Historical Documents in Focus:

Visual and Computational Analysis from Papyri to Inscriptions

Giuseppe De Gregorio: giuseppe.degregorio@unibas.ch

ICDAR2025

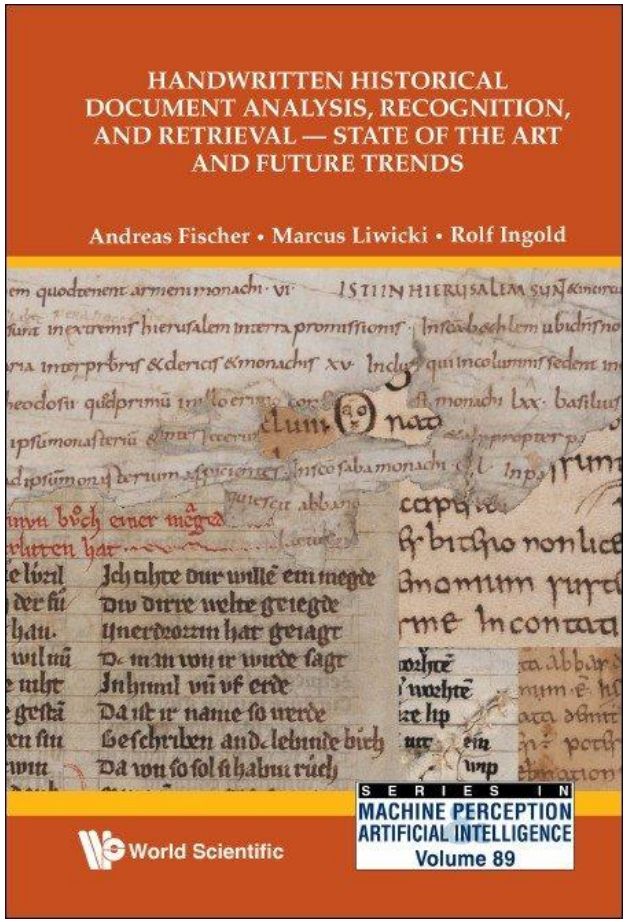
Swiss National Science Foundation

Universität Basel

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  A brief introduction to what historical document analysis is and what the challenges and opportunities it offers.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

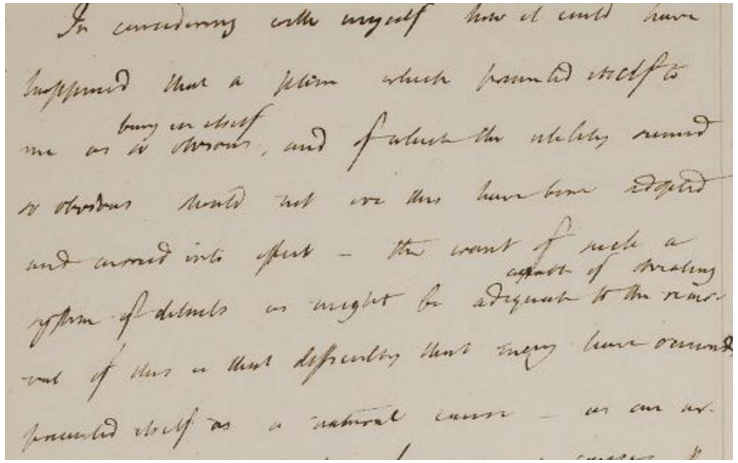


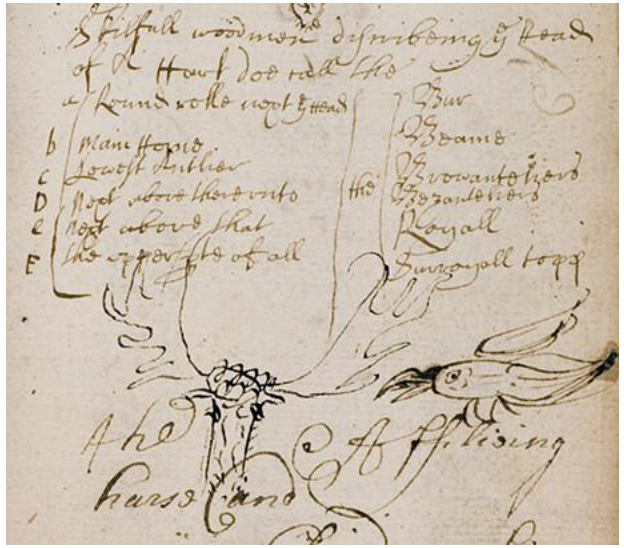
Handwritten Historical Document Analysis, Recognition, and Retrieval — State of the Art and Future Trends
Ed. Andreas Fischer, Marcus Liwicki and Rolf Ingold

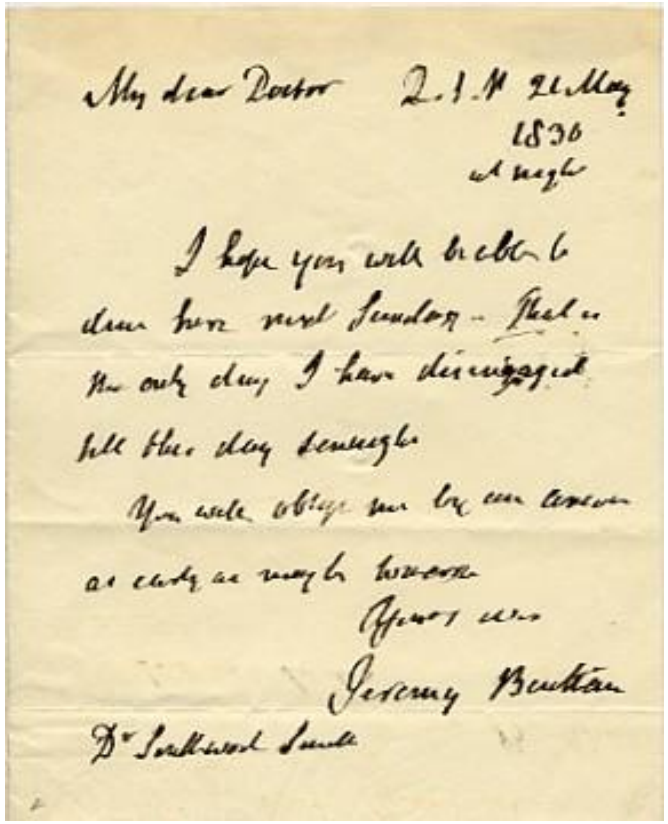
https://doi.org/10.1142/11353

What are Historical Documents?

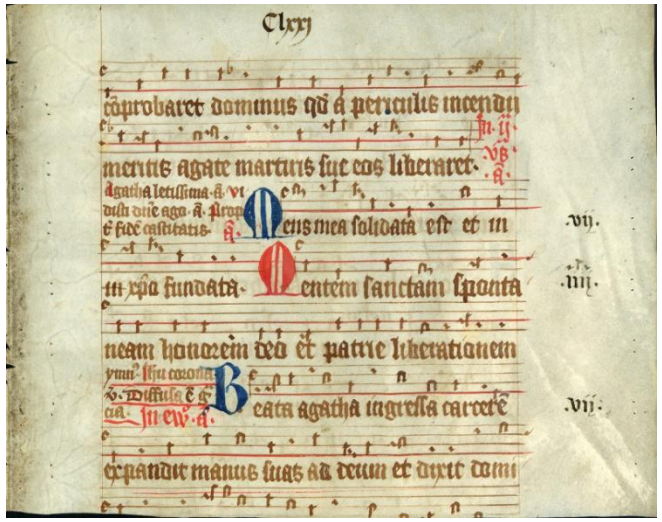
Early Modern Manuscript

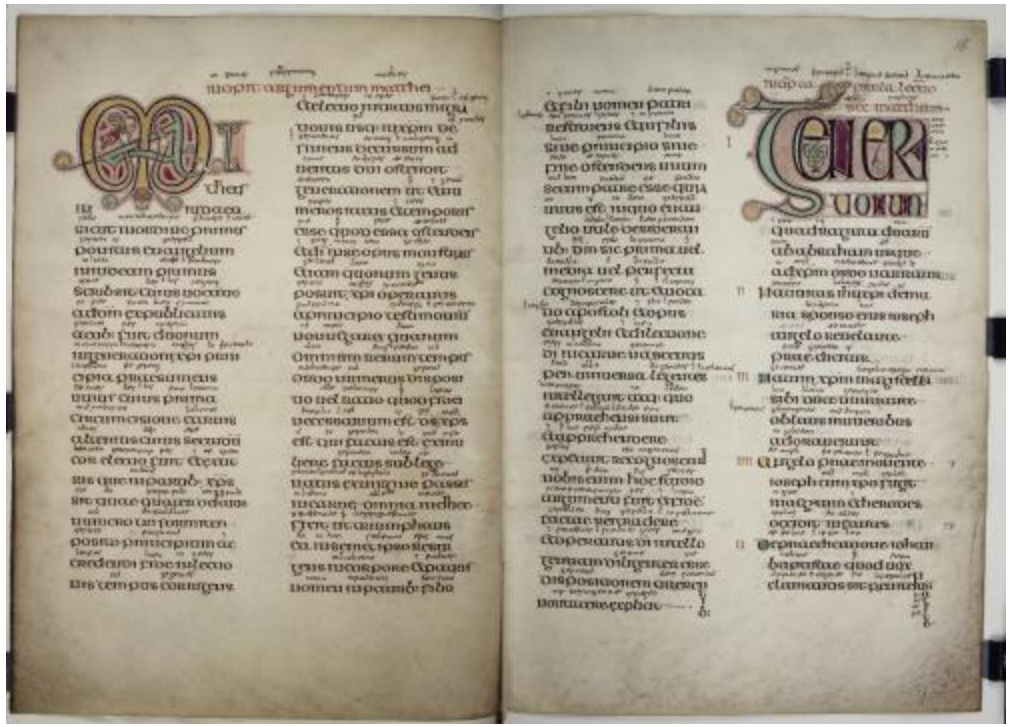




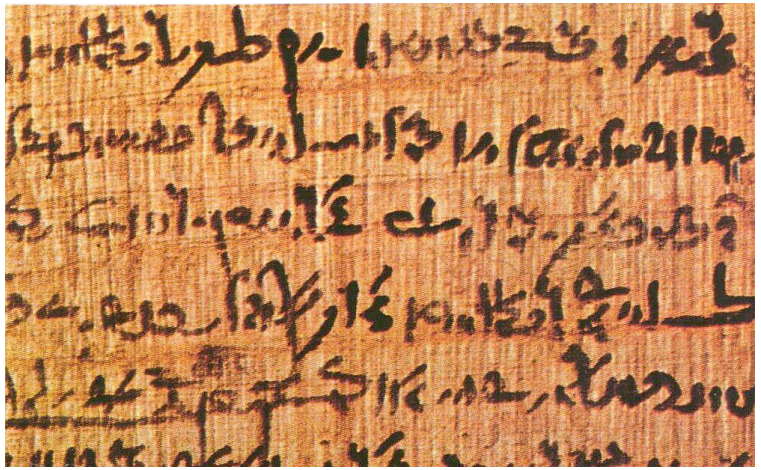


Medieval Manuscripts

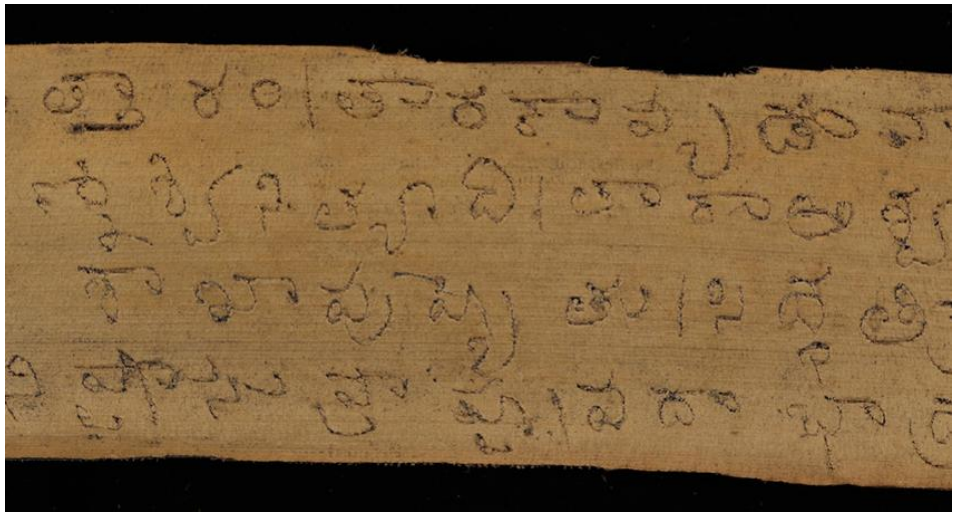




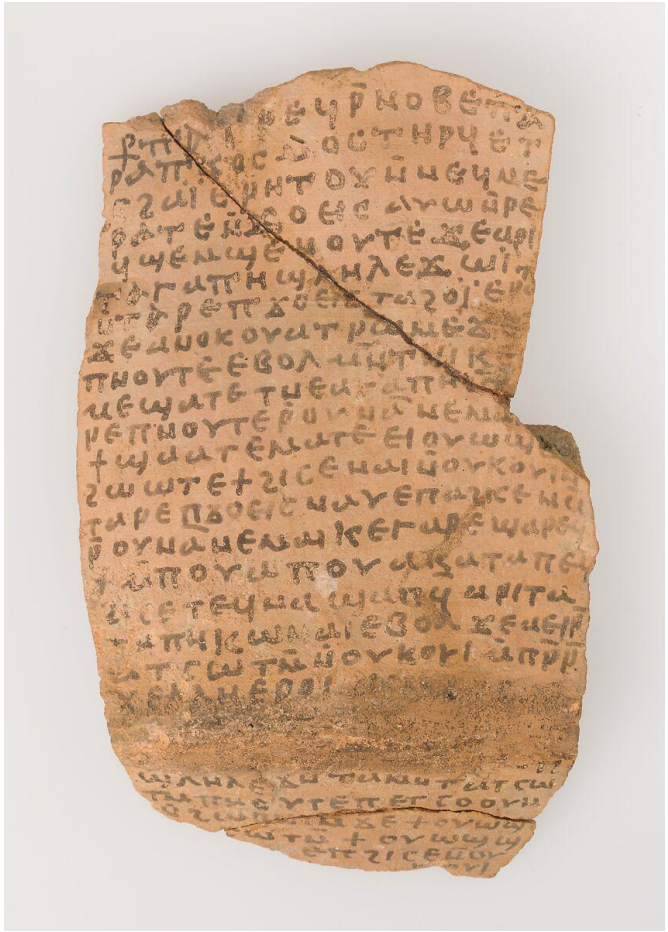
Ancient Papyri



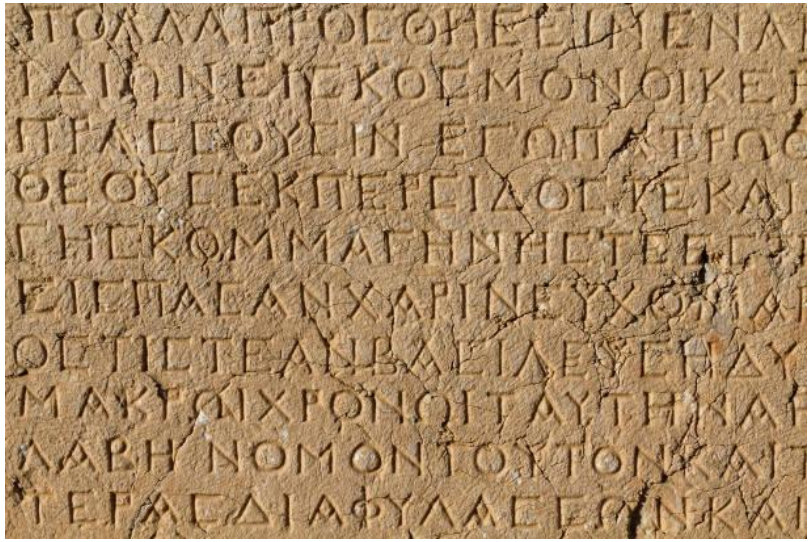
Ancient Leaf Document

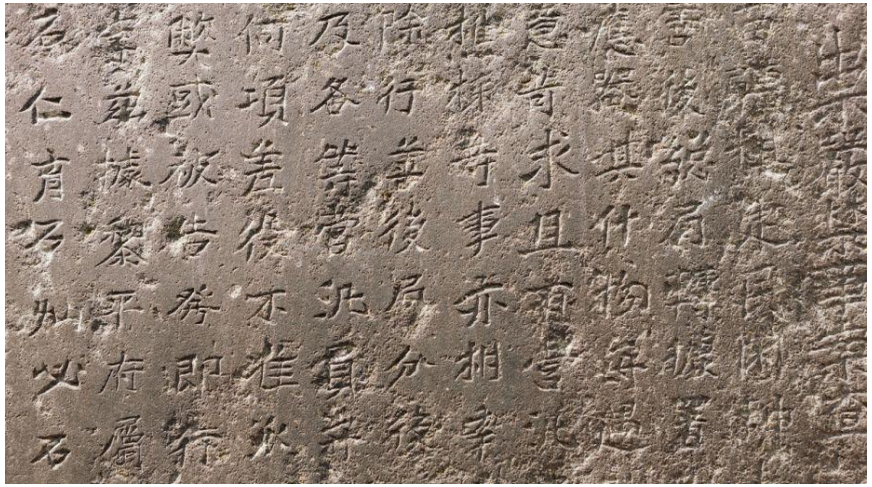


Ostraca



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

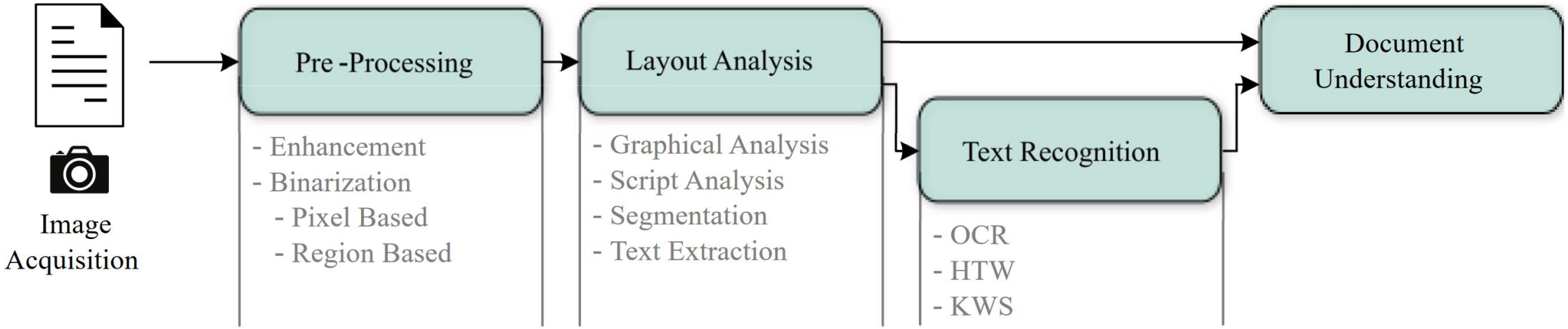
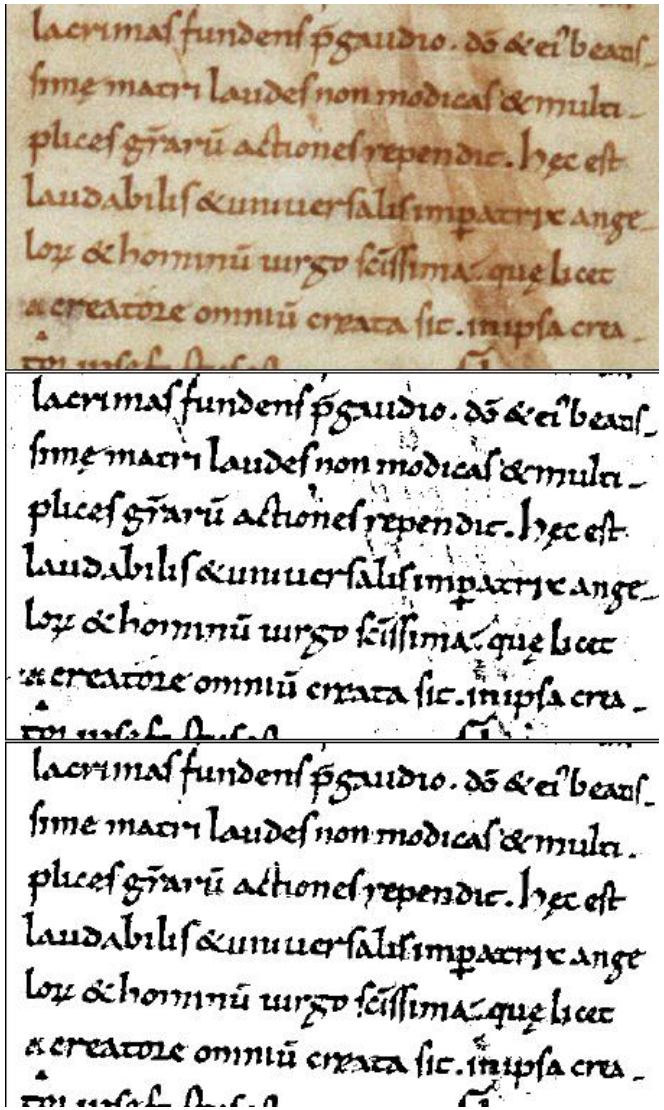
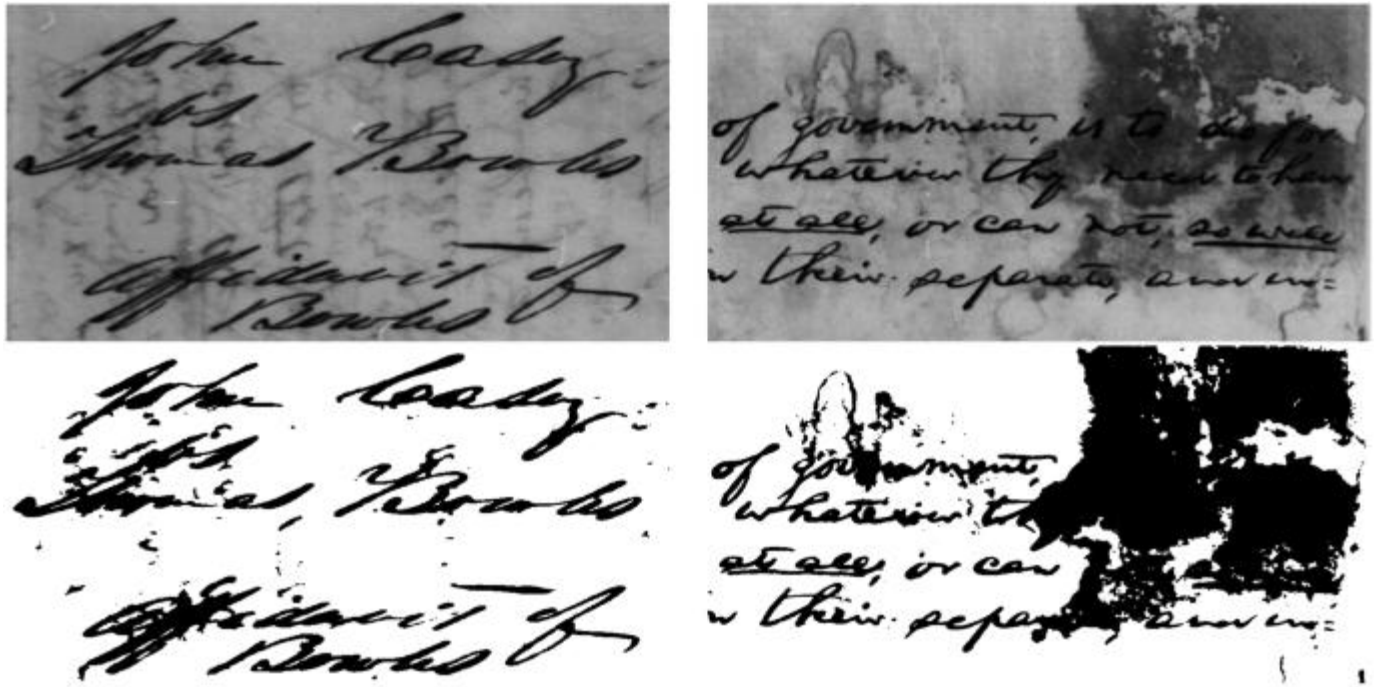


Figure 2.5: Typical Historical Document Processing Workflow.

Pre-Processing - Binarization



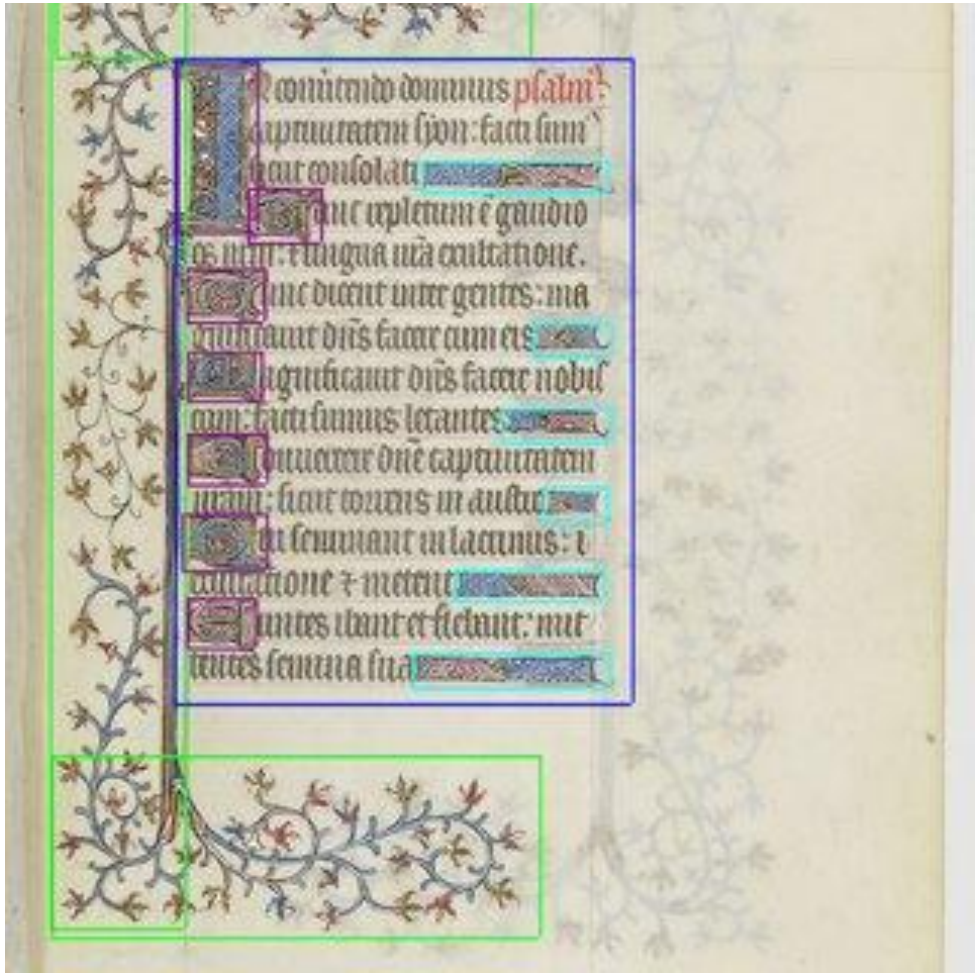


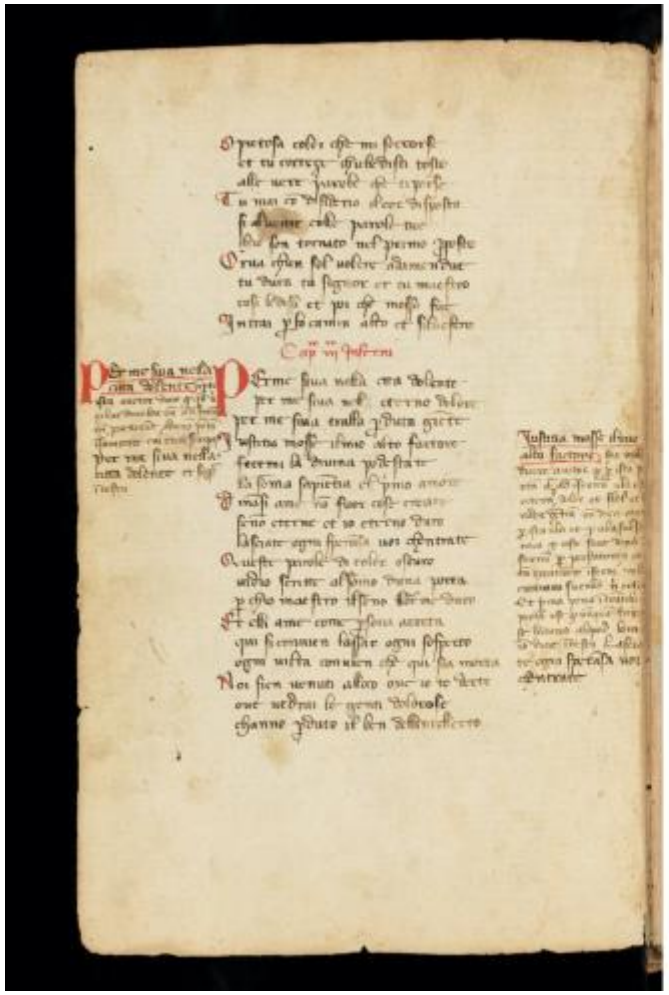
Pre-Processing - Binarization

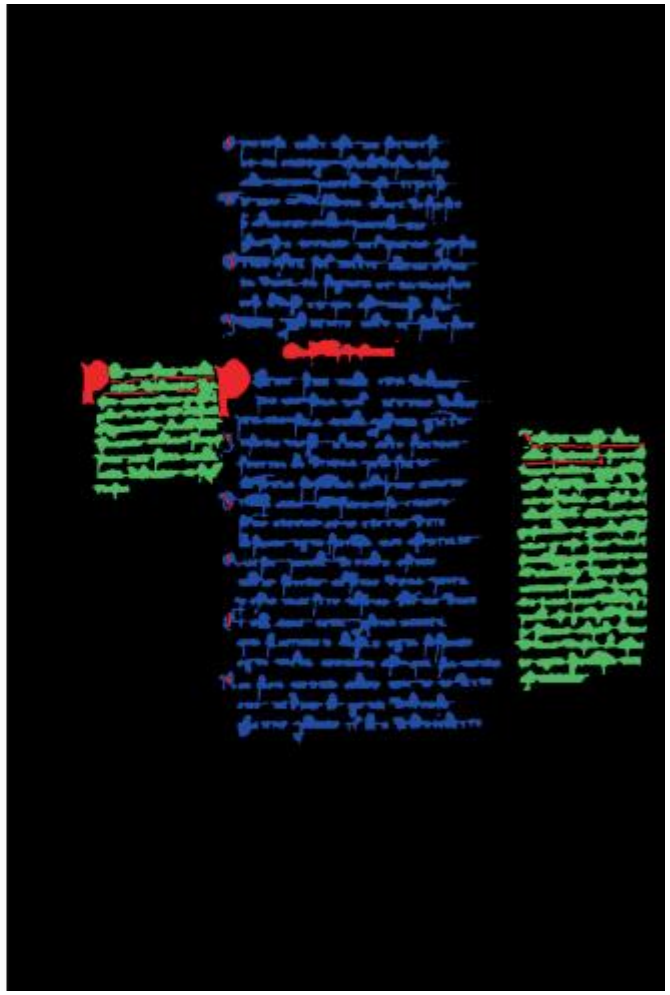
Deep Learning-Based Approaches (End-to-End)

- **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







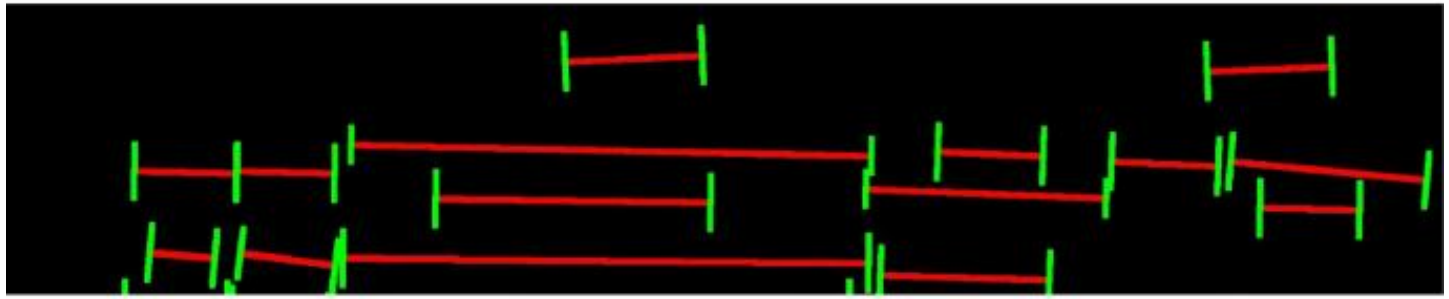
Layout Analysis – Page Segmentation

1. **CNN-based semantic segmentation**
   - **U-Net e 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 sor small datasets.

Layout Analysis – Line Segmentation

Layout Analysis – Baseline Detection

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

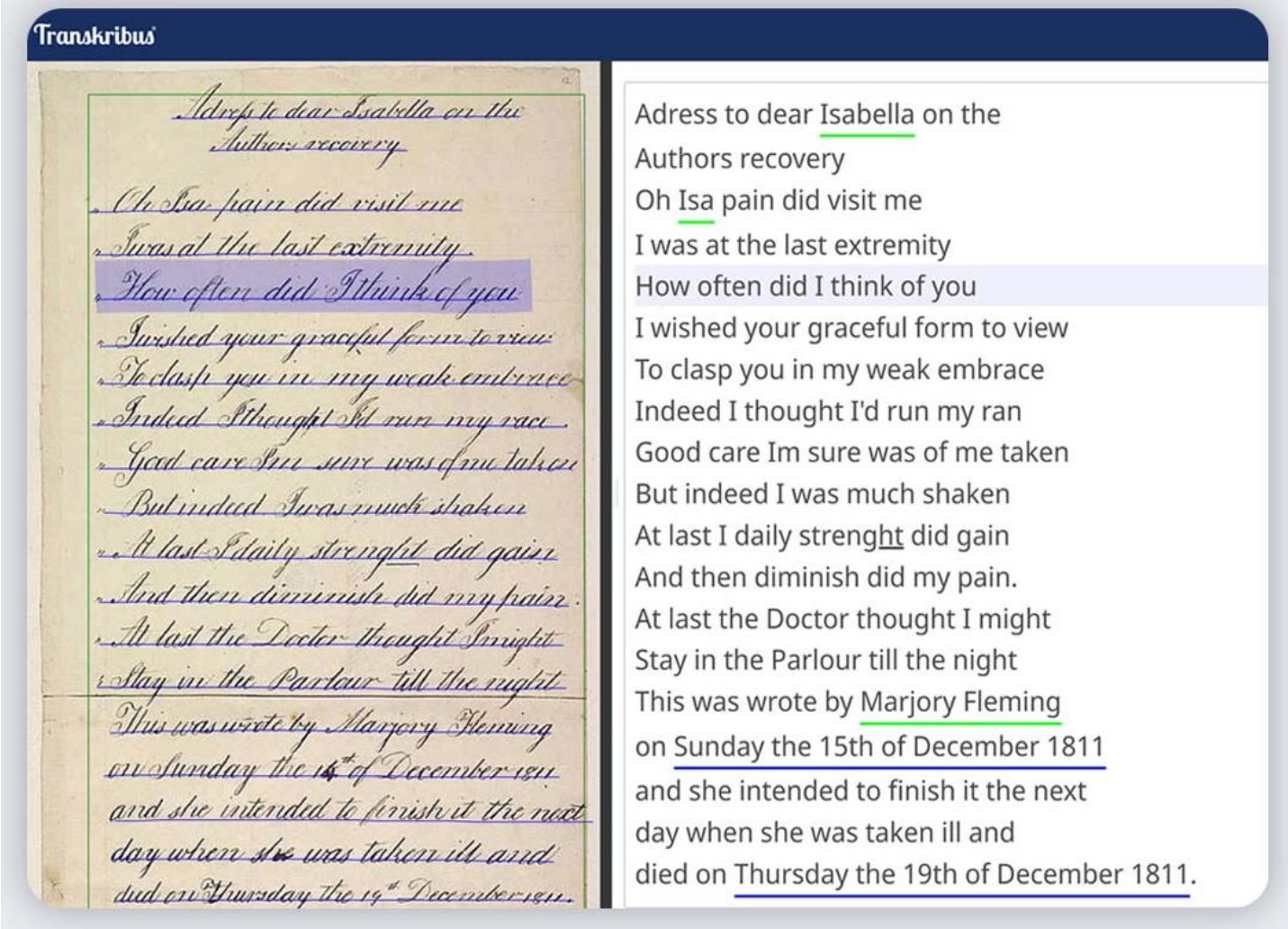


(b) Pixel ground truth – Green encodes the separator class, red the baseline class and black the "other" class. See supplements for an algorithm to automatically generate such GT.

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

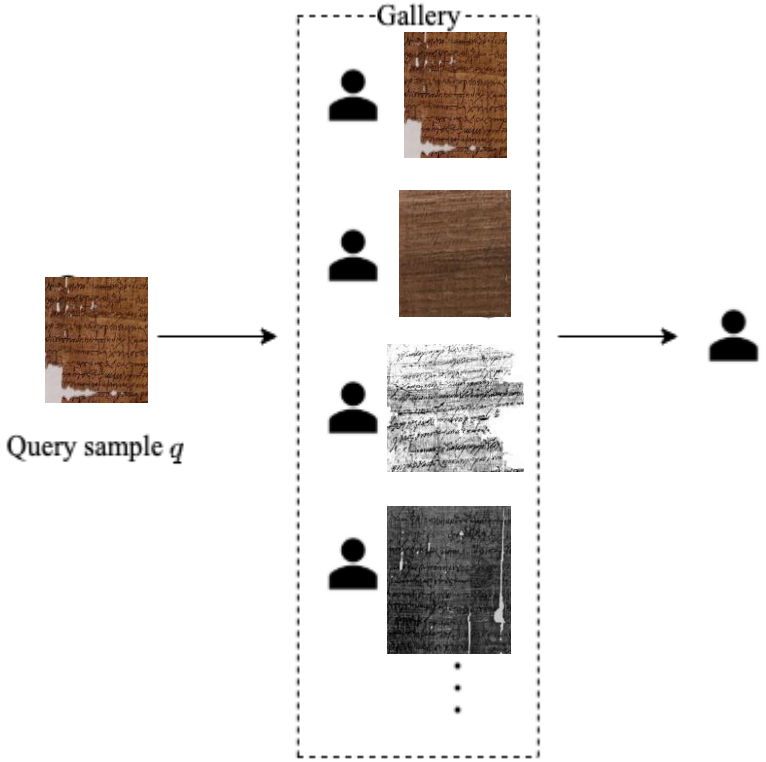


Text Recognition and Transcription

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

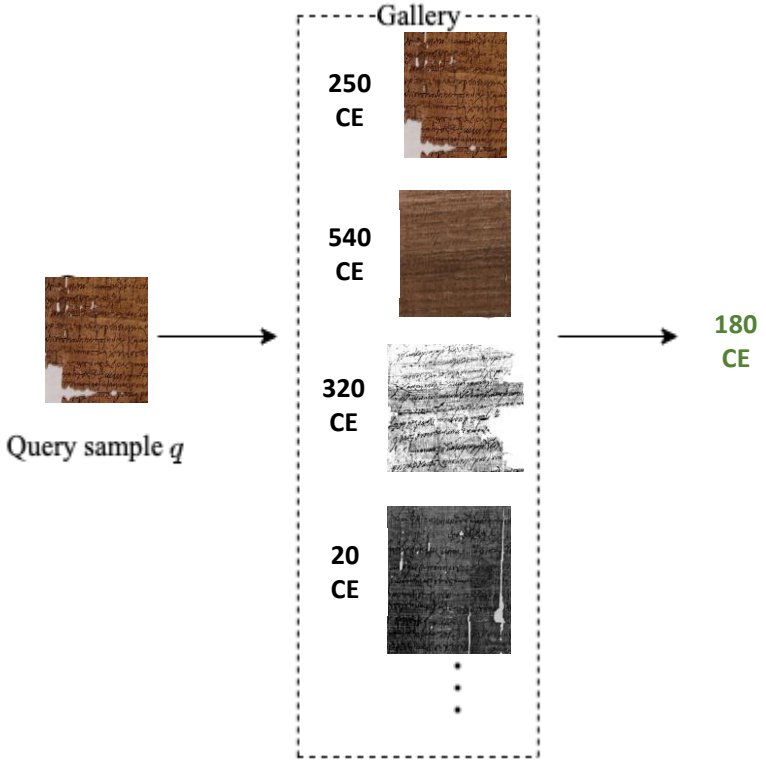
Classification and Attribution – Writer Identification

- Identify the hand of a particular scribe

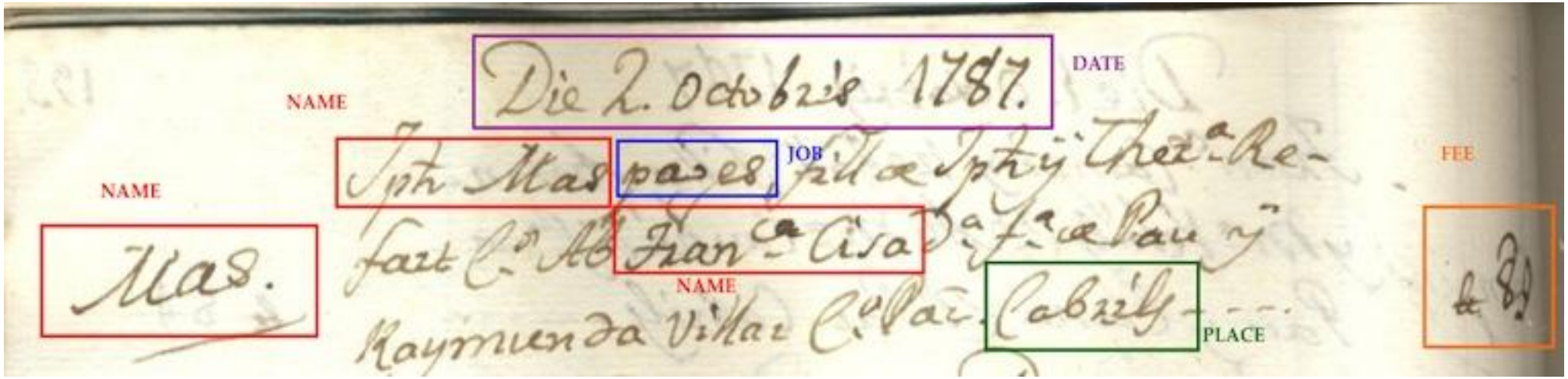


Classification and Attribution – Dating

- Estimate the creation date of a document



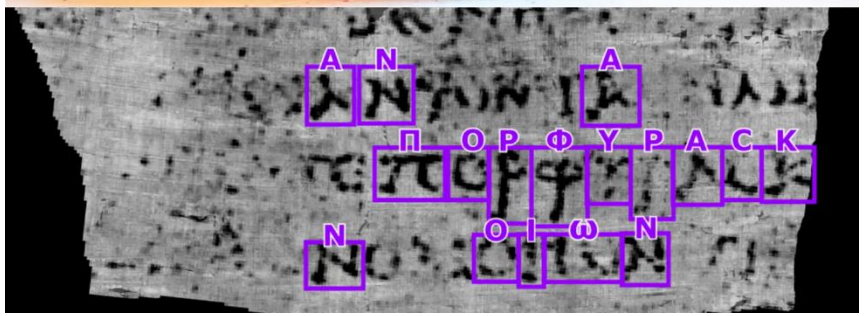
Linguistic and Semantic Analysis

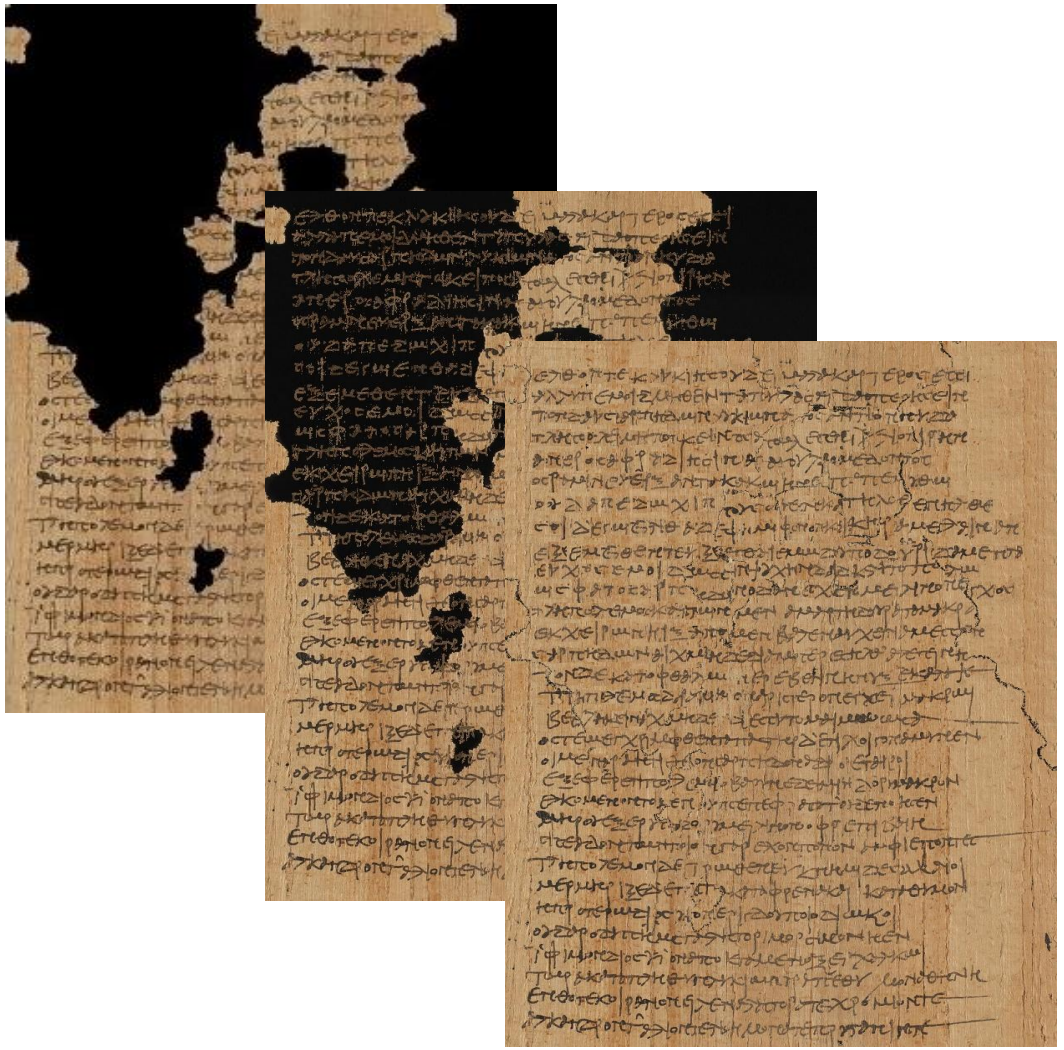


Linguistic and Semantic Analysis

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

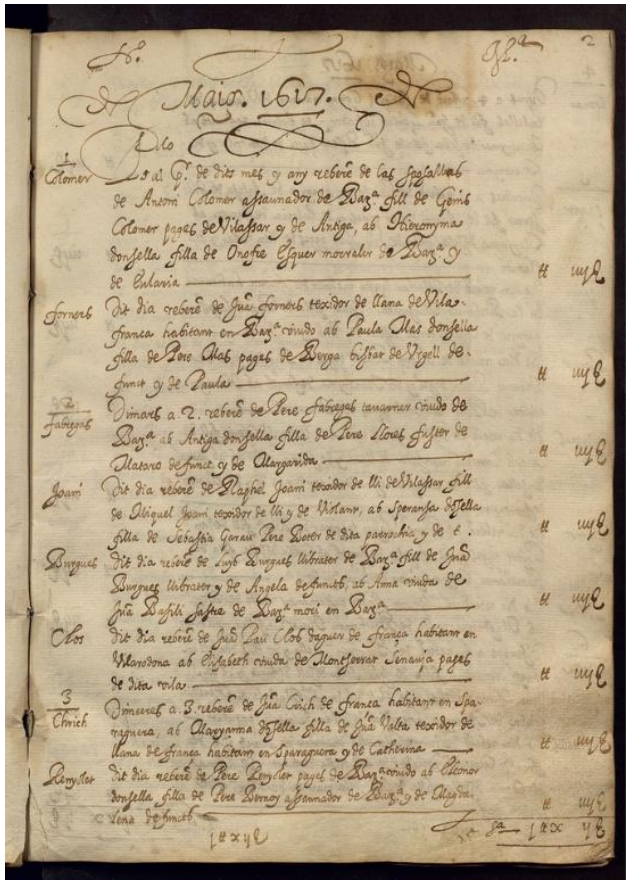
Early Modern to 20th Century - Layout Analysis - Different European languages

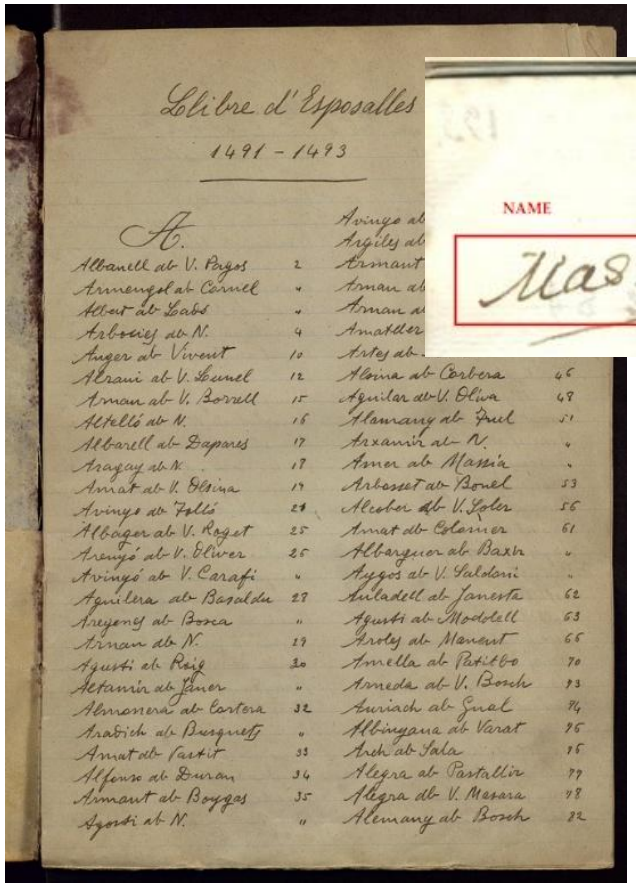
https://www.digitisation.eu/impact-dataset/

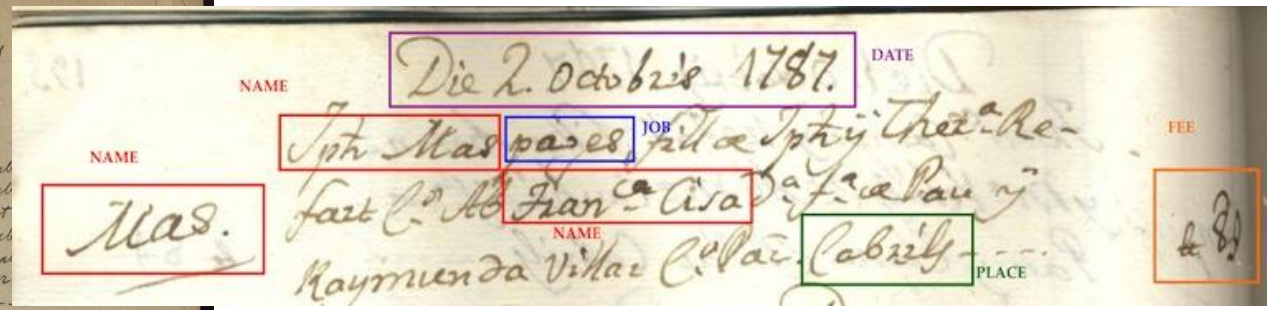
Esposalles

14th to 20th Century - Semantic Analysis - Spanish

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



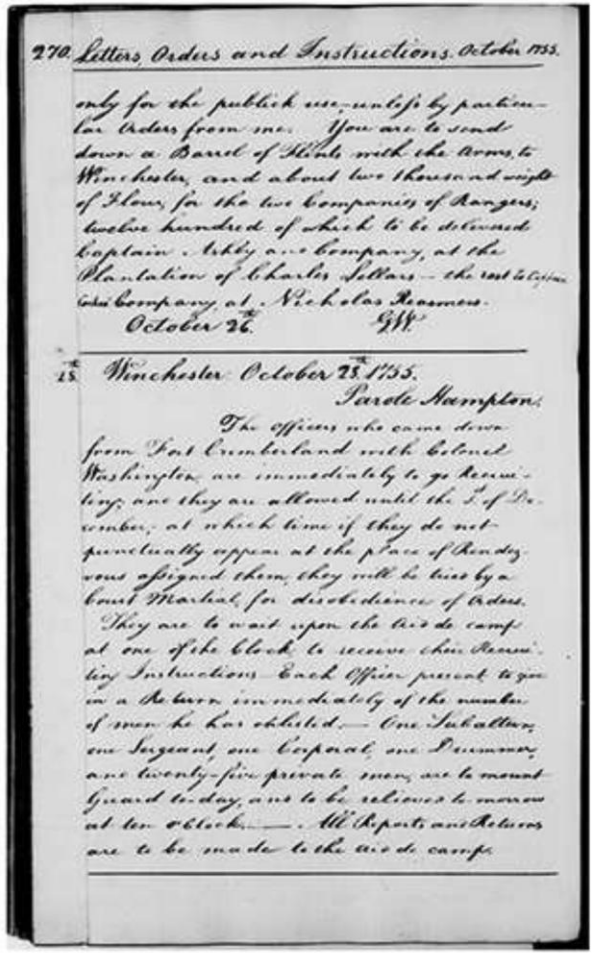




Washington

18th Century - Keyword Spotting / HTR - English

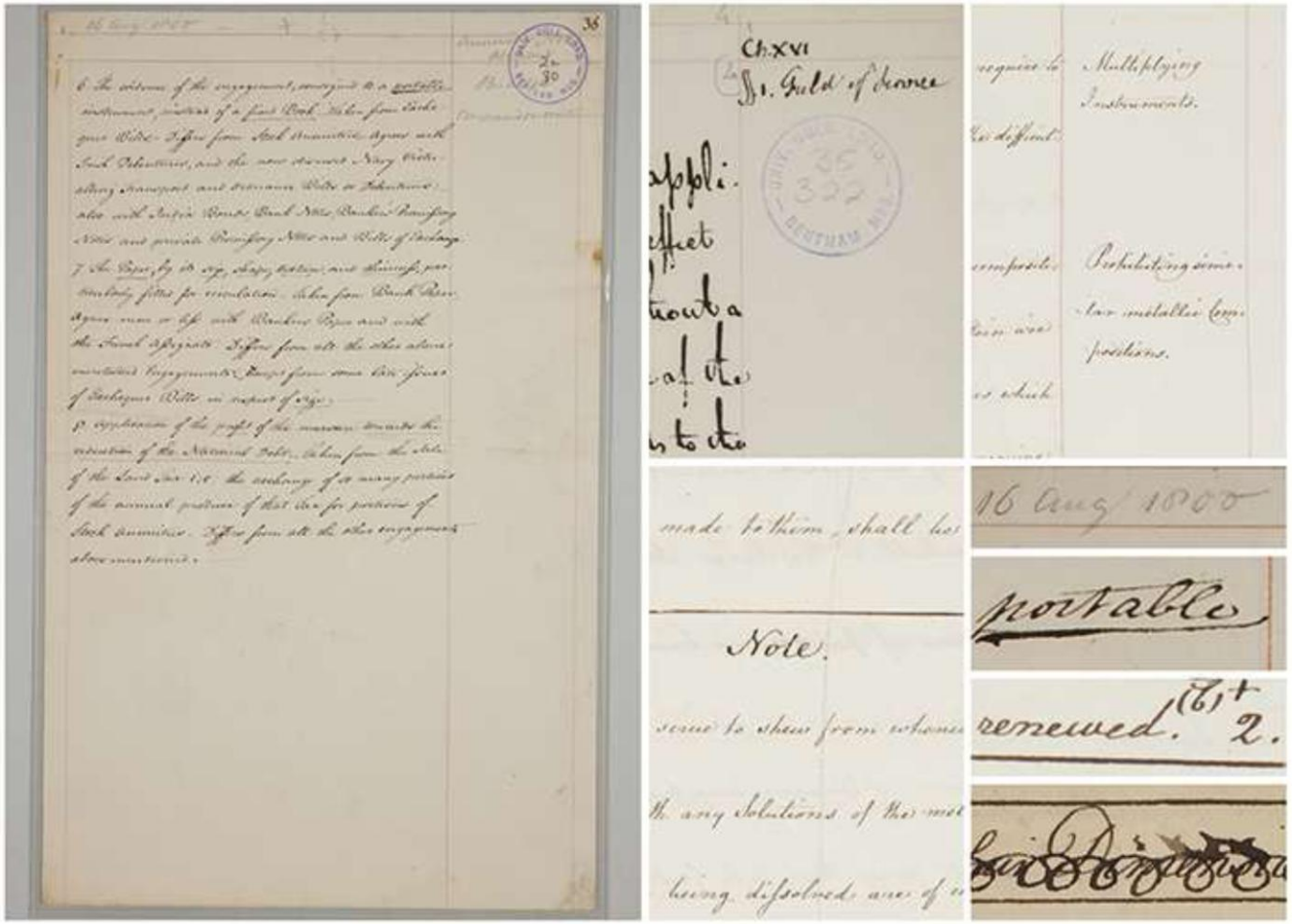
https://fki.tic.heia-fr.ch/databases/washington-database



Bentham Collection

18th Century - HTR - English

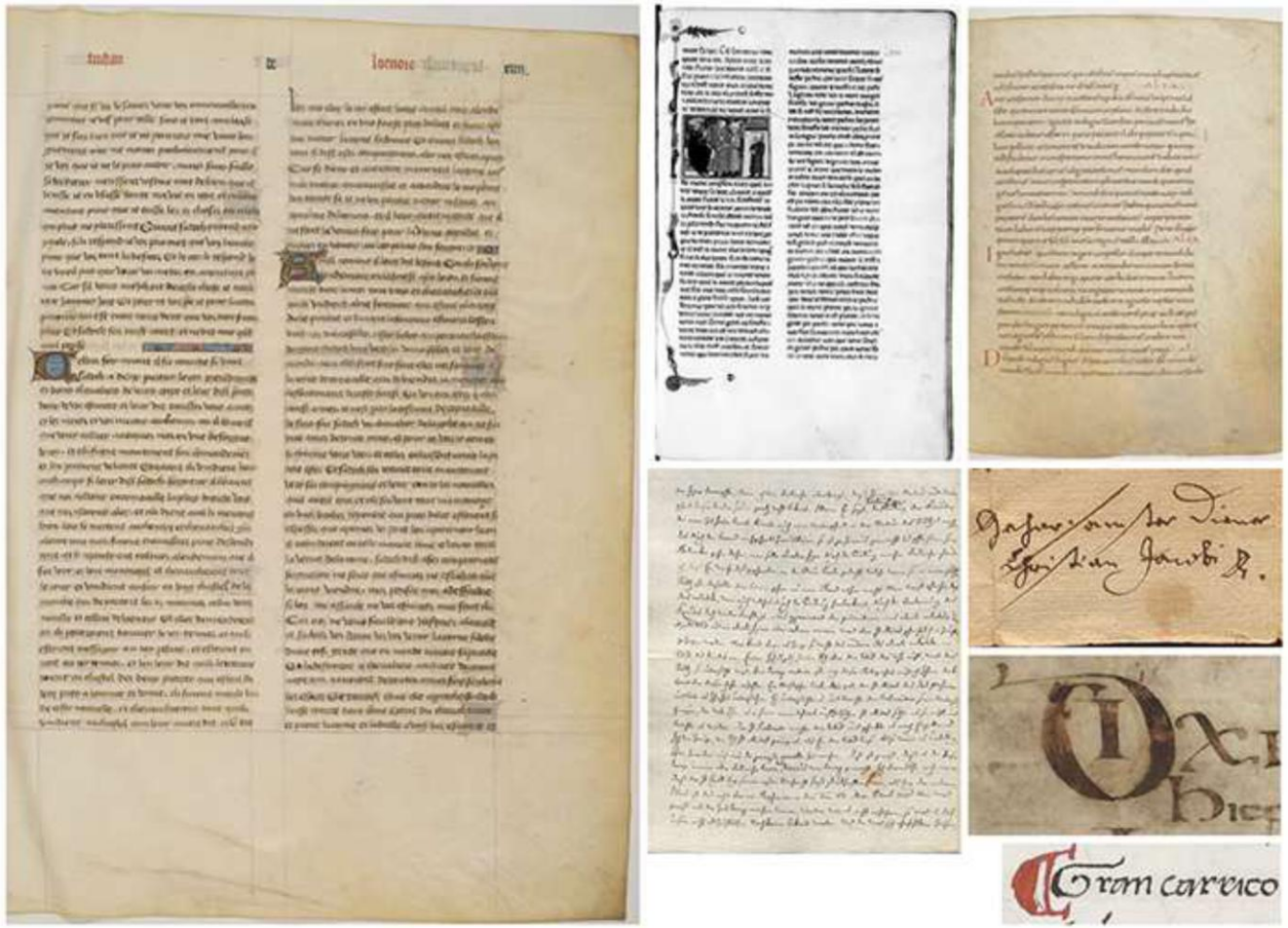
https://zenodo.org/records/44519



HisIR19

Medieval (9th to 15th) - Image Retrieval - Different European languages

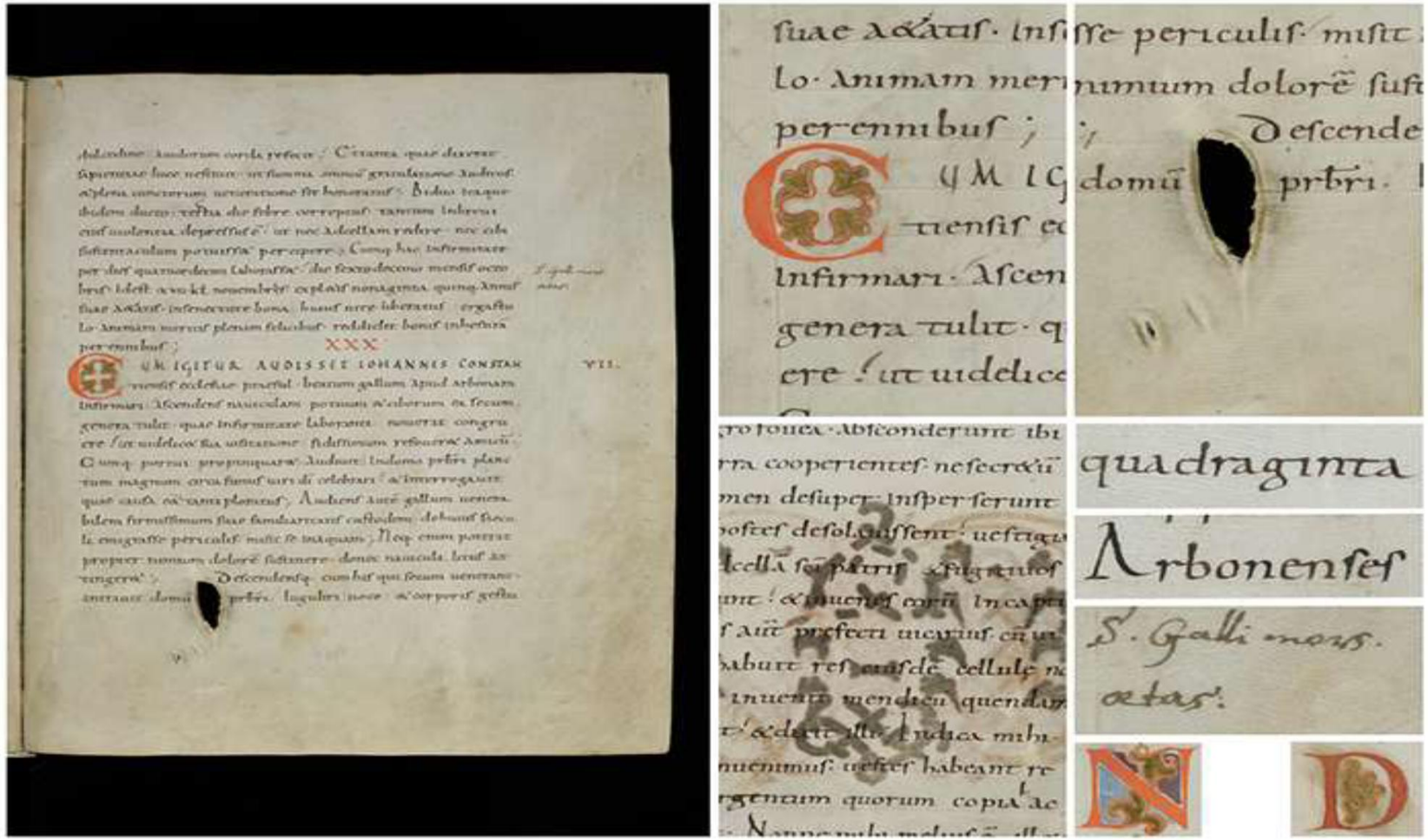
https://tc11.cvc.uab.es/datasets/HisIR19_1



Saint Gall

9th century - HTR / Line segmentation - Latin

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



Parzival

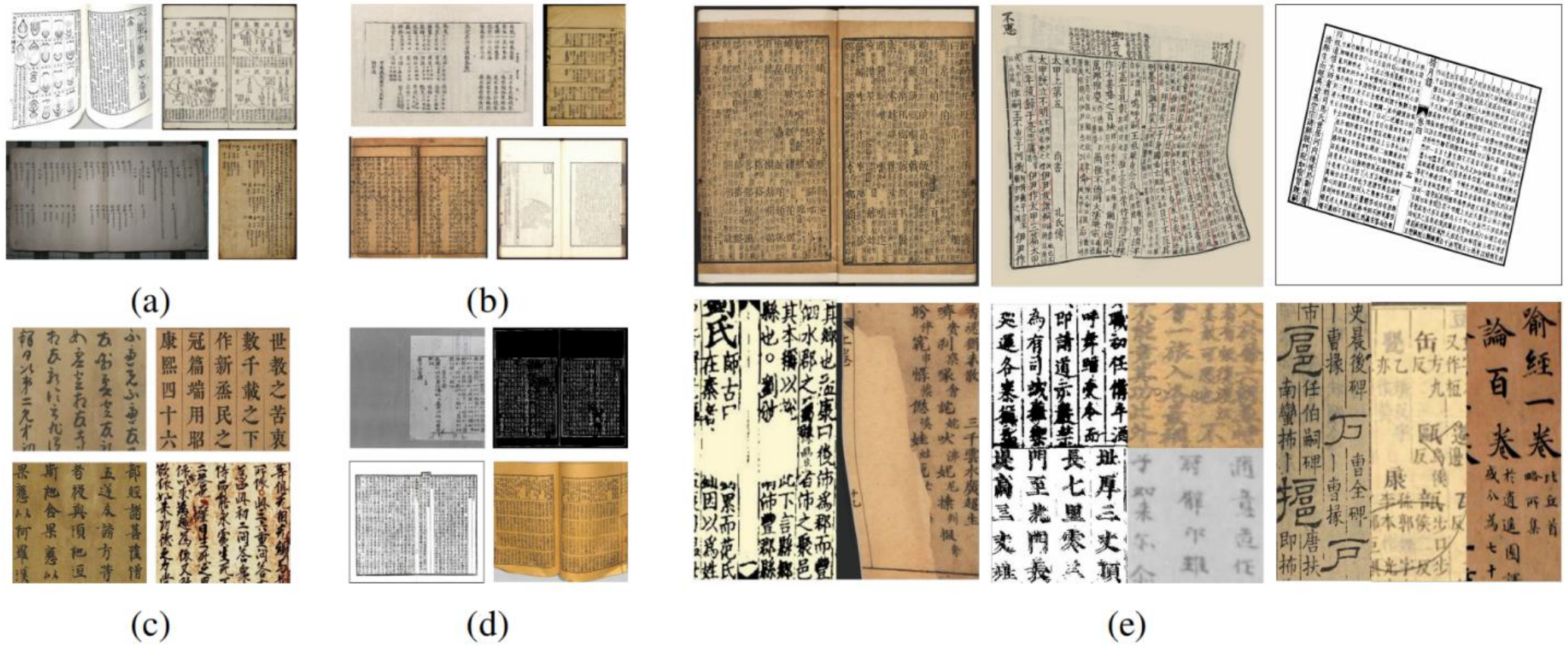
13th to 15th century - HTR - Old German

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

M5HisDoc

13th to 15th century - HTR - Chinese

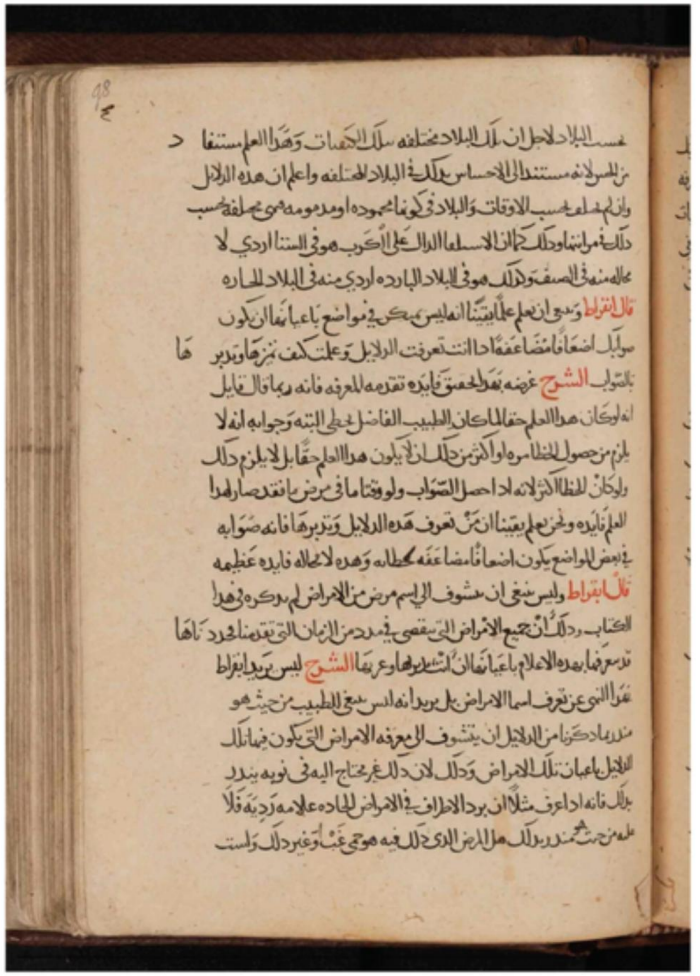
https://github.com/HCIILAB/M5HisDoc



KERTAS

1st to 14th century - HTR / Dating - Arab

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



GRK-Papyri

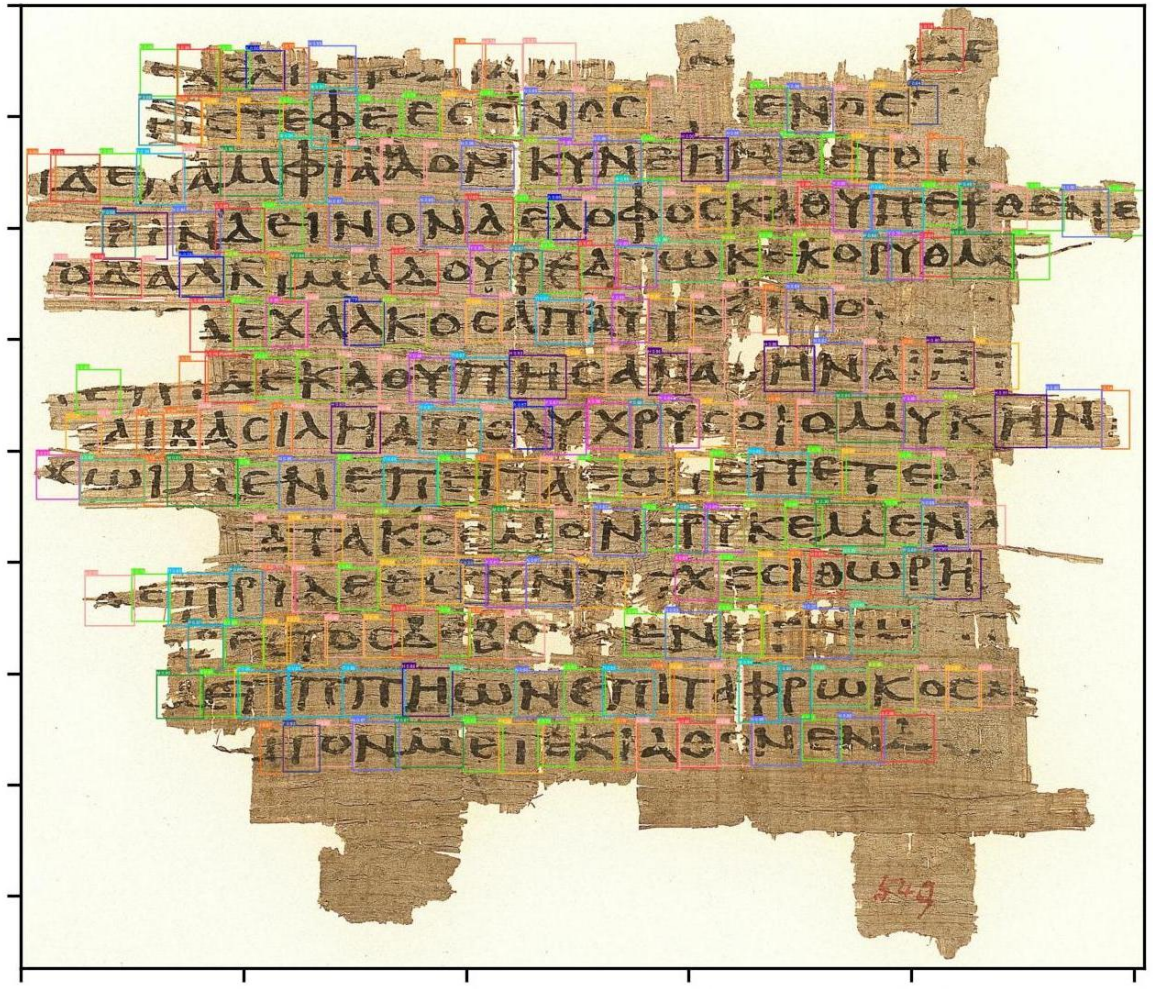
6th century - Writer Identification - Ancient Greek

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

DROGLOP

1st to 4th century - Character Detection - Ancient Greek

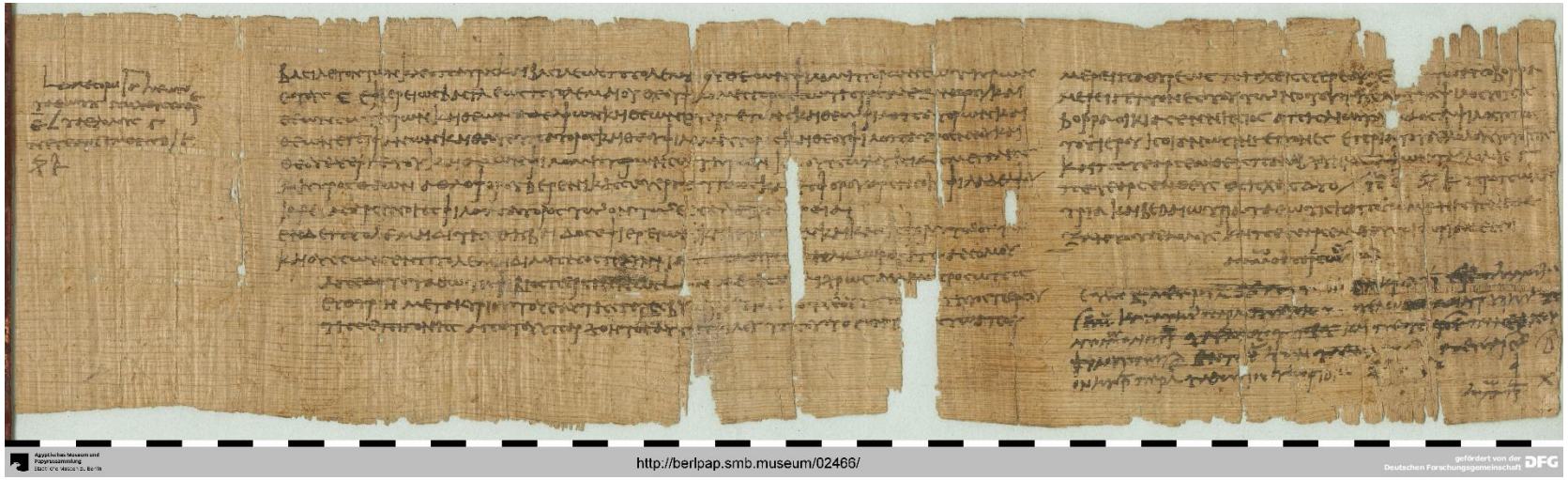
https://qspace.qu.edu.qa/handle/10576/12024

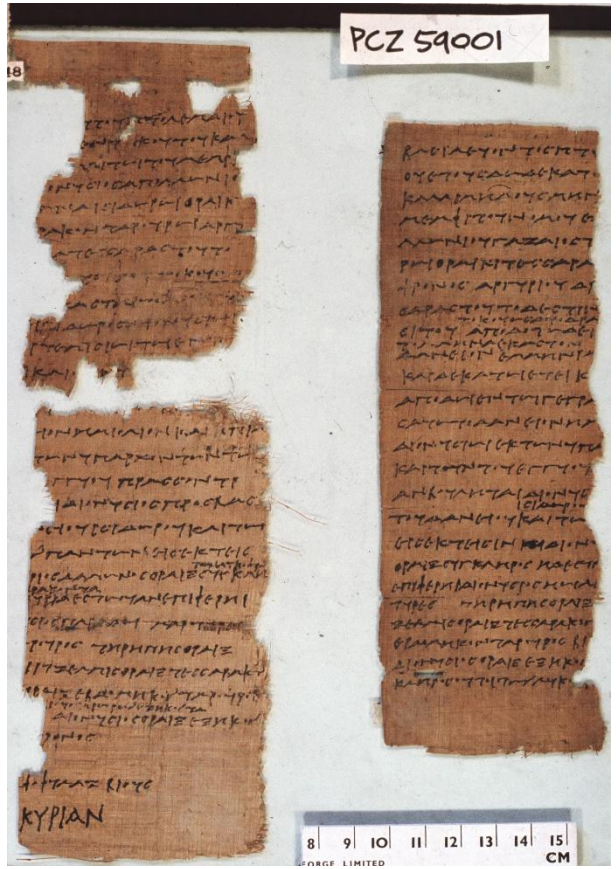


HellDate

3rd to 1st century BCE - Dating - Ancient Greek

https://zenodo.org/records/15830980





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

Contribute content

Partners

APIS
DDbDP
HGV
BP
Trismegistos
APD

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

Ἀπλονοῦς Θερ[μουθ]ᾶτι
τῇ μητρὶ πλεῖσ[τ]α χα[ίρε]ιν·
πρὸ μὲν πάντων εὐχόμεθά σε
3,ms ⸢αρα
ὑγενιν(*) σὺν Ἀπλοναρίῳ. θέλω [σ]ε γι-
νώσκιν(*) ὅτι ἤκουσα παρὰ τῶν {ο} ἡ-
κώτων(*) μοι ὅτι ἠσθένηκος(*),
ἐχάρην δὲ ἀκούσασ[α] ὅτι κωμ-
σῶς(*) ἔσχηκος(*). ἐρωτῶ σε μεγά-
λως καὶ παρακαλῶ, ἐπιμέλου
ἑατῆς(*) ἅμα καὶ τῆς μικρᾶς ὡς
παρέλθ[η]τε τὸν χιμῶνα(*), εἵ-
να(*) εὕρομον(*) ἡμᾶς(*) υἱένωντος(*).
καὶ εἰμῖς(*) γὰρ πάντος(*) ὑγε[ν]ωμον(*).
καὶ περὶ τῆς Συρίας ἕ[ω]ς ἄρτι οὐδ[ὲ]ν
κακόν. ἐρωτῶ σε ἐὰν ἀκούσῃς πε-
ρὶ ⟦Τ⟧ Θερμουθᾶτος πέμψον μοι φάσ[ι]ν.
ἐρωτῶ σε, οὐ πρᾶγμά ἐστιν, ἐάν
τινα εὕρηται(*) καταβαί⟦ ⟧νωντα(*),
ἀποστῖλε(*) ὑμῖν(*) φάσιν περὶ τῆς
ὑγείας(*) ἡμῶν(*) καὶ τῆς μικρᾶς.

Image [open in new window]

Online Archives - Be careful

The Sigma Case

Σ

According to some analysis:

«*the letter sigma (Σ) results in being very rare in the Hellenistic-Roman period…*»

Online Archives - Be careful

The Sigma Case

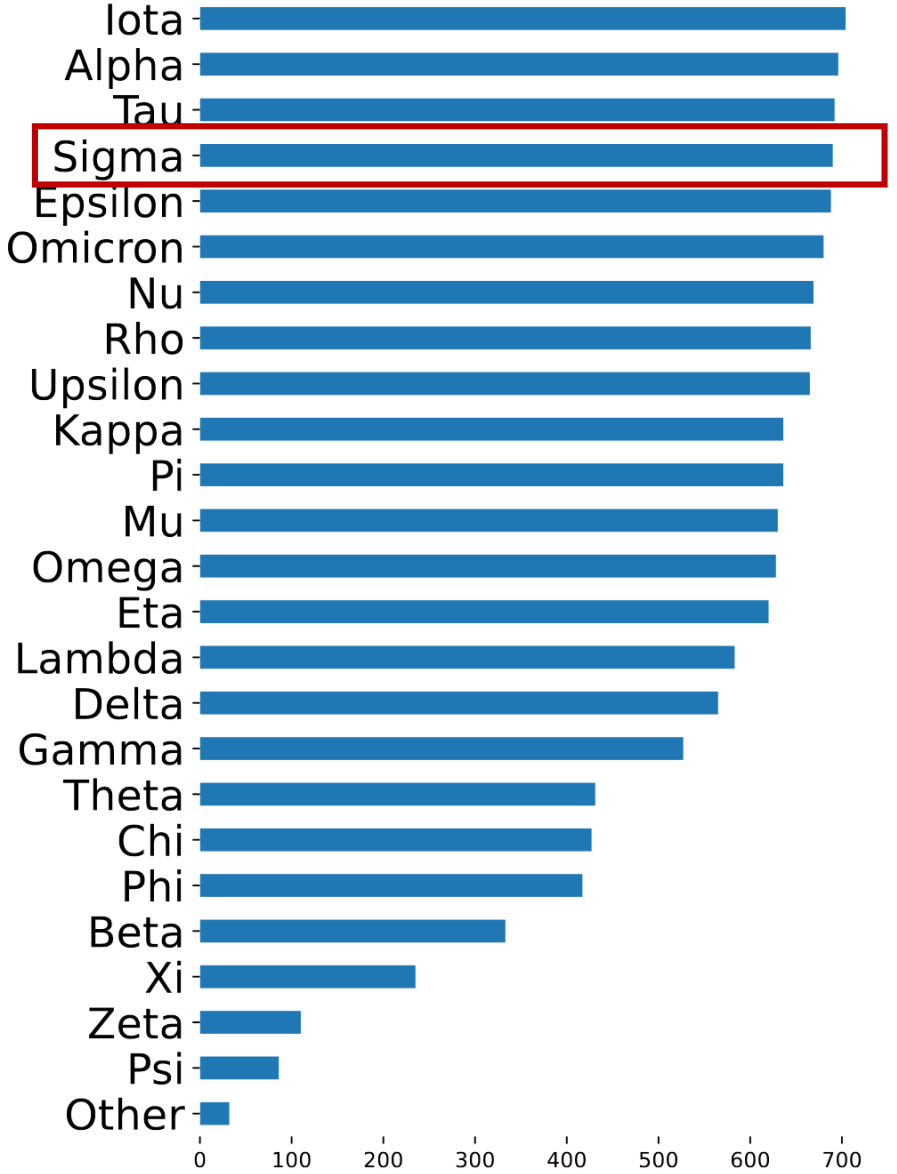


Figure 1. Letter frequency in our Hell-Char dataset

GitHub Tutorial:

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

GliFix Tool:

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