

# Transcrire (automatiquement)

Alexandre Bartz, Simon Gabay



# Exporter

# Données

Il faut réussir décrire les documents OCRisés, afin de reconstruire l'apparence originelle sur la base des informations conservées. On privilégie pour cela des documents XML, *page driven*.

```
<document>
  <page>
    <zone>
      <ligne coordonnées="points">
        <mot coordonnées="points">exemple</mot>
      <ligne>
    </zone>
  </page>
</document>
```

# Données

Afin de faire le lien entre l'image et le texte, on doit donner une information géométrique. Celle-ci peut être de trois ordres: ligne, bloc, ou polygone.

Il existe des documents de niveau page, paragraphe, ligne ou mot.

Il existe aussi plusieurs formats: hOCR, Alto, PageXML... Ces formats sont normalement utilisés avec METS (*Metadata Encoding and Transmission Standard*) pour la description de l'objet numérisé.

# Exemple 1: Alto

ALTO: *Analyzed Layout and Text Object*

Développé lors du projet européen METAe (*Meta Data engine*, 2000-2003) et publié en 2004

Trois éléments centraux:

- <Description> contient les métadonnées
- <Styles> contient le texte
  - <TextStyle> contient les informations sur les fontes (famille, type, taille...)
  - <ParagraphStyle> contient la description des paragraphes (alignement gauche/droite, intelrigne)
- <Layout> contient le contenu, divisé en <Page>

```
<?xml version="1.0"?>
<alto>
    <Description>
        <MeasurementUnit/>
        <sourceImageInformation/>
        <Processing/>
    </Description>
    <Styles>
        <TextStyle FONTSIZE="10.0"/>
        <ParagraphStyle ALIGN="Left"/>
    </Styles>
    <Layout>
        <Page ID="P1" WIDTH="123" HPOS="123" VPOS="123">
            <PrintSpace WIDTH="123" HPOS="123" VPOS="123">
                <TextBlock ID="P1_TB1" WIDTH="123" ...>
                    <TextLine WIDTH="123" HPOS="123" ...>
                        <String WIDTH="123" ... CONTENT="Un">
                            <sp WIDTH="123" HPOS="123" VPOS="123">
                                <String WIDTH="123" ... CONTENT="Exemple">
                            </TextLine>
                        </TextBlock>
                    </PrintSpace>
                </Page>
            </Layout>
        </alto>
```

## Exemple 2: PageXML

PAGE: *Page Analysis and Ground-truth Elements*

Format créé lors du projet IMPACT EU (2010)

Contrairement à l'ALTO, PageXML conserve des informations sur le *pre-processing* (binarisation, deskew, dewarping...) et l'évaluation du layout.

```
<PcGts>
  <Metadata>...</Metadata>
  <page>
    <TextRegion type="paragraph" id="r_1">
      <Coords points="1474,486 3684,486 3684,900...">
      <TextLine id="l_1">
        <Coords points="1475,487 3683,487 3683,635...">
        <Baseline points="1475,635 1587,635 2061...">
        <Word id="w1">
          <Coords points="1475,497 1587,497 1587...">
          <TextEquiv>
            <Unicode>Un</Unicode>
          </TextEquiv>
        </Word>
        <Word id="w2">
          <Coords points="1935,497 2061,497 2061,619...">
          <TextEquiv>
            <Unicode>exemple</Unicode>
          </TextEquiv>
        </Word>
        <TextEquiv>
          <Unicode>Un exemple</Unicode>
        </TextEquiv>
      </TextLine>
    </TextRegion>
  </page>
</PcGts>
```

## Exemple 3: hOCR

Format XML *embedded* dans du XHTML/HTML

Trois grandes classes associées aux éléments html `<div>` , `<p>` ,  
`<span>`

- `ocr_page` pour les pages
- `ocr_par` pour les paragraphes
- `ocrx_line` pour les lignes
- `ocrx_word` pour les mots

L'information géométrique est stockée dans une bbox

```
<?xml version="1.0" encoding="UTF-8"?>
<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <title></title>
    <meta name='ocr-system' content='tesseract'/>
  </head>
  <body>
    <div class='ocr_page' id='page_1'
        title='bbox 0 0 1926 3102'>
      <div class='ocr_carea' id='block_1_1'
          title="bbox 638 108 756 147">
        <p class='ocr_par' id='par_1_1' lang='eng'
            title="bbox 638 108 756 147">
          <span class='ocr_line' id='line_1_1'
              title="bbox 638 108 756 147;
                      baseline 0 0">
            <span class='ocrx_word' id='word_1_1'
                title='bbox 638 108 756 147'>
              exemple
            </span>
          </span>
        </p>
      </div>
    </div>
  </body>
```

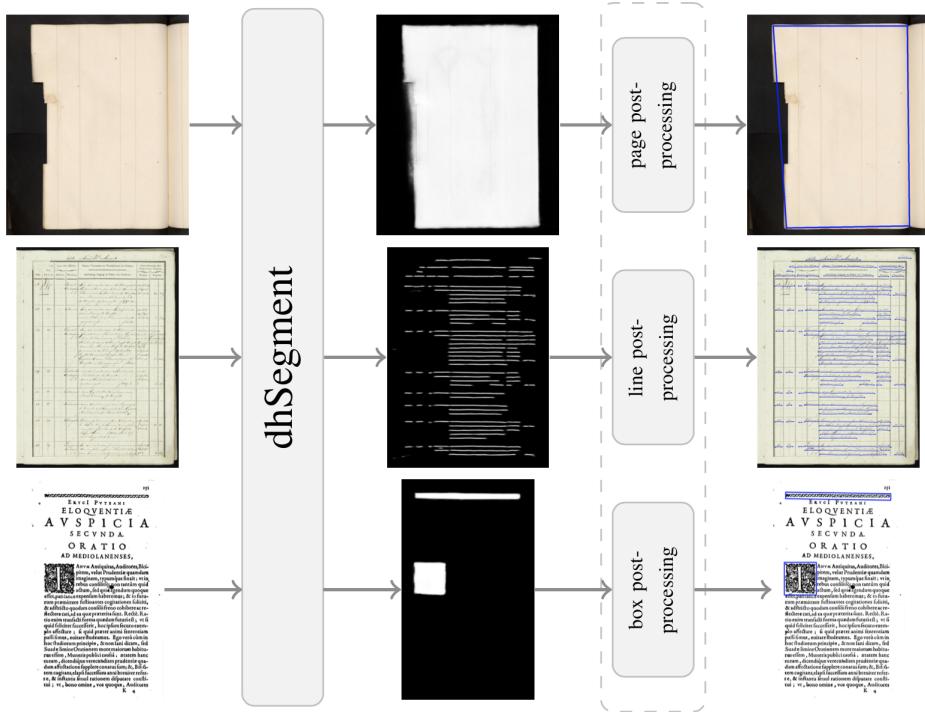
# Dans la jungle des outils

# Outils

- Tesseract
- Ocropy
- Kraken
- Calamari
- DHsegment
- ...

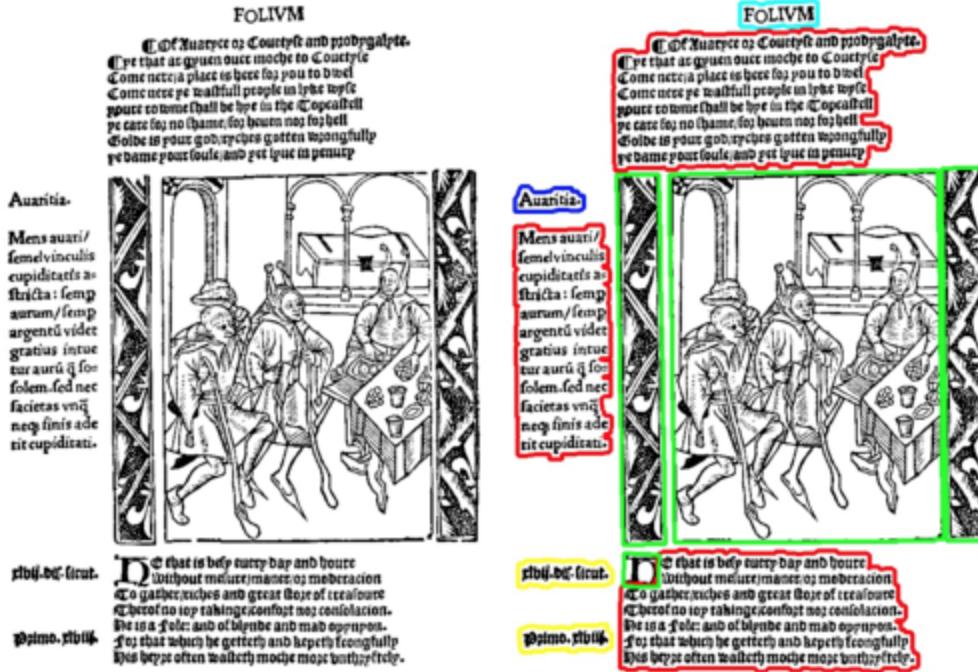
Il est souhaitable de préférer une solution qui intègre les différentes étapes nécessaires à l'OCrisation

# DHsegment



Sofia Ares Oliveira, Benoit Seguin, Frederic Kaplan, "dhSegment: A generic deep-learning approach for document segmentation", v.2, [arXiv:1804.10371](https://arxiv.org/abs/1804.10371)

# Larex



Reul, C., Springmann, U., and Puppe, F., "LAREX - A semi-automatic open-source Tool for Layout Analysis and Region Extraction on Early Printed Books", [arXiv:1701.07396](https://arxiv.org/abs/1701.07396)

# OCR4all

**Fueillet**

elij. di. que  
scamp igle.

je font desquesles ie me tais. O poures folz belisfrés qui de robes censp qui n'ot  
pas du pain a manger et aluenture quiz n'osent demander de honte les dieilles  
gens / poures deues / ladires / auengles / helas penes y / car certes vous en ten/  
tredes compte devant ceuluy qui nous crea

**¶ Des conditioñs courroñx es grandes mauuaisties des femeñs**

Si colligo  
op runcida  
nemo magi  
gaudei q̄ re  
ra

¶ Pluieñts asnes chevaucheroient  
Le nest que monte y eust femme  
Au moyen de quoy ne vouldroient  
Monte dessus po euse disfame

puer. esti  
Eccl. ges.  
puerl. ges.

Quiz ont fait z grief infame  
Au pouure asne z grans toymes  
Car luy ont toz tous ses bons mes  
Car luy ont toz touo fcs bons mes

**G** Ntendes folz e/  
ffourdis e vous  
congnostres les  
mauaisties des  
femeñs. Aussy femeñs appio  
cher doz doz orres bone ma  
tiere. Mes versetz sitiez ces  
criptz vousdroiet des femeñs

LOAD SAVE

ä	æ	b	ɛ	æ	d	ə
ē	ɛ	ff	ft	g	g	ø
í	ɪ	ll	m	n	ð	œ
ō	ɒ	p	p	p	p	P
ö	ɔ	ç	ç	ç	ç	ç
ü	ü	q	q	q	q	q
ñ	ñ	ñ	ñ	ñ	ñ	ñ
í	í	í	í	í	í	í
ó	ó	ó	ó	ó	ó	ó
y	y	y	w	x	y	z
ö	ö	ö	-	ö	ö	ö
ü	ü	ü	-	ü	ü	ü

**Quiz ont fait z grief infame**  
Ouilz ont fait z grief infame

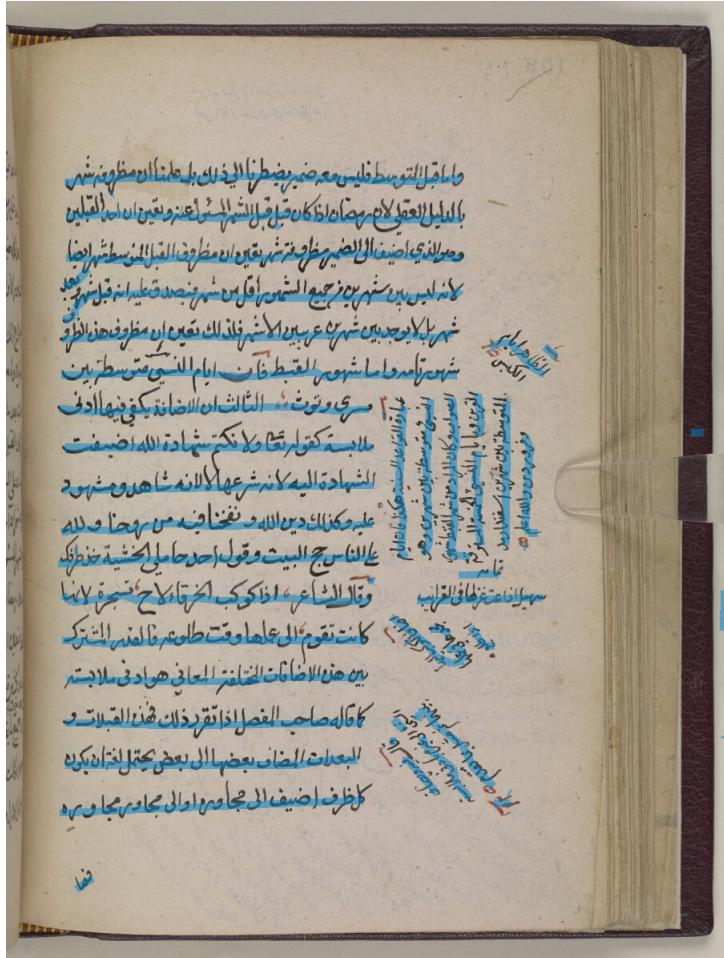
**Au pouure asne z grans toymes**  
Au pouure asne z grans toymes

**Car luy ont toz tous ses bons mes**  
Car luy ont toz tous ses bons mes

**Ntendes folz e/**  
Ntendes folz e=

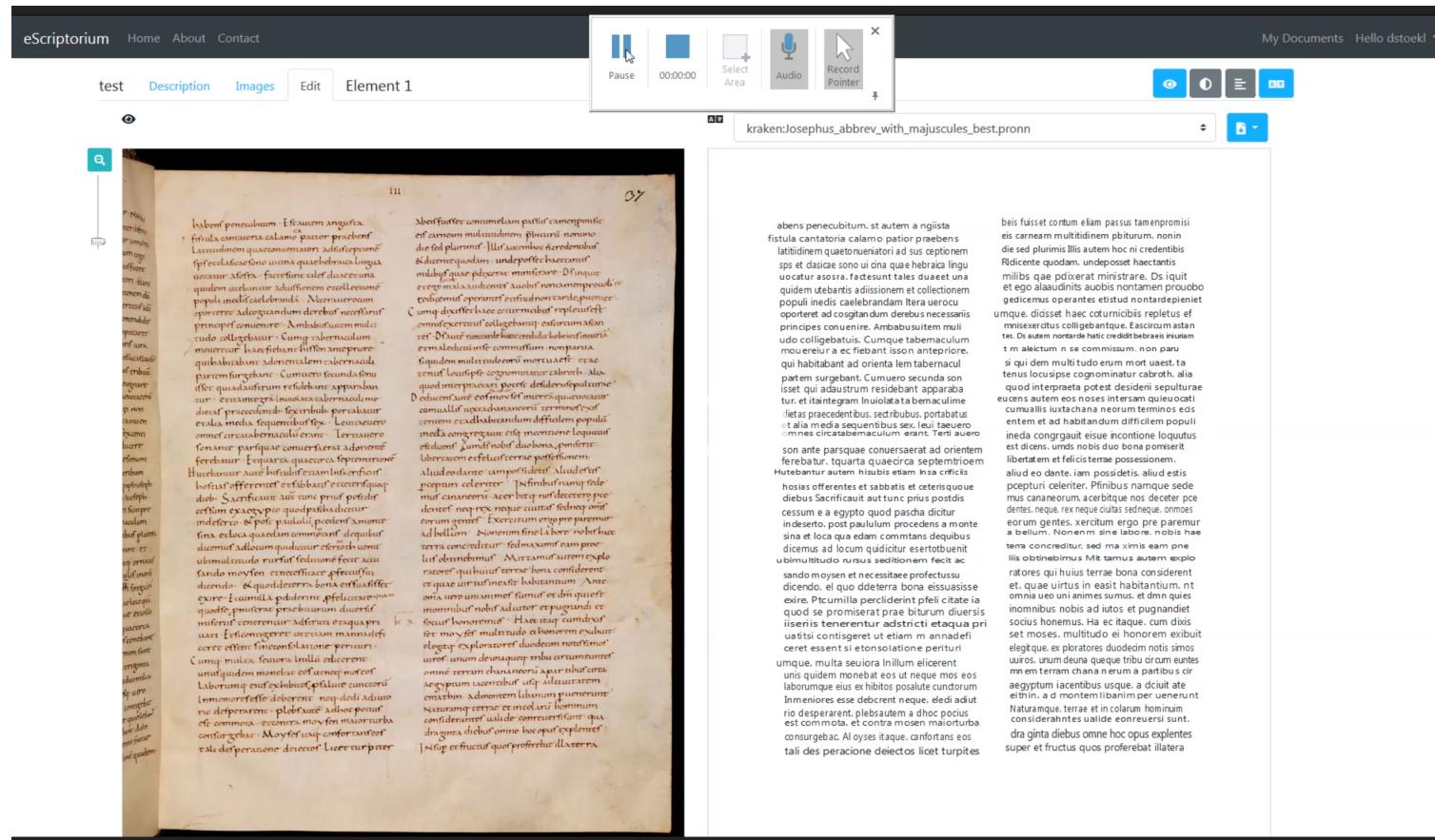
Reul C, Christ D, Hartelt A, Balbach N, Wehner M, Springmann U, Wick C, Grundig C, Büttner A, Puppe F. "OCR4all—An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Paintings". *Applied Sciences*. 2019; 9(22):4853. DOI: 10.3390/app9224853

# Kraken



Kiessling, B., Stökl Ben Ezra, D., Miller M., "BADAM: A Public Dataset for Baseline Detection in Arabic-script Manuscripts", HIP@ICDAR 2019.  
[arXiv:1907.04041](https://arxiv.org/abs/1907.04041)

# eScriptorium



Kiessling, B.; Tissot, R., Stökl Ben Ezra, D., Stokes P. "eScriptorium: An Open Source Platform for Historical Document Analysis", OST@ICDAR 2019 (2019) <https://ieeexplore.ieee.org/document/8893029>

## *Pipeline*

Il existe plusieurs solutions qui articulent tous les éléments nécessaires pour l'OCRisation

- Web: [eScriptorium](#)
- Docker: [ocr4all](#)
- Java: [Transkribus](#)

## *Pipeline : Comment choisir?*

Il existe plusieurs solutions qui articulent tous les éléments nécessaires pour l'OCRisation

- Web: [eScriptorium](#), *open source*
- Docker: [ocr4all](#), *open source*
- Java: [Transkribus](#), *non open source*

