# Gallic(orpor)a : Processing Gallica's historical sources

Simon Gabay, Ariane Pinche, Kelly Christensen

# GALLIC(ORPOR)A: PROCESSING *GALLICA*'S HISTORICAL SOURCES

**Simon Gabay,[1] Ariane Pinche,[2] Kelly Christensen,[3]**
**[1]Université de Genève (CH), [2]CNRS CIHAM, Lyon (FR), and [3]Sciences-Po Paris (FR).**

## INTRODUCTION

The apparition of digital libraries have brought together unprecedented amounts of data and provided human sciences the means with which to integrate a new scientific paradigm. Such a transition implies overcoming a technological hurdle, particularly for historical documents which pose a significant number of challenges: manuscript documentation, difficult-to-read hands, variation-rich varieties of language... The *Gallic(orpor)a* project tries to overcome these obstacles.

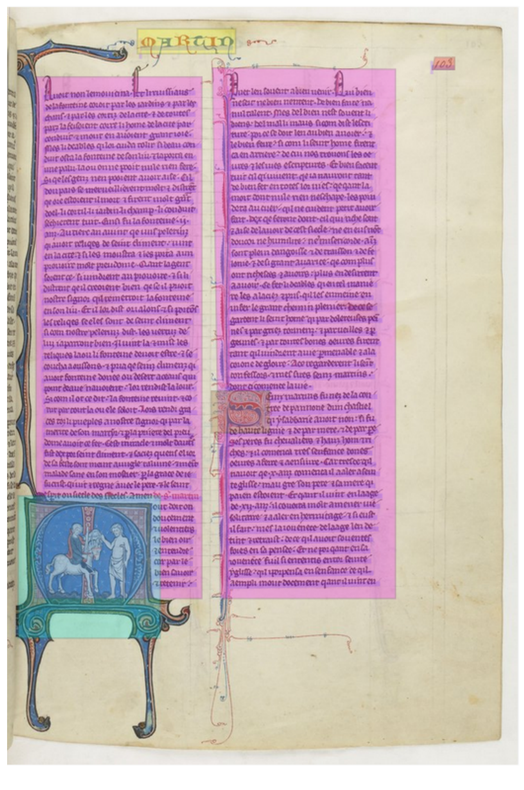
## TRAINING MODELS

One of the main objective of the project is the creation of AI models to extract information from digital facsimiles. Because such a task is not trivial, empirical experiments are conducted to found the best configurations (network architectures, training data, scope of models...)

| **Model** | **Accuracy** |
|---|---|
| Baseline | 98.61% |
| Gallic(orpor)a+ | 98.66% |
| Gallic(orpor)a Antiqua | 91.10% |
| Gallic(orpor)a Antiqua/gothic | 96.74% |

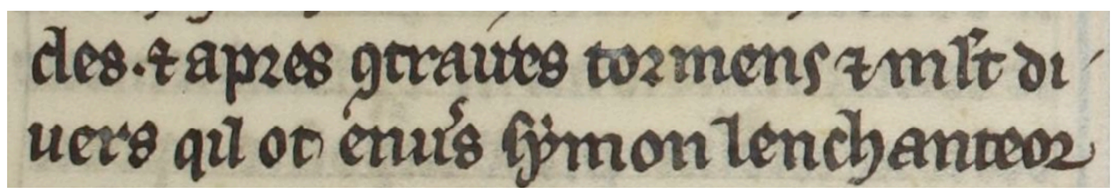
Because AI requires tremendous of data, the harmonisation of the production is a key problem.

## HARMONISATION OF LAYOUT ANALYSIS



The SegmOnto controlled vocabulary is based on the assumption that most of the textual sources can be described in the same way if we use a codicological perspective focused on material aspects (running title, pagination/foliation, headings/rubrics, drop capitals, a main textual zone and potentially notes). This succeeds whether the sources are historical prints or manuscripts.

## HARMONISATION OF TRANSCRIPTIONS



How to transcribe these two lines? (*[...] ꝛ apꝛes ꝯtraires toꝛmens ꝛ ml̓t di / uers qil ot enuś ſymon lenchanteoꝛ*). We offer recommendations regarding:

- Abbreviations such as *ml̓t* for *moult* or *enuś* for *envers*
- Tyronian notes such as *ꝛ* for *et*, or *ꝯ* for *con*.
- Allographetic (or graphic) variation such as *ſ* vs *s* or *ꝛ* for *r*

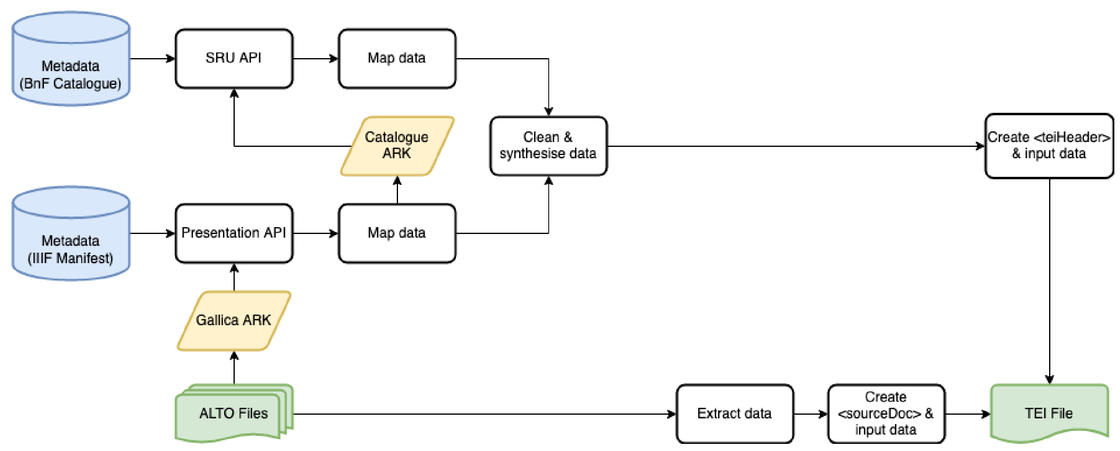
## DATA MODELLING

The information (text+layout) retrieved by HTR engines is encoded in ALTO-XML, which is not a format used by digital humanists. It needs to be converted in a standard format for text edition and publication: XML-TEI. Research had to be conducted on the optimal modelling, especially for the <sourceDoc> element.

```
<sourceDoc>
   <!-- Page -->
   <surface xml:id="f1" n="0" cert="gold"
            ulx="0" uly="0" lrx="3800" lry="5600">
      <graphic url="url"/>
      <!-- TextBlock "MainZone:column#1"-->
      <zone xml:id="f1_z1" type="MainZone" subtype="column"
            n="1" points="800,580 800,4700 2100,4700
            2100,580" source="url">
         <!-- TextLine "DefaultLine"-->
         <zone xml:id="f1_z1_l1" type="DefaultLine"
               subtype="subtype"  n="1"
               points="1150,730 1140,640 2000,600
               2030,700 2030,740 1150,759"
               source="url">
            <!-- baseline -->
            <path xml:id="f1_z1_l1_p"
               points="1150,730 2030,700"/>
            <!-- transcription -->
            <line xml:id="f1_z1_l1_t">A transcription</line>
         </zone>
      </zone>
   </surface>
</sourceDoc>
```

## TOWARDS AUTOMATED CORPUS BUILDING

The advantage of the TEI is the extended flexibility and the richness of its <teiHeader> (i.e. metadata). Via API such as the BnF's SRU service, it is possible to automate the metadata retrieval along with the text contained in the digital facsimile to create automatically a corpus from a list of ARK identifiers (i.e. unique identifiers of the digital library).




## FURTHER WORK

- Our models are not optimal for all the French sources: resources for incunabula, cursive writings, etc should be developed;
- Research is currently being made on the linguistic annotation of the data: lemmatisation, part-of-speech tagging, named entity recognition, linguistic normalisation...;
- Producing data is not enough: an efficient web application should be offered to give access to the information;
- The philological exploitation of the data remains the main objective of this project. Books and manuscripts processed by our pipeline will be used to answer linguistic and linguistic questions using data-driven approach, such as stylometry or natural language processing.