FoNDUE (FOrmes Numerisees et Detection Unifiee des Ecritures)

An Handwritting Text Recognition infrastructure for Geneva

Simon Gabay[1] and the SCIentific COmputing Support (SciCoS)[1]

[1]Université de Genève

cc BY

HTR

HTR augmented

HTR augmented : macro-level

HTR augmented : micro-level

Beyond HTR

HTR infrastructure

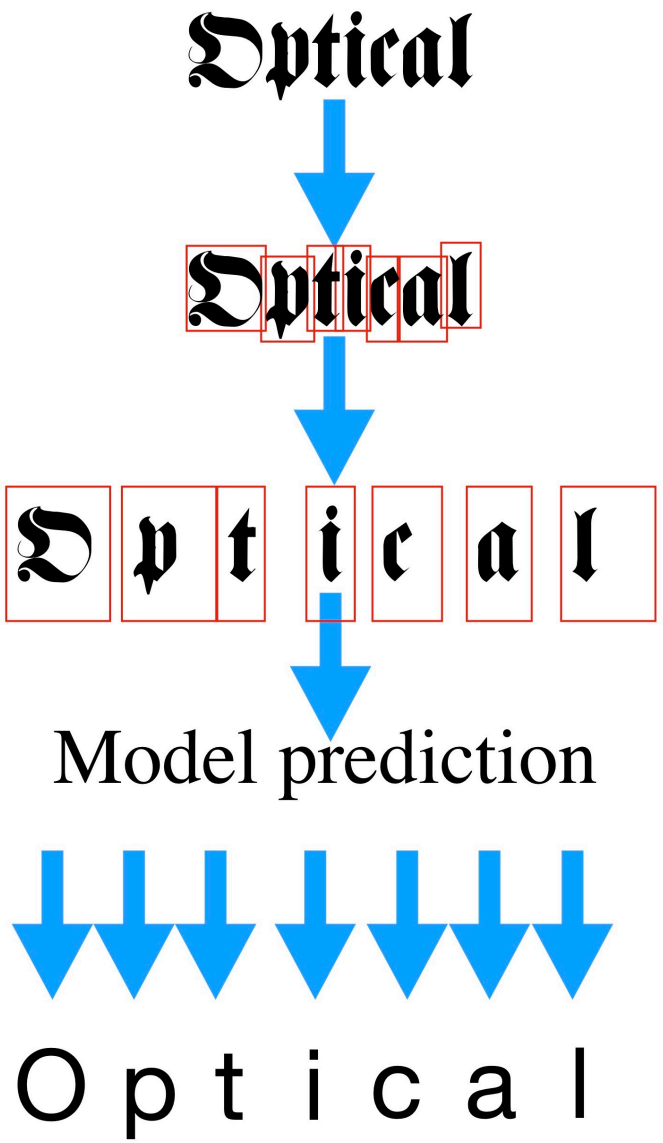
HTR

Definitions

- OCR : Optical Character Recognition
- HTR : Handwritten Text Recognition

Main difference is that OCR traditionally focuses on individual letters, while HTR processes larger chunks of texts, usually lines.

OCR



HTR

Problem : how can we segment letters with cursive scripts, the definition of which being precisely that (some) characters are written joined together in a flowing manner ?

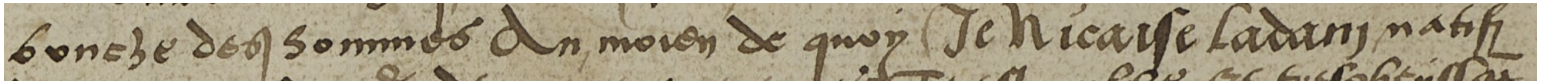
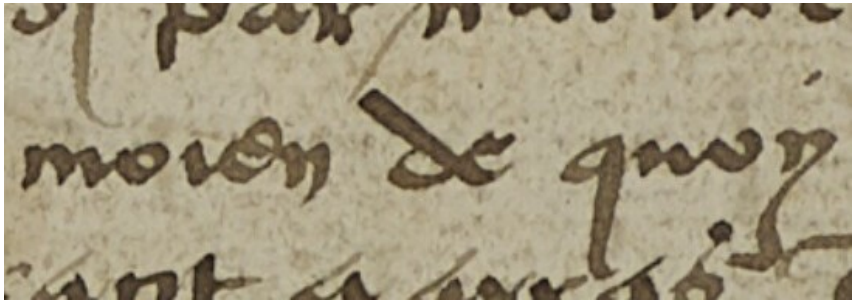


FIGURE – Ms Arras BM 187, *Chronique* de Nicaise Ladam, 16th c.



It is impossible to work at the letter level.

[Images from B. Salvati's PhD thesis, UniL]

Creating GT

Machine learning being primarily machine teaching, we are gathering as much transcription as possible to create models.

The question of the transcription, and especially complex glyphs (such as the long *s* *ſ*), remains an important problem.

Metadata for (OCR) models

Current metadata as proposed for the OCR-D project

```
<mets:xmlData>
    <gt:gt  xmlns:gt="http://www.ocr-d.de/GT/">
        <gt:state prop="acquisition"/>
        <gt:state prop="granularity/physical/document-related/page"/>
        <gt:state prop="granularity/physical/document-related/region"/>
        <gt:state prop="granularity/physical/document-related/text-line"/>
        <gt:state prop="granularity/physical/document-related/word"/>
        <gt:state prop="condition/production-related/document-faults/ink-from-facing"/>
        <gt:state prop="condition/wear/additions/informative/stamps"/>
        <gt:state prop="data-attributes/document-related/visual/text/font/typeface/blackletter"/>
        <gt:state prop="data-attributes/document-related/visual/text/font/multi-font/font-sizes"/>
        <gt:state prop="data-attributes/document-related/visual/text/font/multi-font/typefaces"/>
    </gt:gt>
</mets:xmlData>
```

FIGURE – XML Metadata

[Konstantin Baierer, Maria Federbusch, Clemens Neudecker, "Labelling OCR Ground Truth for Usage in Repositories", *DATeCH2019*.]

Metadata for (OCR) models

It is important to improve the quality of metadata, including information about the transcription using controlled vocabularies (*e.g.* TEI).


```
{
    "dc:title": "Oraison funebre de Marie Terese d'Austriche",
    "dc:creator": "Jacques Bénigne Bossuet",
    "dc:publisher": "Sébastien Mabre-Cramoisy, Paris",
    "dc:date": "1683",
    "dc:language": "fr",
    "dc:identifier": "https://catalogue.bnf.fr/ark:/12148/cb36575655n",
    "dc:type": "Ground truth",
    "dc:description": [
        {
            "gt:prop": [
                "data-attributes/document-related/visual/text/script/latin",
                "data-attributes/document-related/visual/text/font/typeface/antiqua",
                "data-attributes/document-related/visual/columns/one"
            ],
            "tei:normalization": "none",
            "tei:segmentation": "original"
        }
    ]
}
```


FIGURE – JSON-LD Metadata

[Simon Gabay, "Une introduction linguistique pour les données de Machine learning ?", *Humanistica 2020*.]

HTR augmented

HTR

HTR augmented

HTR augmented : macro-level

HTR augmented : micro-level

Beyond HTR

HTR infrastructure

HTR augmented : macro-level

Zone detection

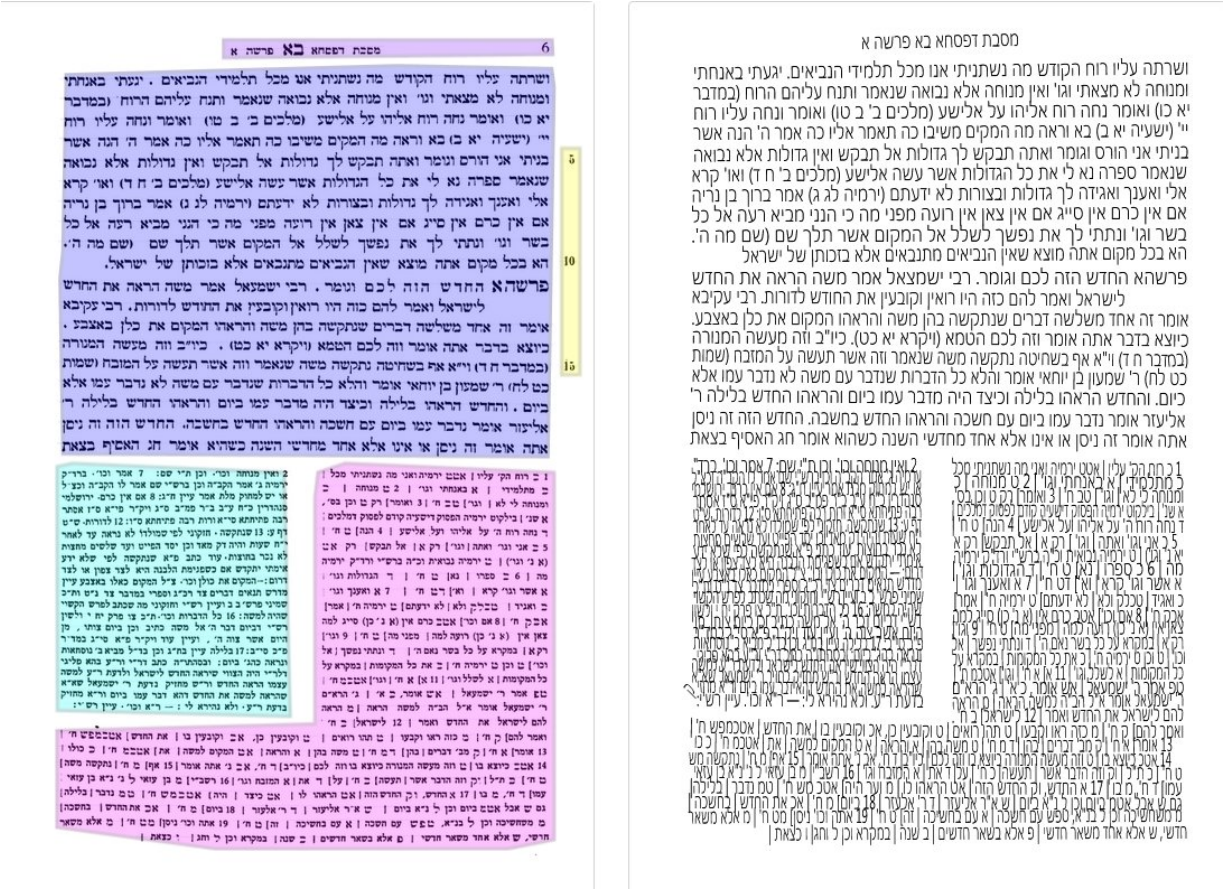
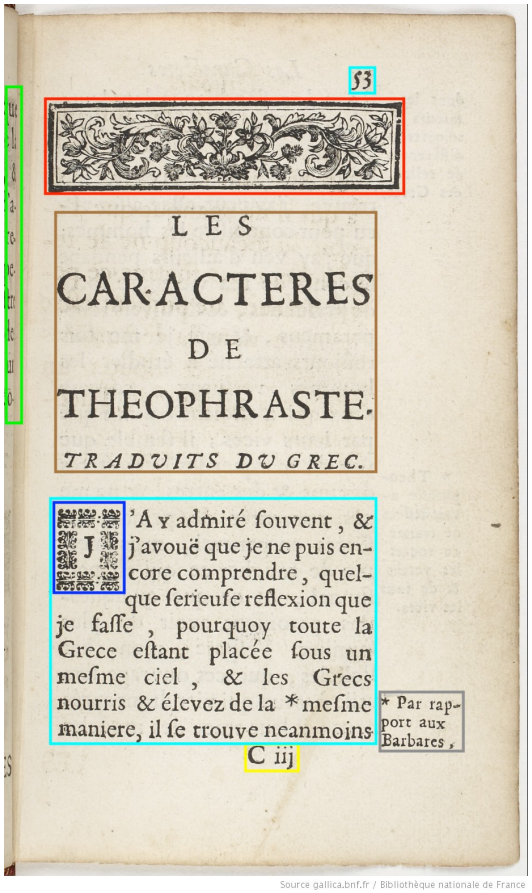


FIGURE – Zone detection on complex manuscripts. Courtesy of D. Stökl

[B. Kiessling, "A Modular Region and Text Line Layout Analysis System", *ICFHR 2020*.]

Image detection and categorisation


53

LES
CARACTERES
DE
THEOPHRASTE.
TRADUITS DU GREC.

J'AY admiré souvent, & j'avouë que je ne puis encore comprendre, quelque serieuse reflexion que je fasse, pourquoy toute la Grece estant placée sous un mesme ciel, & les Grecs nourris & élevez de la *mesme maniere, il se trouve neanmoins

* Par rapport aux Barbares,

C iij

Source gallica.bnf.fr / Bibliothèque nationale de France

FIGURE – La Bruyère, *Caractètres*

It is important not only to detect images but :

- Could we create a controlled vocabulary (better than the one developed for PageXML)
- Should we differentiate a illustration and and an initial letter ?
- Can an image contain text ?

Image and text

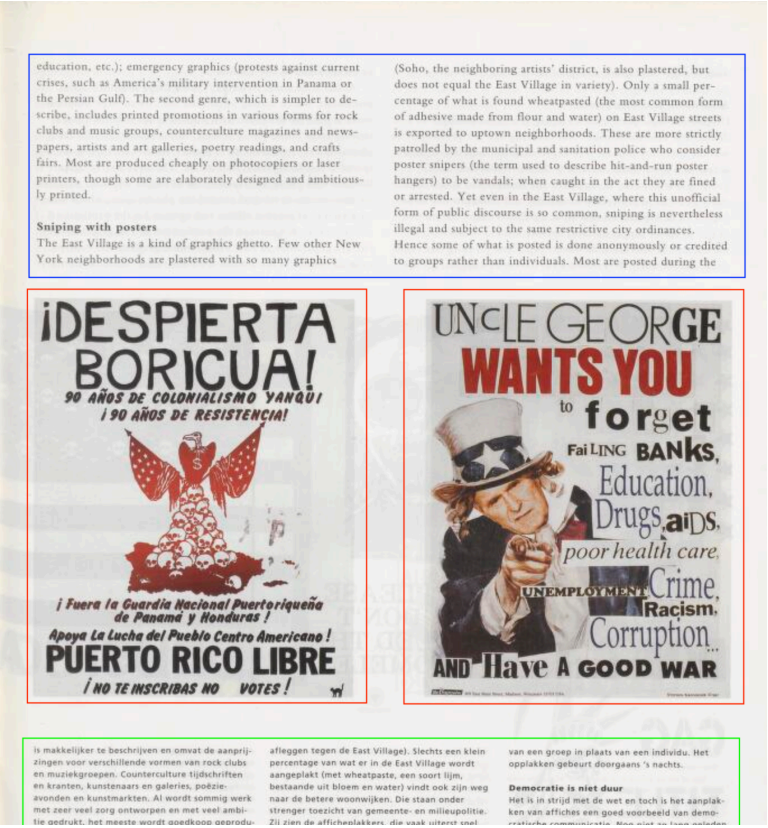


FIGURE – *Affiche*, 1 mars 1993

We need to

- Detect illustrations, which is a type of zone
- Use semantic rather than descriptive tagging
- Link the image with its textual context

HTR augmented : micro-level

HTR

HTR augmented

HTR augmented : macro-level

HTR augmented : micro-level

Beyond HTR

HTR infrastructure

Typographical information

We need to capture typographical information such as bold, italic, font size, font type. . .

mandrītēs einsiedler.

Nfr. *mandrite* m. „ermite, solitaire, moine“ (t. de l'Eglise d'Orient)“ (seit Enc 1765). — Nfr. *archimandrite* „titre de supérieur de couvent (t. de l'Eglise d'Orient)“ (seit Fur 1701); dazu *archimandritat* „bénéfice, revenu d'un archimandrite“ (seit Trév 1771), „titre d'archimandrite“ (seit Schwan 1787).

Lehnwörter aus auch ins spätlatein eingedrungenen termini der griech.-orthod. kirche. ThesLL; DC. — Keller.

FIGURE – "mandrites", *Französisches etymologisches Wörterbuch*.

Beyond HTR

HTR

HTR augmented

HTR augmented : macro-level

HTR augmented : micro-level

Beyond HTR

HTR infrastructure

Workflow

We believe that OCR is a first step toward further complex processing of data.
We would like to present here an example of post-processing of HTRised data :

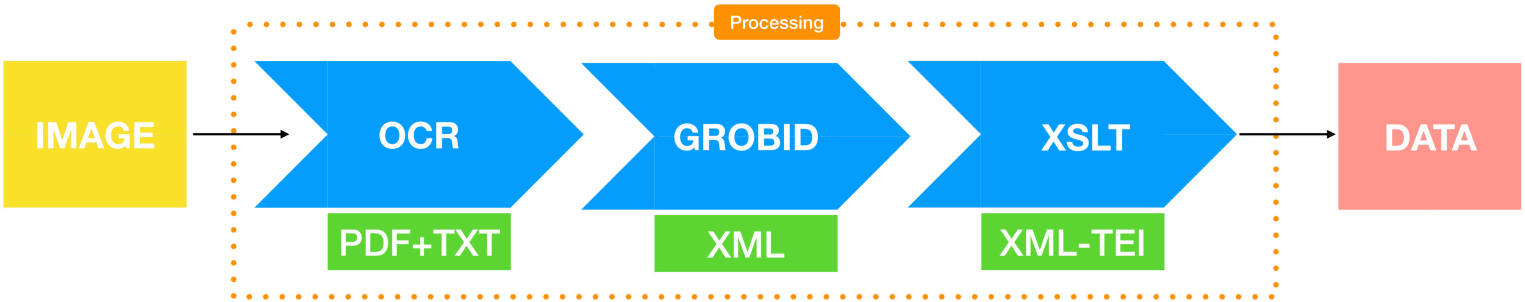


FIGURE – Workflow including OCR

Toward a database

Our goal is to transform an image :

55 **Scudéry** (Madeleine de), célèbre romancière du XVIIe siècle, créatrice du genre précieux, surnommée la *Sapho moderne*, née au Havre. — L. a. s. à Huet, (1686), 3 p. in-4. 100 »

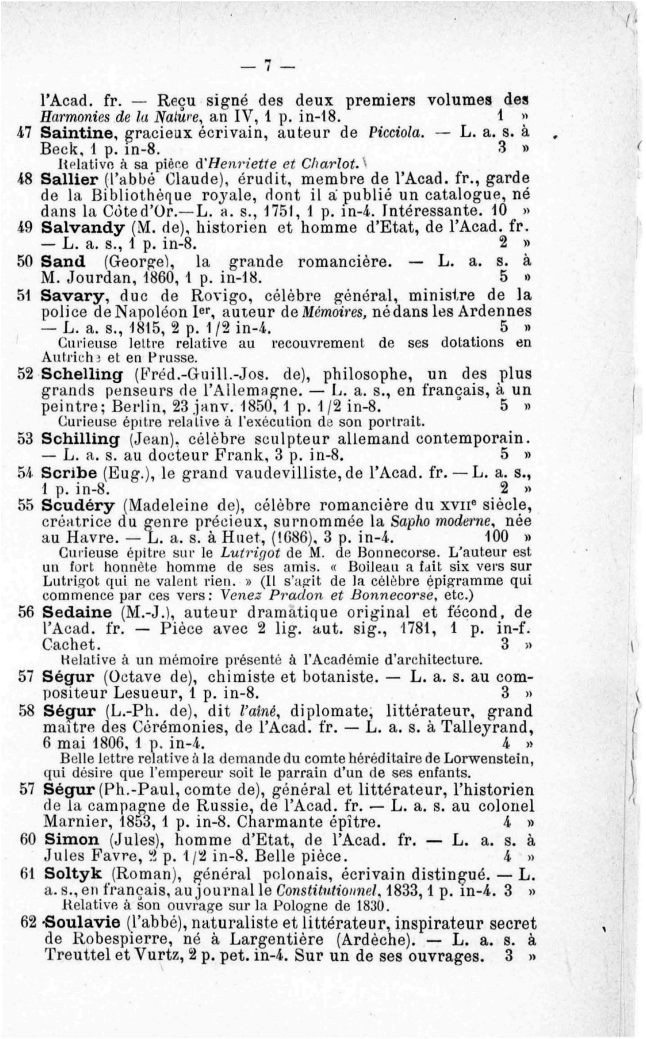
Curieuse épître sur le *Lutrigot* de M. de Bonnecorse. L'auteur est un fort honnête homme de ses amis. « Boileau a fait six vers sur Lutrigot qui ne valent rien. » (Il s'agit de la célèbre épigramme qui commence par ces vers : *Venez Pradon et Bonnecorse*, etc.)

Into a database encoded in a standard format like TEI to allow complex queries :

```
<item n="55" xml:id="CAT_000037_e55">
  <num>55</num>
  <name type="author">Scudéry (Madeleine de),</name>
  <trait>
    <p>célèbre romancière du XVIIe siècle, créatrice du genre précieux,
    surnommée la  Sapho moderne, née au Havre.</p>
  </trait>
  <desc>L.a.s. à Huet, (1686), 3 p. in-4. 100 ż</desc>
  <note>Curieuse épître sur le Lutrigot de M. de Bonnecorse. L'auteur
  est un fort honnête homme de ses amis. ń Boileau a fait six vers sur
  Lutrigot qui ne valent rien.ż  (Il s'agit de la célèbre épigramme
  qui commence par ces vers: Venez Pradon et Bonnecorse, etc.)</note>
</item>
```

FIGURE – TEI version of the Scudery entry

An example



— 7 —

l'Acad. fr. — Reçu signé des deux premiers volumes des *Harmonies de la Nature*, an IV, 1 p. in-18. 1 »

47 **Saintine**, gracieux écrivain, auteur de *Picciola*. — L. a. s. à Beck, 1 p. in-8. 3 »
Relative à sa pièce d'*Henriette et Charlot*.

48 **Sallier** (l'abbé Claude), érudit, membre de l'Acad. fr., garde de la Bibliothèque royale, dont il a publié un catalogue, né dans la Côte d'Or.—L. a. s., 1751, 1 p. in-4. Intéressante. 10 »

49 **Salvandy** (M. de), historien et homme d'Etat, de l'Acad. fr. — L. a. s., 1 p. in-8. 2 »

50 **Sand** (George), la grande romancière. — L. a. s. à M. Jourdan, 1860, 1 p. in-18. 5 »

51 **Savary**, duc de Rovigo, célèbre général, ministre de la police de Napoléon Ier, auteur de *Mémoires*, né dans les Ardennes — L. a. s., 1815, 2 p. 1/2 in-4. 5 »
Curieuse lettre relative au recouvrement de ses dotations en Autriche et en Prusse.

52 **Schelling** (Fréd.-Guill.-Jos. de), philosophe, un des plus grands penseurs de l'Allemagne. — L. a. s., en français, à un peintre; Berlin, 23 janv. 1850, 1 p. 1/2 in-8. 5 »
Curieuse épître relative à l'exécution de son portrait.

53 **Schilling** (Jean), célèbre sculpteur allemand contemporain. — L. a. s. au docteur Frank, 3 p. in-8. 5 »

54 **Scribe** (Eug.), le grand vaudevilliste, de l'Acad. fr. — L. a. s., 1 p. in-8. 2 »

55 **Scudéry** (Madeleine de), célèbre romancière du XVIIe siècle, créatrice du genre précieux, surnommée la *Sapho moderne*, née au Havre. — L. a. s. à Huet, (1686), 3 p. in-4. 100 »
Curieuse épître sur le *Lutrigot* de M. de Bonnecorse. L'auteur est un fort honnête homme de ses amis. « Boileau a fait six vers sur Lutrigot qui ne valent rien. » (Il s'agit de la célèbre épigramme qui commence par ces vers: *Venez Pradon et Bonnecorse*, etc.)

56 **Sedaine** (M.-J.), auteur dramatique original et fécond, de l'Acad. fr. — Pièce avec 2 lig. aut. sig., 1781, 1 p. in-f. Cachet. 3 »
Relative à un mémoire présenté à l'Académie d'architecture.

57 **Ségur** (Octave de), chimiste et botaniste. — L. a. s. au compositeur Lesueur, 1 p. in-8. 3 »

58 **Ségur** (L.-Ph. de), dit *l'aîné*, diplomate, littérateur, grand maître des Cérémonies, de l'Acad. fr. — L. a. s. à Talleyrand, 6 mai 1806, 1 p. in-4. 4 »
Belle lettre relative à la demande du comte héréditaire de Lorwenstein, qui désire que l'empereur soit le parrain d'un de ses enfants.

57 **Ségur** (Ph.-Paul, comte de), général et littérateur, l'historien de la campagne de Russie, de l'Acad. fr. — L. a. s. au colonel Marnier, 1853, 1 p. in-8. Charmante épître. 4 »

60 **Simon** (Jules), homme d'Etat, de l'Acad. fr. — L. a. s. à Jules Favre, 2 p. 1/2 in-8. Belle pièce. 4 »

61 **Soltyk** (Roman), général polonais, écrivain distingué. — L. a. s., en français, au journal le *Constitutionnel*, 1833, 1 p. in-4. 3 »
Relative à son ouvrage sur la Pologne de 1830.

62 **Soulavie** (l'abbé), naturaliste et littérateur, inspirateur secret de Robespierre, né à Largentière (Ardèche). — L. a. s. à Treuttel et Vurtz, 2 p. pet. in-4. Sur un de ses ouvrages. 3 »

FIGURE – *Revue des autographes*, No. 67 (March 1881), p. 7.

Data extraction

55 **Scudéry** (Madeleine de), célèbre romancière du XVIIe siècle, créatrice du genre précieux, surnommée la *Sapho moderne*, née au Havre. — L. a. s. à Huet, (1686), 3 p. in-4. 100 »

Curieuse épître sur le *Lutrigot* de M. de Bonnecorse. L'auteur est un fort honnête homme de ses amis. « Boileau a fait six vers sur Lutrigot qui ne valent rien. » (Il s'agit de la célèbre épigramme qui commence par ces vers : *Venez Pradon et Bonnecorse*, etc.)

55 **Scudéry** (Madeleine de), célèbre romancière du XVIIe siècle, créatrice du genre précieux, surnommée la *Sapho moderne*, née au Havre. — L. a. s. à Huet, (1686), 3 p. in-4. 100 »

Curieuse épître sur le *Lutrigot* de M. de Bonnecorse. L'auteur est un fort honnête homme de ses amis. « Boileau a fait six vers sur Lutrigot qui ne valent rien. » (Il s'agit de la célèbre épigramme qui commence par ces vers : *Venez Pradon et Bonnecorse*, etc.)

55 **Scudéry** (Madeleine de), célèbre romancière du XVIIe siècle, créatrice du genre précieux, surnommée la *Sapho moderne*, née au Havre. — L. a. s. à Huet, (1686), 3 p. in-4. 100 »

Curieuse épître sur le *Lutrigot* de M. de Bonnecorse. L'auteur est un fort honnête homme de ses amis. « Boileau a fait six vers sur Lutrigot qui ne valent rien. » (Il s'agit de la célèbre épigramme qui commence par ces vers : *Venez Pradon et Bonnecorse*, etc.)

The *Katabase* project

The *Katabase* project digitise sale catalogues of manuscripts :

`https://katabase.huma-num.fr`

All data are available online

`https://github.com/katabase`

It involves much more data processing, for instance to reconcile similar entries (same ms sold multiple times) and trace their history in time (history of collection, evolution of prices. . .)

HTR infrastructure

HTR

HTR augmented

HTR augmented : macro-level

HTR augmented : micro-level

Beyond HTR

HTR infrastructure

Providing services

We believe an important part of research in DH is to provide infrastructure to colleagues.

We are committed in using open-source and free solution that everyone can benefit from.

An application for HTR

We want to provide a simple interface for non-specialists.
We have decided to use *eScriptorium*, developed in Paris.

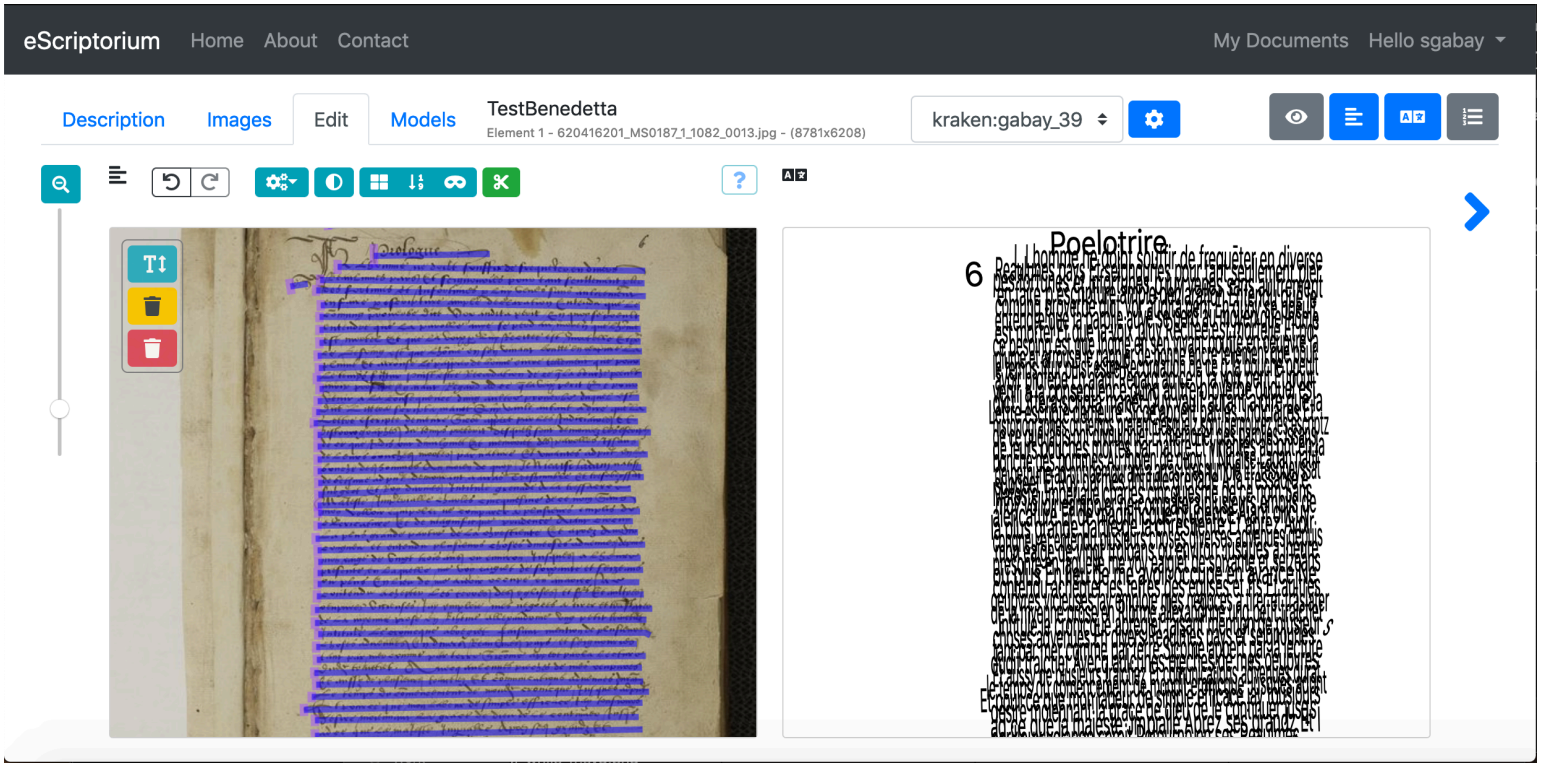


FIGURE – *escriptorium*

[Kiessling, B. ; Tissot, R., Stökl Ben Ezra, D., Stokes, "P. eScriptorium : An Open Source Platform for Historical Document Analysis", OST@ICDAR 2019]

Building an infrastructure in Geneva ?

In collaboration with Parisian teams (INRIA+PSL) we would like to develop an infrastructure for Swiss scholars in Geneva, but also in other universities.

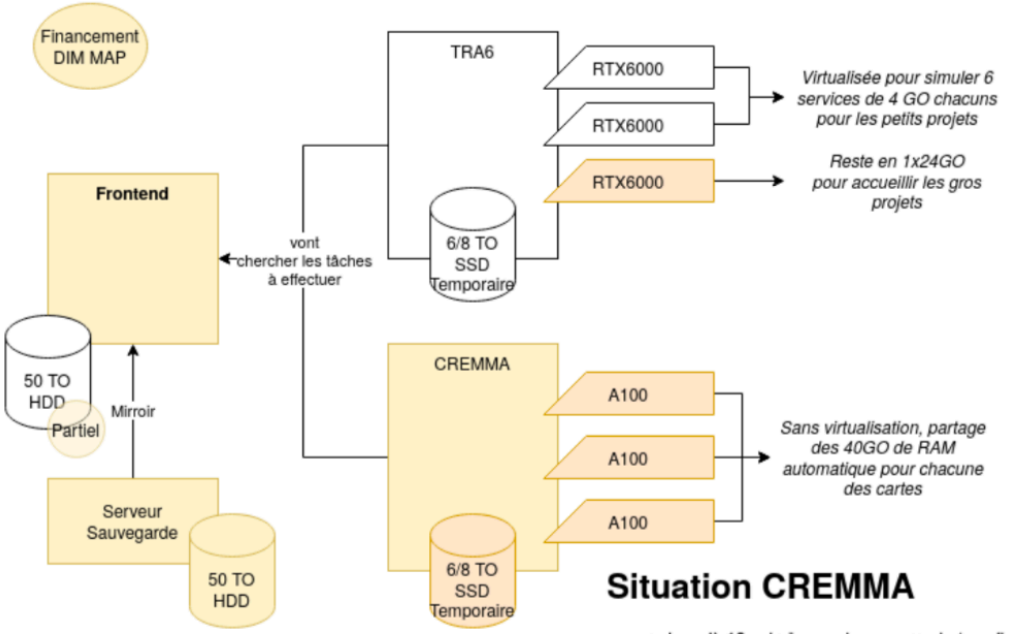


FIGURE – CREMMA infrstructure in Paris

Partners are welcome !

[CREMMA structure drawn by Th. Clerice.]

Building an infrastructure : HTR on HPC

Improving *eScriptorium* in order to run heavy computations on HPC clusters.

Design goals :

- Keeping the user-friendliness of *eScriptorium* ;
- Seamless integration with clusters authentication and accounting system ;
- Contributing code to the open-source codebase.

Development by SciCoS : Christophe Charpilloz, Pierre Kuenzli, Jean-Luc Falcone.