\*), εἴ-  
 να(\*) εὔρομον(\*) ἡμᾶς(\*) υἱένωντος(\*).  
 καὶ είμις(\*) γὰρ πάντος(\*) ὑγε[ν]ιωμον(\*).  
 καὶ περὶ τῆς Συρίας ἔ[ω]ς ἀρτὶ οὐδὶ[έ]ν.  
 15 κακόν. ἐρωτῶ σε ἐὰν ἀκούσῃς πε-  
 ρὶ [Τ] Θεομουθᾶτος πέμψον μοι φάσ[ι]ν.  
 ἐρωτῶ σε, οὐ πρᾶγμά ἐστιν, ἐάν  
 τινα εὑρηται(\*) καταβαί[] νωντα(\*),  
 ἀποστίλε(\*) ὑμῖν(\*) φάσιν περὶ τῆς  
 20 ὑγείας(\*) ἡμῶν(\*) καὶ τῆς μικρᾶς.

## Image [[open in new window](#)]



**Notice:** Each library participating in APIS has its own policy concerning the use and reproduction of digital images included in APIS. Please contact the [owning institution](#) if you wish to use any image in APIS or to publish any material from APIS.

# Online Archives - Be careful

## The Sigma Case

$$\Sigma$$

According to some analysis:

*«the letter sigma ( $\Sigma$ ) results in being very rare in the Hellenistic-Roman period...»*

# Online Archives - Be careful

## The Sigma Case

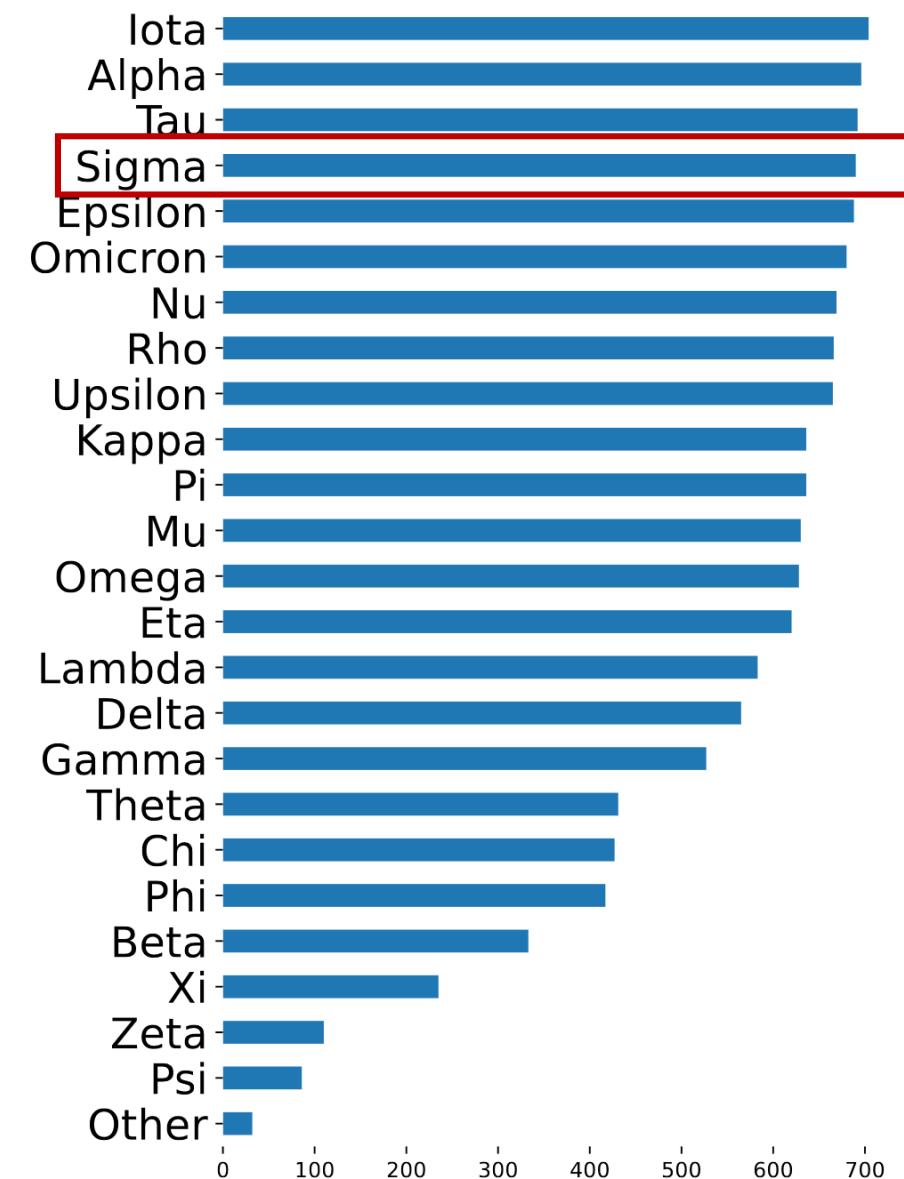


Figure 1. Letter frequency in our Hell-Char dataset



GitHub Tutorial:

<https://github.com/giuseppedeg/ICDAR25---Tutorial-Historical-Documents-in-Focus>

GliFix Tool:

<https://glyfix.scicore.unibas.ch/>

