Assembling Legacy Data for AI: the Case of the Paris Bible Project

Estelle Guéville, Louvre Abu Dhabi
David Joseph Wrisley, NYU Abu Dhabi, @DJWrisley

Futurs Fantastiques, BnF
09/12/2021

جامعة نيويورك أبوظبي
NYU | ABU DHABI

اللوفر أبوظبي
LOUVRE ABU DHABI

LAD 2013.051

اللوفر أبوظبي
LOUVRE ABU DHABI

جامعة نيويورك أبوظبي
NYU | ABU DHABI

OUTLINE

1. Our Source Material & Collaboration
    - What is a Paris Bible?
    - Paris Bibles as a Collection

2 Assembling Legacy Data for AI

- Gathering Data and Images
- Challenges of Legacy Data

3 Understanding Digital Collections with AI

- HTR-centered Workflow
- Some High Level Results

4 Collaboration and AI Research for Culture

Conclusion: What is the Promise of HTR for Other Archival Situations?

BnF Latin 10422
115x90 cm

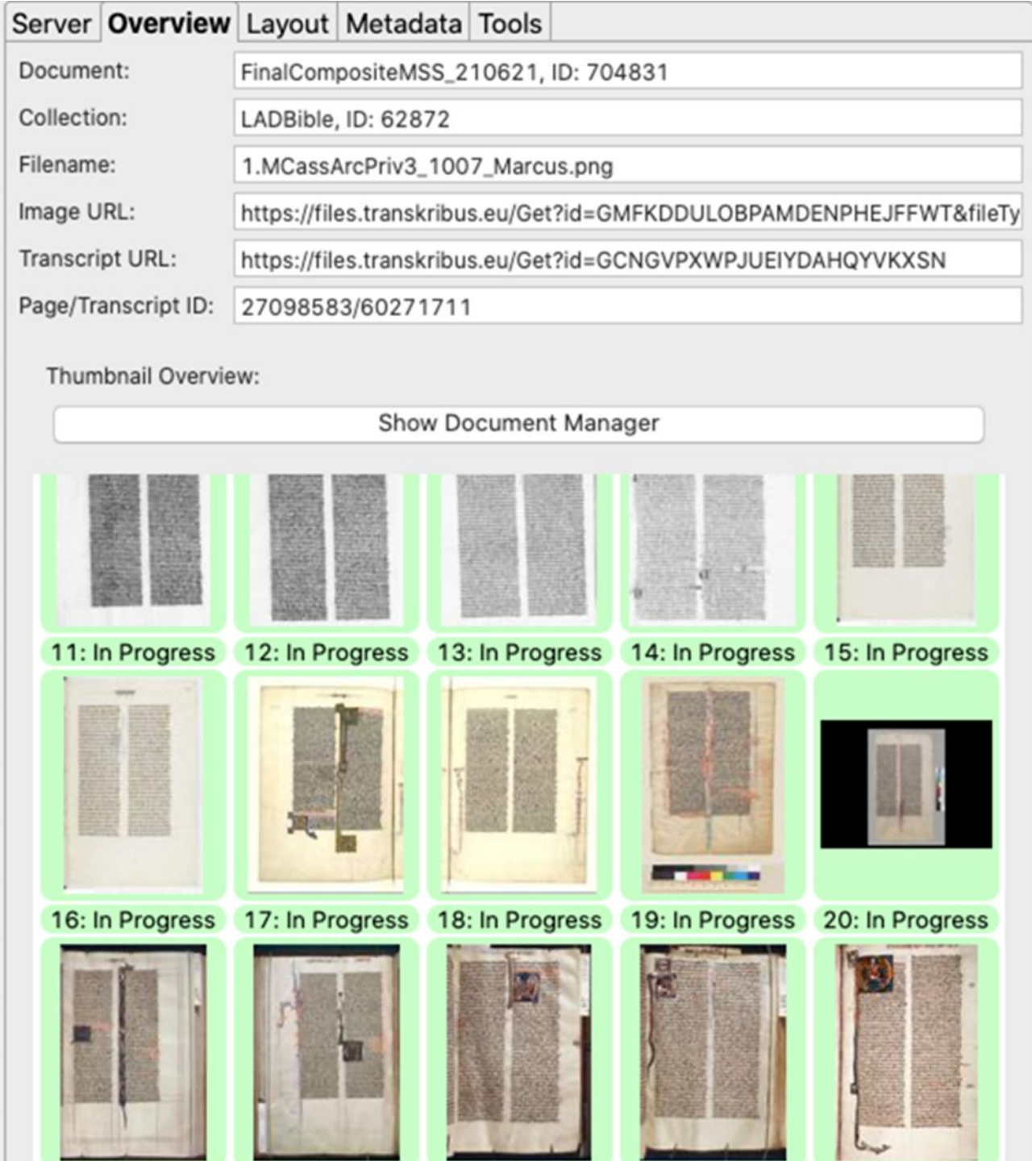
BnF Latin 40
255x180 cm

BnF Latin 11
425x300 cm

Source gallica.bnf.fr / Bibliothèque nationale de France, Département des Manuscrits, Latin 11

A Museum-University Collaboration for Studying Paris Bibles…

- One of the first mass-produced objects.
- Emerged in the 13th century, following the foundation of European Universities, specifically Paris.
- Spread in Europe during two centuries, with thousands n institutions worldwide.
- Books usually not signed, often portrayed as uniform:
    - Shared layout
    - New division of biblical books, following scrupulously the same order.
    - Increased number of abbreviations added to a smaller font size and more lines on the pages with a division in two text columns.
- Script somewhat specific to regions/workshops.
- This diversity in "uniformity" plus the scale of the archive makes it a perfect test case for machine learning
- BUT… we are not working with existing structured data, but creating unstructured textual data from HTR



Paris Bibles as Collection

- **Evolution in book history. Paris Bibles…**
  - were objects for individuals, for studying, teaching or preaching purposes.
  - have not traditionally been a particular medieval collection, or created to be part of a collection. To our knowledge, there have not been collections of fully digitized ones.
  - became a "must-have": today most modern cultural institutions possess at least one, and some historical libraries, many.
  - are not digitised and found in one place..
- **Digital Collections**
  - We are creating a **curated digital collection.**
  - We are **not** building a digital collection in our institution-s, but collecting from across institutions.
  - Our curation principles have machine learning in mind.

Getty

BRITISH LIBRARY

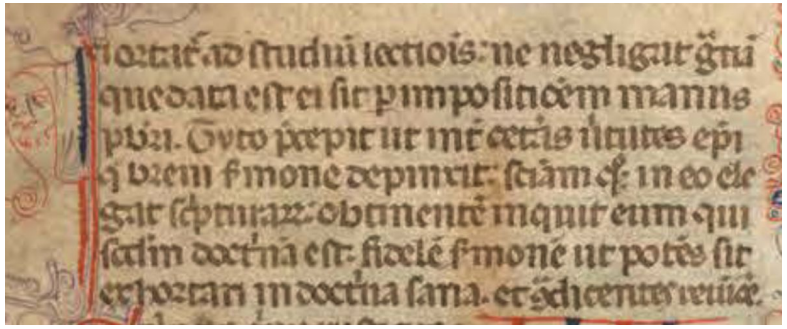
Vatican Apostolic Library

INTERNET ARCHIVE 25

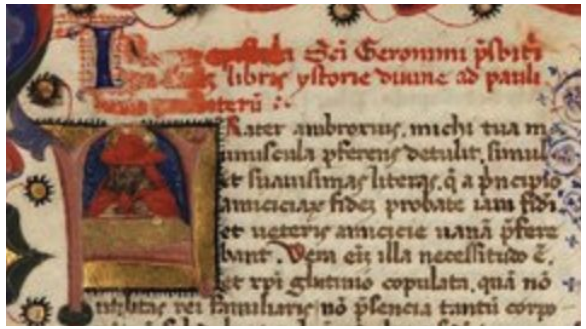
BnF Gallica

Gathering Data and Images

- Understanding where the Paris Bibles are
- Identifying metadata and denominations:
    - Biblia sacra
    - Pariser Normbibel
    - Bibbia dell' università
    - Universitetsbibel
    - Or even general denomination "Bible"
- Choosing sections of Bibles, if fully digitized.
- Isolating text from fragments or partial digitization.
- Challenge: document management and conflicting metadata



Biblioteca digital hispanica: Madrid, BNE 12906, 2r



Biblioteca nazionale centrale di Roma, Ms.Vitt.Em.825, 1r

e-codices: Aargau Cantonal Library, MsWettF 11, 1r

Challenges of Legacy Data

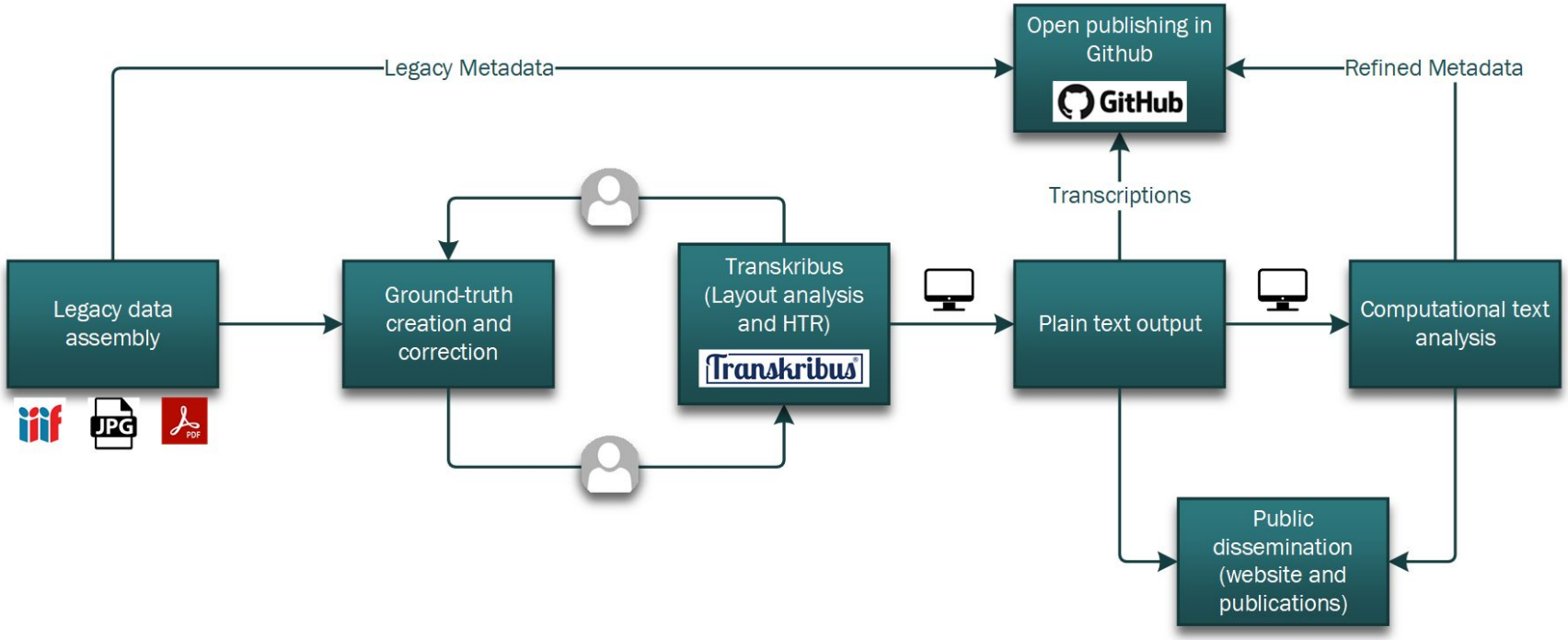
- creating ground truth is a time consuming process (there is little or none with all the features of interest to us)
- with high quality ground truth created based on one or two manuscripts, applying the models to new scenarios led to overfitting
- Ideal of creating a generic model that works across all possible manuscripts is a complex endeavor.
    - is more data the answer?
    - are sub-models the answer?
- On top of these question of the creation of ground truth, many issues are encountered with medieval manuscripts owing to layered histories of digitization:
    - Quality unequal from copy to copy (not so much a machine problem as a human one)
    - Metadata: Can we trust it? Need to start with well known documented mss
    - what is difficult for a machine might not be the same for humans
    - Making the objects comparable, finding a middle ground

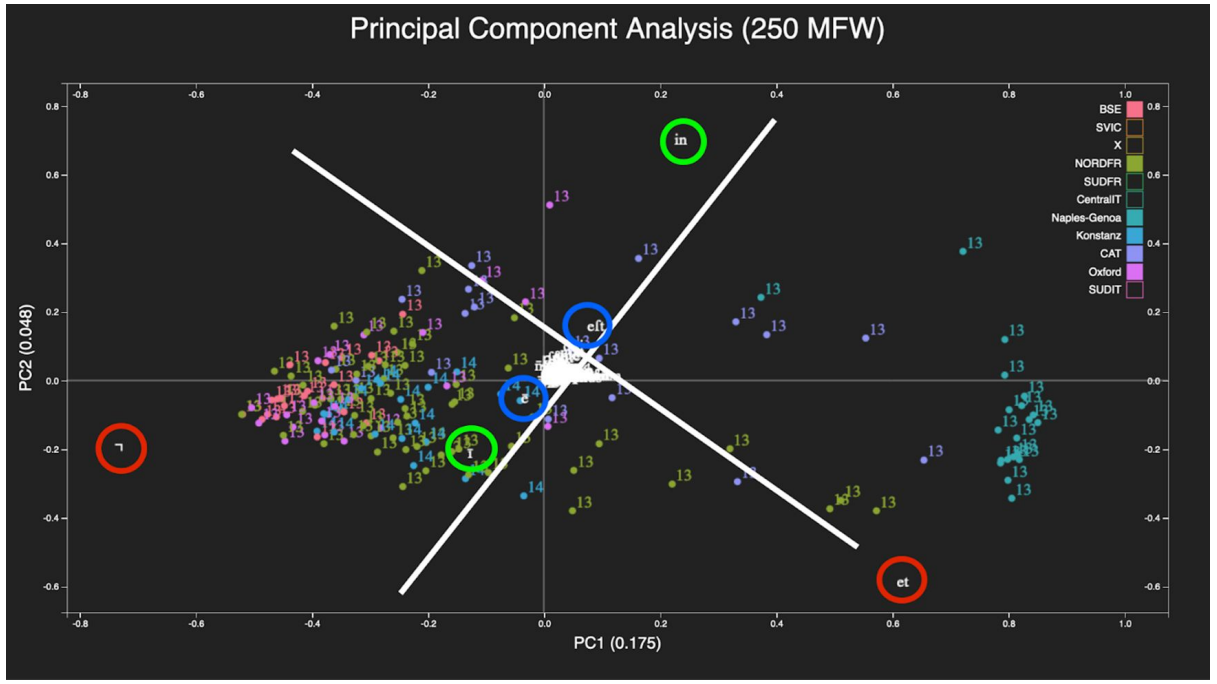
HTR-centered Workflow

- HTR - we have such a controlled corpus (uniformity of layout, of base text, narrow possibilities for paleographic) that we can aim for very low consistent CER.

- Other medieval and non medieval textual examples → higher paleographic difference will mean need for more HTR training and more keyword spotting approach.

- Such work is of a scale that is prohibitive for researchers -- need HTR

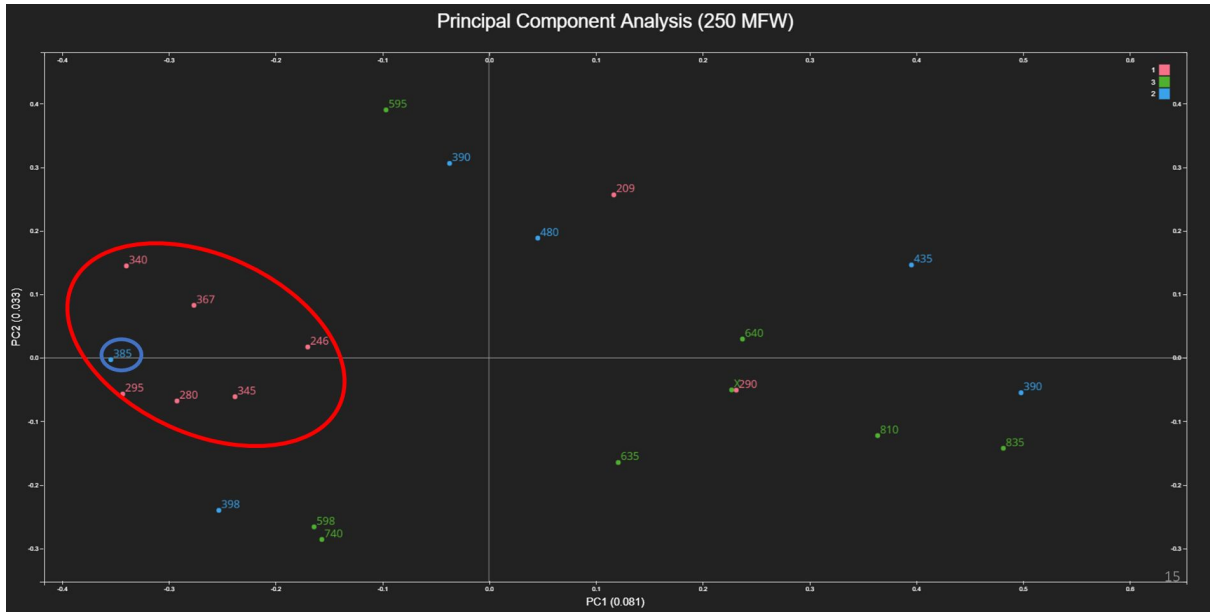


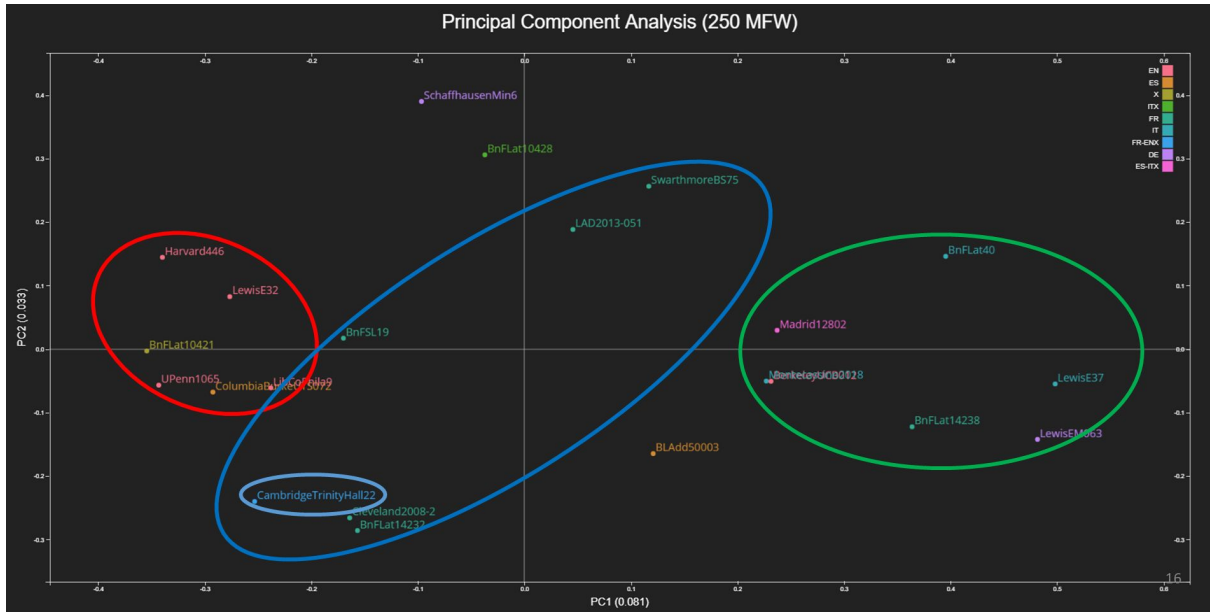


Some High Level Results

- Expectations: The use of abbreviations and special letter forms is specific to every scribe.

- Exciting findings :
  - Size of manuscripts seems to matter
  - regional patterns of abbreviation exist
  - Abbreviation patterns are highly localizable -- to scribe or a workshop





Collaboration and AI Research for Culture

- Timelines for research between university and museum are not the same.
- Funding in different institutions is not balanced.
- Project management for remote work was particularly important.
- Happy consequence of a tricultural project team, for understanding better the ways in which such collections are organized and described.
- *Interlocking question of diversity*: diversity of cultural knowledge about collections and how that diversity can benefit the models we train.
- But having members from those different environments did not assure us that we found everything, including GT from different manuscripts did not assure better output. .

Conclusion: What is the Promise of HTR for Other Archival Situations?

- HTR has great promise for medieval archives, or for any other voluminous handwritten collections
- However, remember that our scenario works relatively well since we are working with a corpus in which the domain is stable and the form is somewhat uniform
- There are plenty of medieval textual traditions with large numbers of manuscripts (*De consolatio philosophiae, Legenda aurea, Roman de la Rose*, etc.)
- BUT these have a very large spatio-temporal imprint, are found in many hands, many layouts, etc.
- You need to have a clear question in mind so that you design your research instrument clearly to know how AI/HTR will give you results which are useful to you, at the level of desired granularity..

Thank you for your attention!

Merci de votre attention!

شكراً لإتبهكم!

اللوفر أبوظبي
LOUVRE ABU DHABI

جامعة نيويورك أبوظبي
NYU | ABU DHABI