From eScriptorium to TEI Publisher

Hugo Scheithauer, Alix Chagué, Laurent Romary

From eScriptorium to TEI Publisher

Brace your digital scholarly edition!
11/19/21

Hugo Scheithauer, Alix Chagué, Laurent Romary

Overview - from source documents to scholarly editions

Context: heterogeneous formats at the various stages of a digitisation workflow

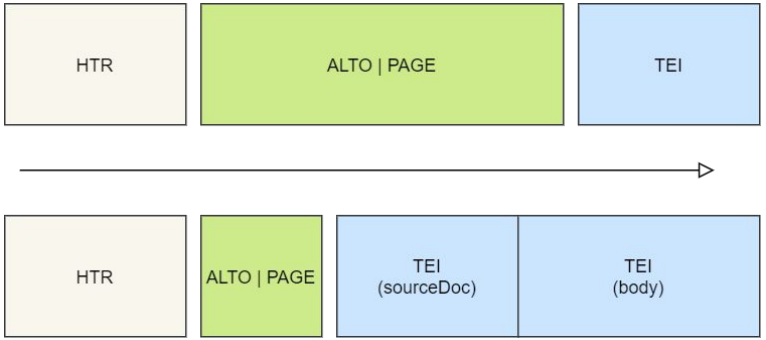
- ALTO/PAGE XML vs. TEI

Objectives:

- Improve reusability of content at various stages
- Hence, easier combination of different software solutions

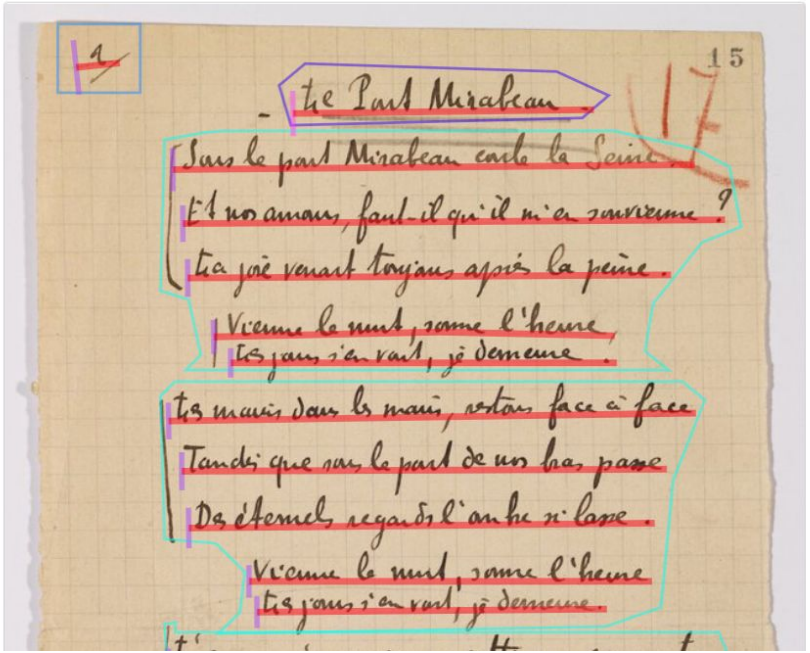
Results:

- Propagation of metadata elements
- Mapping of layout, transcription information
- Architecture for integrating the content in further editions



What is eScriptorium?

- An open-source software, developed by the research team Scripta (PSL)
- Provides an interface for:
  - document segmentation,
  - layout annotation,
  - transcribing (manually or automatically),
  - and training OCR/HTR models.



1

Le Pont Mirabeau.

Sous le pont Mirabeau coule la Seine.

Et nos amours, faut-il qu'il m'en souvienne ?

La joie venait toujours après la peine.

Vienne la nuit, sonne l'heure,
Les jours s'en vont, je demeure.

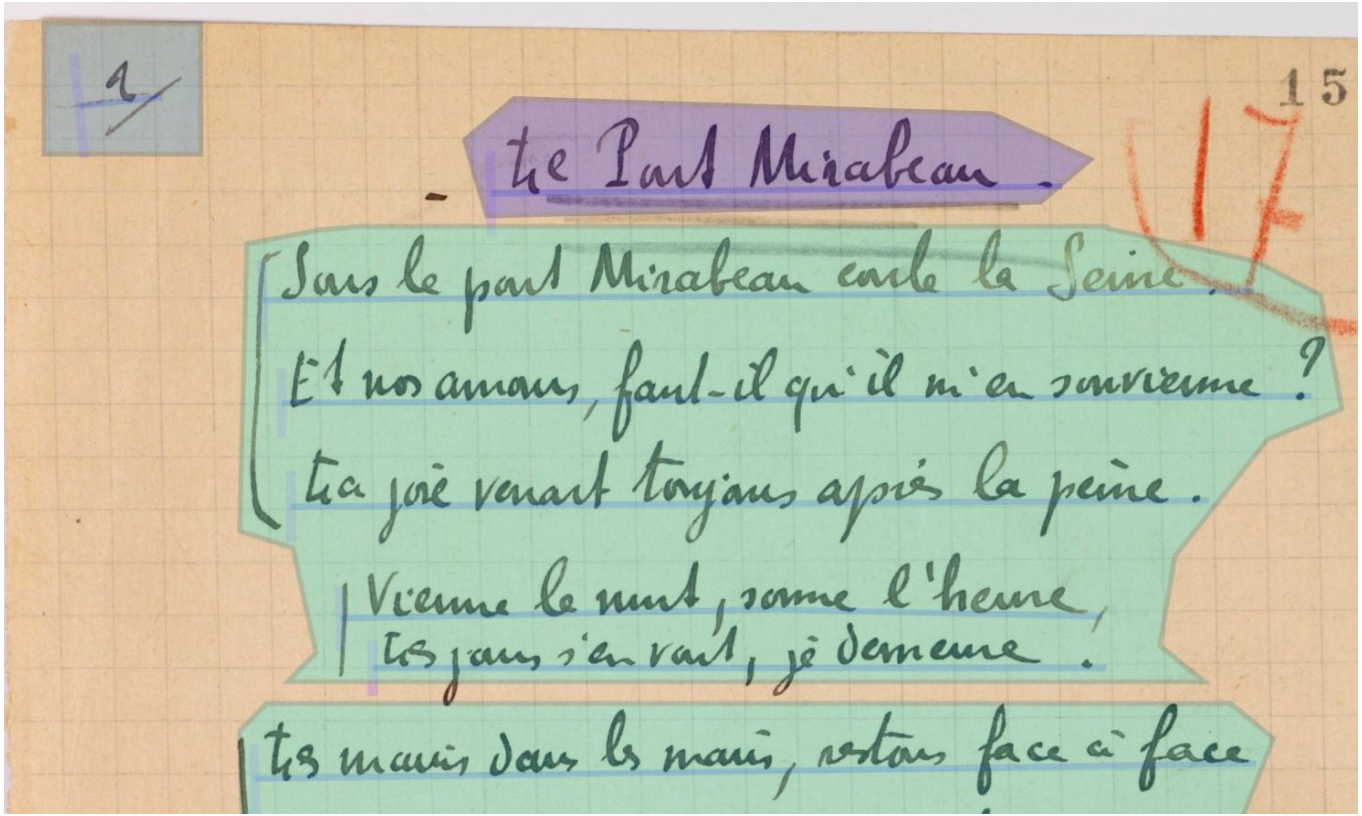
Les mains dans les mains, restons face à face

Tandis que sous le pont de nos bras passe

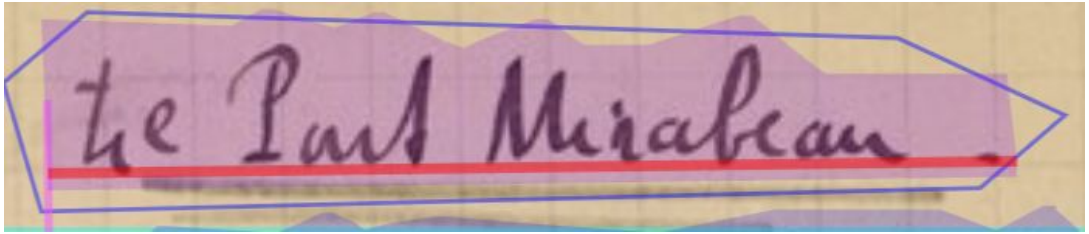
Des éternels regards l'ombre si lasse.

Vienne la nuit, sonne l'heure
Les jours s'en vont, je demeure.

A page layout can be segmented into one or several **text regions**, each with their own coordinates:



Text lines can be nested in a text region.



A line of text combines:

- A baseline (red) or a topline: a line defined by at least 2 points
- A mask (purple): a polygon defined by at least 3 points
- A text node

eScriptorium currently uses three formats for transcription export:

- PAGE XML
- ALTO
- Raw text

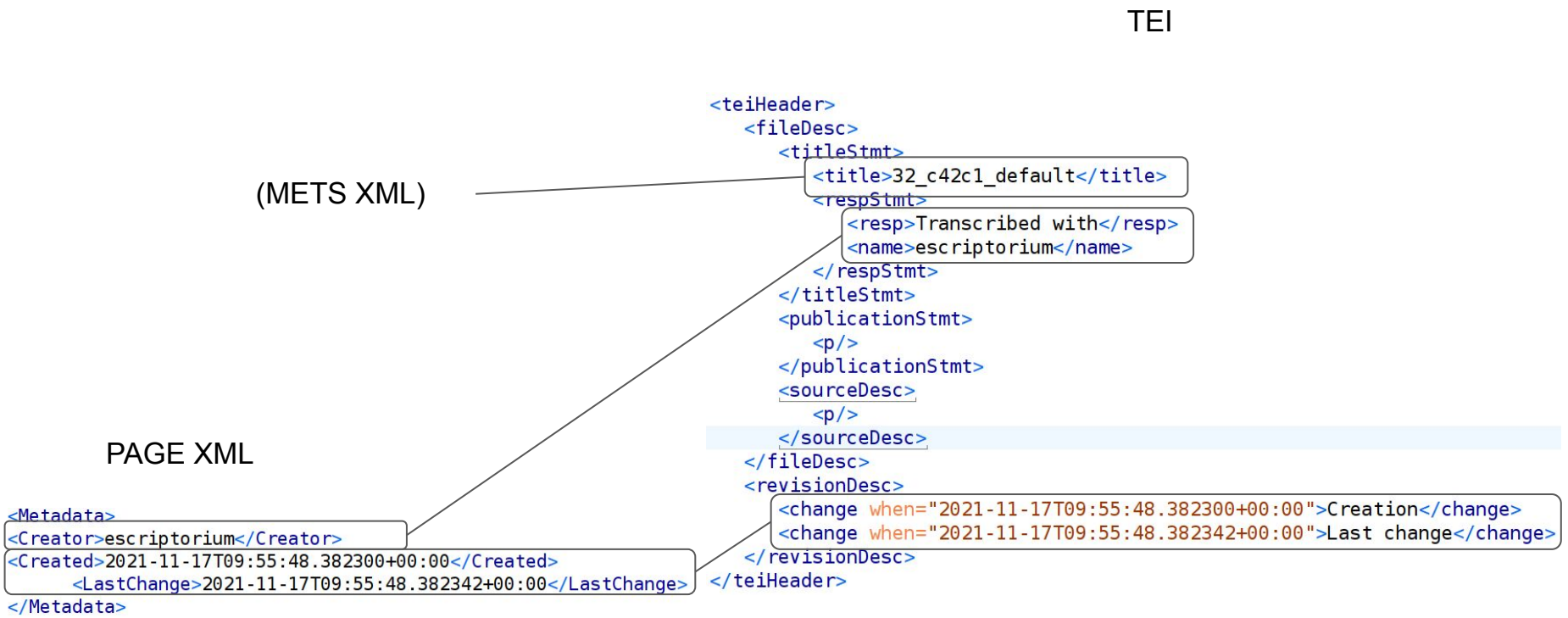
A PAGE XML example

```
<?xml version="1.0" encoding="UTF-8"  standalone="yes"?>
<PcGts xmlns="http://schema.primaresearch.org/PAGE/gts/pagecontent/2019-07-15" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http:/,
  <Metadata>
  <Creator>escriptorium</Creator>
  <Created>2021-11-17T09:55:48.382300+00:00</Created>
        <LastChange>2021-11-17T09:55:48.382342+00:00</LastChange>
  </Metadata>
  <Page imageFilename="32_c42c1_default.jpg" imageWidth="2312" imageHeight="3469">
    <TextRegion id="eSc_textblock_dde8b8e9"  custom="structure {type:numbering;}">
      <Coords points="154,59 154,257 392,257 392,59"/>
      <TextLine id="eSc_line_106ba71b" >
        <Coords points="205,182 199,118 237,118 240,109 251,95 254,95 254,92 254,92 257,92 257,92 257,92 260,92 260,92 260,92 263,92 324,98 332,170 338,251 211
        <Baseline points="207,184 333,171"/>
        <TextEquiv>
          <Unicode>1</Unicode>
        </TextEquiv>
      </TextLine>
    </TextRegion>
    <TextRegion id="eSc_textblock_e94aafae"  custom="structure {type:line_group;}">
      <Coords points="1770,370 1995,420 2093,471 2126,644 2013,690 1872,867 1849,1003 1917,1058 524,1058 567,1014 532,855 451,829 462,393 576,372 633,367 1769,
      <TextLine id="eSc_line_c4880d79" >
        <Coords points="486,479 480,358 549,364 549,364 552,364 552,364 552,364 555,364 555,364 555,367 590,401 659,401 700,361 700,358 700,358 700,358 703,358
        <Baseline points="486,479 1993,465"/>
        <TextEquiv>
          <Unicode>Sous le pont Mirabeau coule la Seine.</Unicode>
        </TextEquiv>
      </TextLine>
      <TextLine id="eSc_line_8af91efd" >
        <Coords points="515,635 509,537 636,534 639,534 639,534 642,534 642,534 656,540 737,569 778,555 778,555 781,555 781,555 784,555 784,555 784,555 787,555
        <Baseline points="515,635 2080,624"/>
        <TextEquiv>
          <Unicode>Et nos amours, faut-il qu&amp;#39;il m&amp;#39;en souvienne ?</Unicode>
        </TextEquiv>
      </TextLine>
      <TextLine id="eSc_line_4ab1ac7a" >
```

From page to content

Which TEI representation for HTR output?

Documenting a transcription: metadata representation with TEI



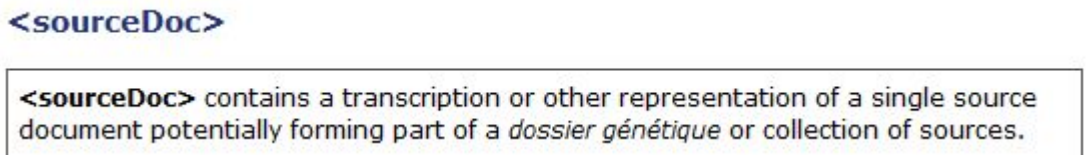
Any metadata missing for documenting an automatic transcription?

→ The transcription model
→ Documenting automatic and manual post-processing, such as correction
→ Information regarding how the transcription was produced

Which TEI representation for the transcription itself?

What does the TEI have to offer for the representation of data resulting from HTR/OCR?

Beyond facsimiles, the **<sourceDoc>** element:



Screenshot from the TEI guidelines for the <sourceDoc> element (https://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-sourceDoc.html)

How are we using the <sourceDoc>?

Two key principles:

→ The <sourceDoc> must be the strict transposition of all automatic transcription output elements.

→ All elements in <body> are the user’s responsibility and contain their interpretation/edition of the transcription.

Representation of a page with TEI using the **<sourceDoc>** element

The PAGE XML <Page> element, with basic metadata, for instance:

```
<Page imageFilename="32_c42c1_default.jpg" imageWidth="2312" imageHeight="3469">
```

becomes a <sourceDoc> element in TEI:

```
<sourceDoc>
   <graphic url="32_c42c1_default.jpg" width="2312px" height="3469px"/>
```

PAGE XML and TEI structure of an image content: **text regions and baselines**

PAGE XML

```
<TextRegion id="eSc_textblock_dde8b8e9"  custom="structure {type:numbering;}">
  <Coords points="154,59 154,257 392,257 392,59"/>
  <TextLine id="eSc_line_106ba71b" >
    <Coords points="205,182 199,118 237,118 240,109 251,95 254,95 254,92 254,92
    <Baseline points="207,184 333,171"/>
    <TextEquiv>
      <Unicode>1</Unicode>
    </TextEquiv>
  </TextLine>
</TextRegion>
      ...
```

TEI

```
<surfaceGrp>
   <surface xml:id="eSc_textblock_dde8b8e9"
            type="structure_{type:numbering;}"
            points="154,59 154,257 392,257 392,59">
      <zone xml:id="eSc_line_106ba71b"
            type="mask"
            points="205,182 199,118 237,118 240,109 251,95 254,95
         <path type="baseline" points="207,184 333,171"/>
         <line>1</line>
      </zone>
   </surface>
        ...
```

All TEI elements together:

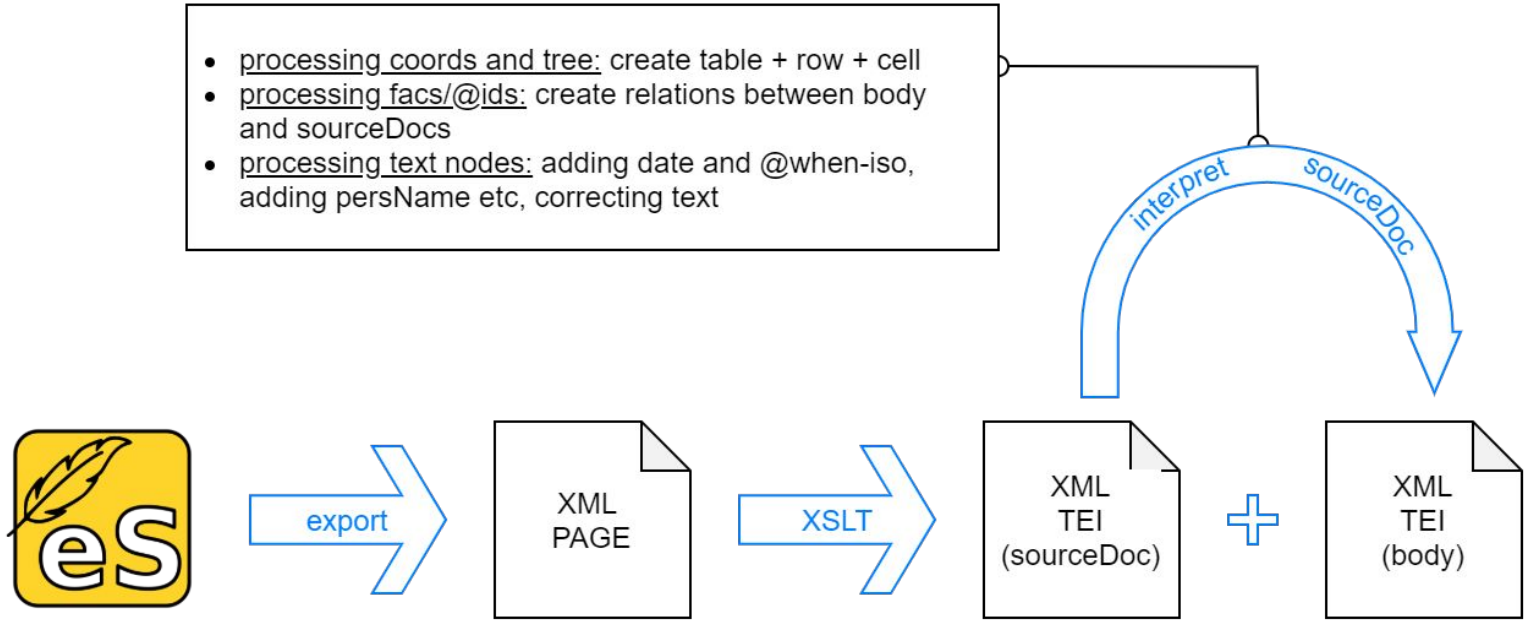
```
<sourceDoc>
   <graphic url="32_c42c1_default.jpg" width="2312px" height="3469px"/>
   <surfaceGrp>
      <surface xml:id="eSc_textblock_dde8b8e9"
               type="structure_{type:numbering;}"
               points="154,59 154,257 392,257 392,59">
         <zone xml:id="eSc_line_106ba71b"
               type="mask"
               points="205,182 199,118 237,118 240,109 251,95 254,95 254,92 254,92 257,92 257,92 257,92 260,92 260,92 260,92 263,92 324,98 332,170 338,251 211,268">
            <path type="baseline" points="207,184 333,171"/>
            <line>1</line>
         </zone>
      </surface>
      <surface xml:id="eSc_textblock_e94aafae"
               type="structure_{type:line_group;}"
               points="1770,370 1995,420 2093,471 2126,644 2013,690 1872,867 1849,1003 1917,1058 524,1058 567,1014 532,855 451,829 462,393 576,372 633,367 1769,367">
         <zone xml:id="eSc_line_c4880d79"
               type="mask"
               points="486,479 480,358 549,364 549,364 552,364 552,364 552,364 555,364 555,364 555,367 590,401 659,401 700,361 700,358 700,358 700,358 703,358 703,358
703,358 706,358 706,358 708,358 708,358 708,358 711,358 711,358 714,358 798,398 891,398 931,364 931,364 934,364 934,364 934,364 937,364 937,364 937,364 940,364 1038,370
1059,361 1079,352 1079,352 1082,352 1082,352 1082,352 1085,352 1085,352 1085,352 1088,352 1088,352 1088,352 1108,361 1116,367 1142,361 1183,355 1183,355 1183,355 1186,355
1296,364 1380,370 1409,364 1484,349 1484,349 1487,349 1629,349 1629,349 1629,349 1632,349 1632,349 1681,367 1681,367 1684,367 1724,346 1724,346 1724,346 1727,346 1727,346
1730,346 1730,346 1730,346 1733,346 1733,346 1733,346 1736,346 1736,346 1736,349 1756,367 1776,390 1961,390 1976,370 1993,465 1987,500 1788,482 1655,503 1655,503 1652,503
1652,503 1522,485 972,505 969,505 969,505 969,505 897,491 784,511 784,511 784,511 781,511 781,511 781,511 590,494 486,514">
            <path type="baseline" points="486,479 1993,465"/>
            <line>Sous le pont Mirabeau coule la Seine.</line>
         </zone>

            ...
```

Keeping the link with IIIF image servers?

```
      <graphic url="https://gallica.bnf.fr/iiif/ark:/12148/btv1b525056707/f33/full/full/0/native.jpg"
source="https://gallica.bnf.fr/iiif/ark:/12148/btv1b525056707/manifest.json" width="2312px"
height="3469px"/>
```

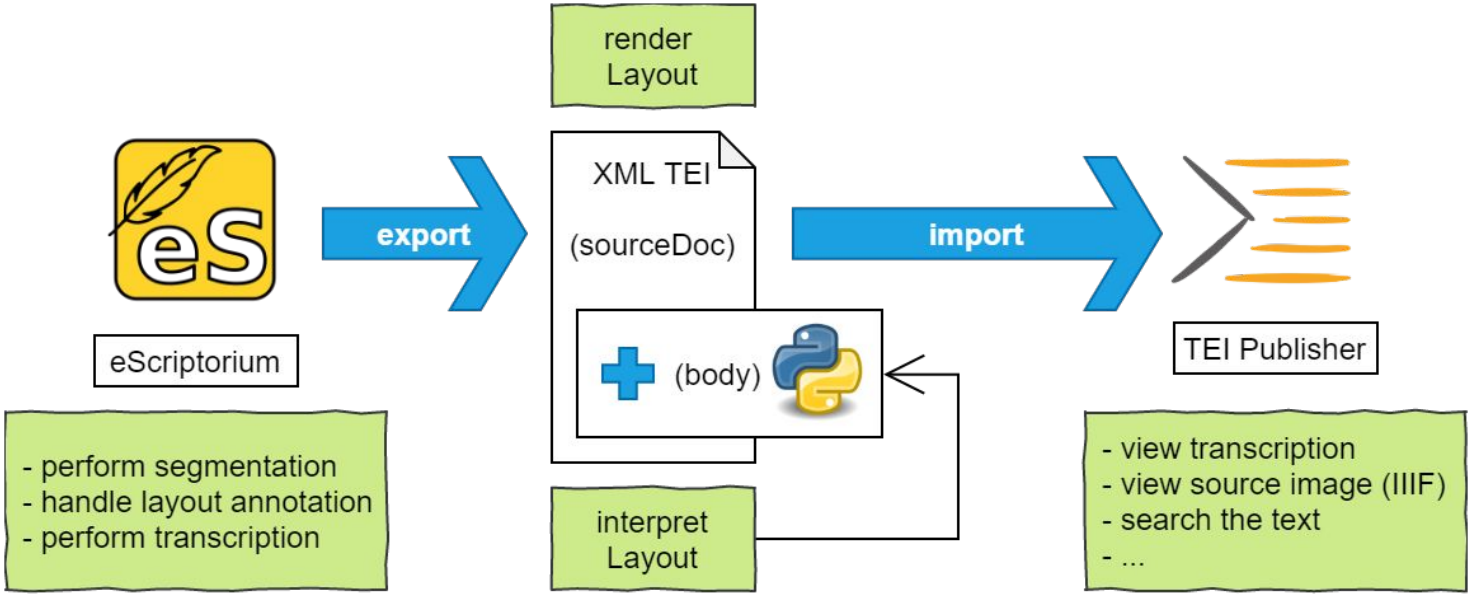
Processing the <sourceDoc>

Linking interpreted content in the <text> elements with the various components available in <sourceDoc>



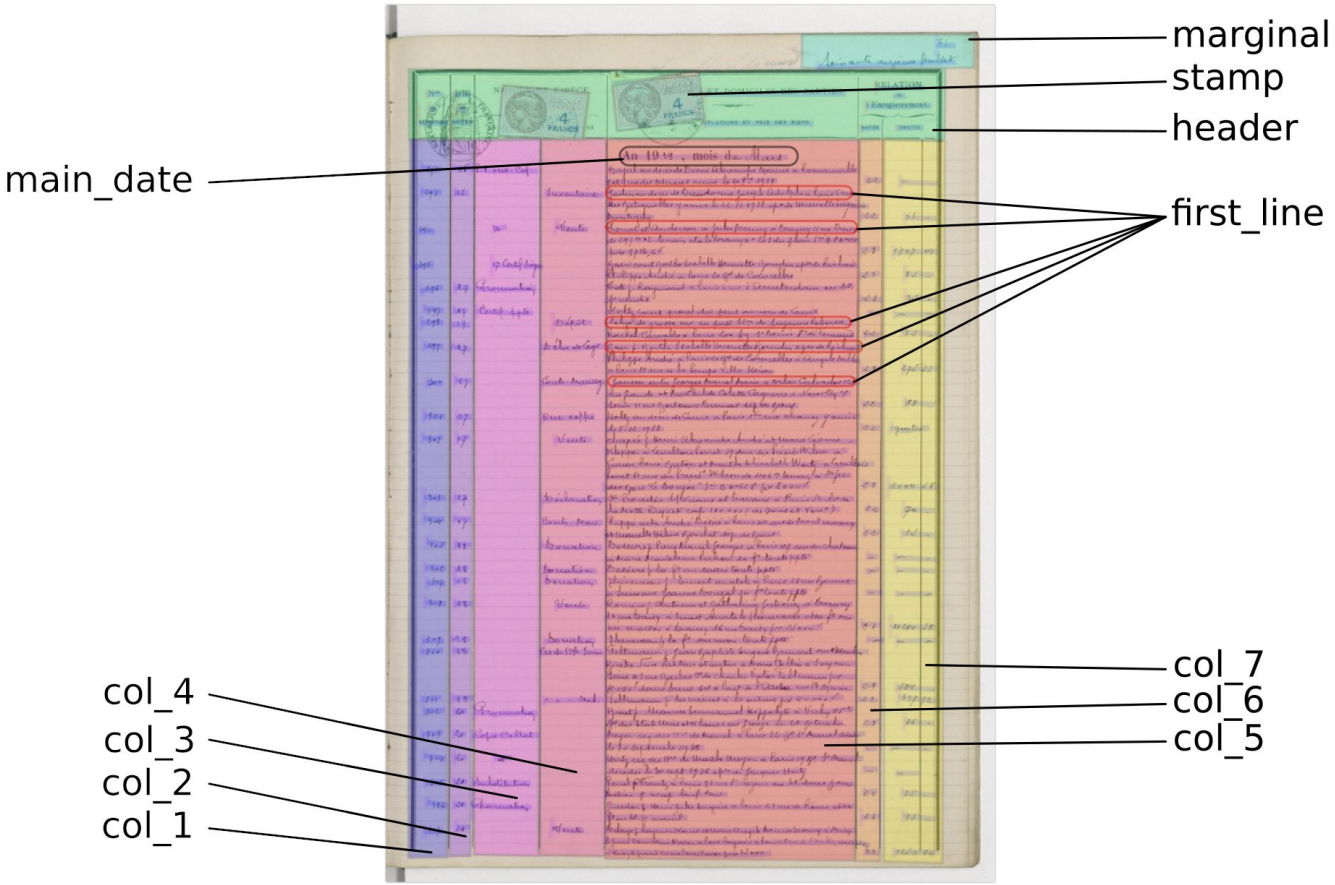
Simplification of the LEPIDEMO workflow

Linking interpreted content in the <text> elements with the various components available in <sourceDoc>

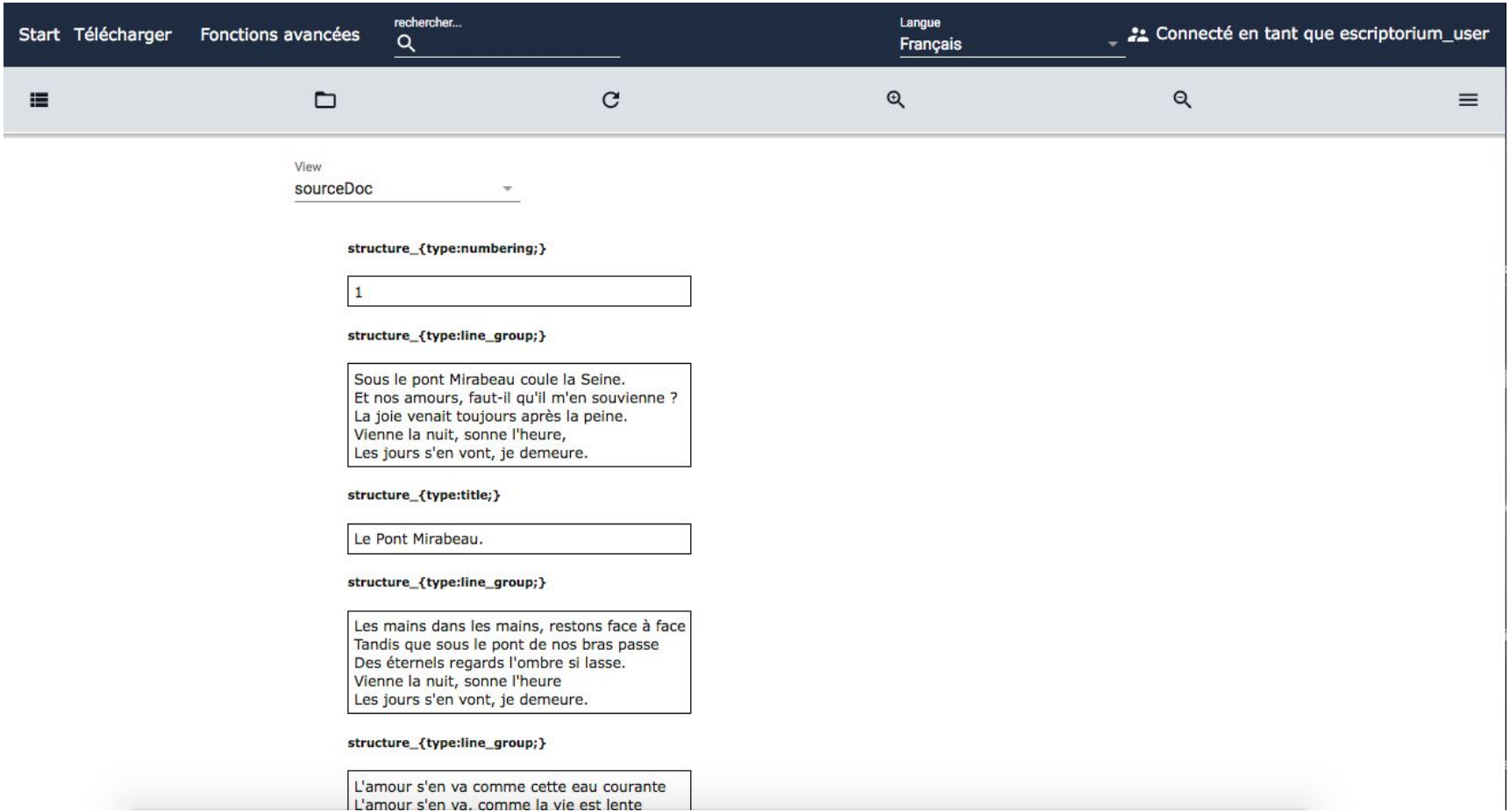


Simplification of the LEPIDEMO workflow

A complex layout finely annotated in eScriptorium



Visualizing the <sourceDoc> and the processed <body> with TEI Publisher



FRAN_0025_3056_L-0.tei

View
sourceDoc

Unable to open [object Object]: HTTP 404 attempting to load TileSource

structure_{type:col_1;}

198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216

structure_{type:col_3;}

Procuration
Autorisation
Certificat de ppté

FRAN_0025_3056_L-0.tei

View
sourceDoc

Unable to open [object Object]: HTTP 404 attempting to load TileSource

structure_{type:col_3;}

Procuration
Autorisation
Certificat de ppté
Procuration
Mainlevée
de signification

structure_{type:col_5;}

An 1914, mois d Fevrier
à Paris 34 rueSedaine
Lamour (par Jn Bte) employé à laCaserne de laGde Républicaine
à Paris rue St- Etienne du Mont en blanc pour recueillir SSon
Boettcher (par Henry Emile) Ingr Conseil à Paris 39 Bd St. Martin
à Léontine Jeanne Blanchot safe dt avec lui pour gérer Comce-
Boileau-Descombres (parlaSté) à Paris 73 rue Lafayette d'inscn. cf
Sophie Marie Louise Dessoliers épse de Edouard Lavelaine de Maubeuge
Gront (par Jean Eugène André) Directr d'usine à Bobigny route des
Petits Ponts à André Valern Fortin indel à Paris 34 rueSedaine
Valter (par Léon) imprimr à Paris rue Grange aux Belles d'ins.
c/ Jean Louis Perréal à Montigny leBretenneux (S. & O.)
Billema (par Lucien) négt dt à Paris rue d'Aumale 13bis// d'insc.
c/ Edouard Eugène Lebourcq & Eugène Pauline Potel 104 r. St Maur
Kopenhague (par Andrée Esther Eugénie) à Neuilly s/ S. 14 rue
St- Pierre à Augustine Marie Bonchot rentière à Paris 41 rue
de Maubeuge de 26000 f - remboursable le 9 février 1920, hyp. s/

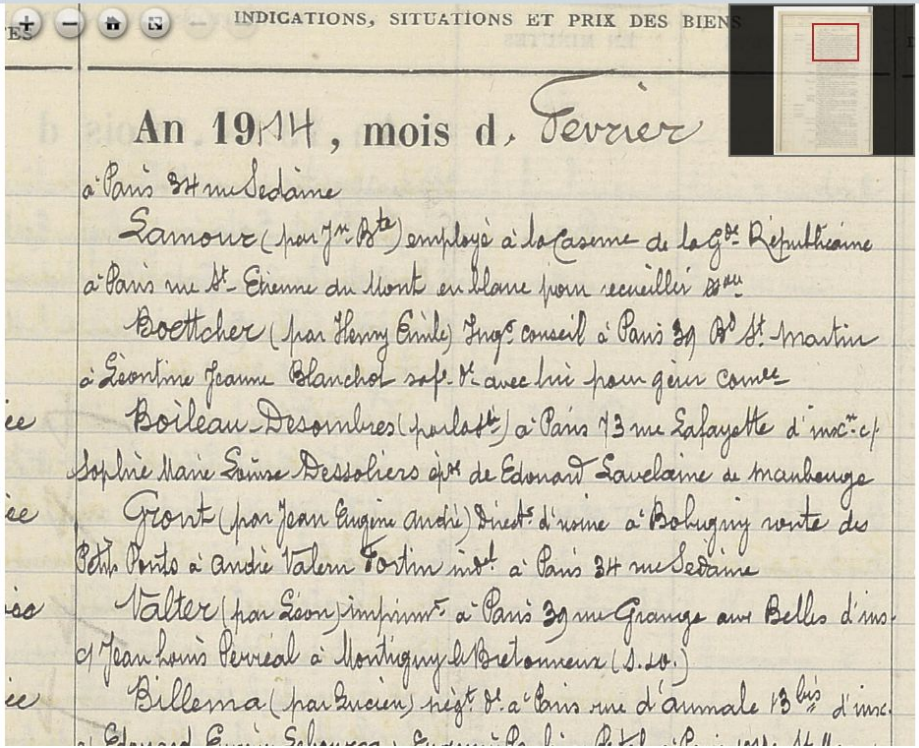
Processing <sourceDoc> to create an interpreted content in <body>

- processing coordinates and sourceDoc
  - create table, row and cell in body
- processing facs and ids:
  - create relations between sourceDoc and body//cell
- processing text nodes:
  - add annotation layer (when-iso to dates, persName, etc)
  - correct the text

```
<row>
 <cell n="1" role="col1">
  <lb facs="#eSc_line_e24d89c4"/> 440 </cell>
 <cell n="2" role="col2">
  <lb facs="#eSc_line_eb98aedd"/>
  <date cert="high" when-iso="1875-05-28,1875-06-28"> 28 </date>
 </cell>
 <cell n="3" role="col3">
  <lb facs="#eSc_line_a42fbb0d"/> Suite du 8 juillet 1869 </cell>
 <cell n="4" role="col4">
  <lb facs="#eSc_line_58d48f3d"/> Dépôt </cell>
 <cell n="5" role="col5">
  <lb facs="#eSc_line_f60606c2" n="1"/> An 1875 mois de Mai
  <lb facs="#eSc_line_3fc0d6ed" n="2"/> Chemins de fer Normands
  (de pièces de publications de délibérations </cell>
 <cell n="6" role="col6">
  <lb facs="#eSc_line_f6960fc1"/>
  <date cert="high" when-iso="1875-05-29,1875-06-29"> 29 </date>
 </cell>
 <cell n="7" role="col7"/>
 <cell n="8" role="misc"/>
</row>
```

FRAN_0025_3056_L-0-tei

Numéros du répertoire	Dates des actes	Actes en brevets	Actes en minutes	Noms, prénms et domiciles des parties ; indication, situations et prix des biens	Date de l'enregistrement	Droits de l'enregistrement	Autres
198	9	Procuration		Lamour (par Jn Bte) employé à laCaserne de laGde Républicaine à Paris rue St-Etienne du Mont en blanc pour recueillir SSon Boettcher (par Henry Emile) Ingr Conseil à	10	3.75	



Conclusion

- All elements from PAGE XML and ALTO files can be mapped to a <sourceDoc> element
- Switching to TEI earlier in the pipeline simplifies the workflow. It ensures that data resulting from OCR/HTR are documented and easily accessible with an edition software such as TEI Publisher.
- We present a proof of concept and a series of rather stable specifications and call for feedback from the community

Thank you for your attention, feedback is welcome and appreciated!

Contacts:

- hugo.scheithauer[at]inria.fr
- alix.chague[at]inria.fr
- laurent.romary[at]inria.fr

Resources:

- https://github.com/lectaurep/page2tei (XSLT PAGEXML to TEI and transformation exemples)
- https://github.com/lectaurep/lepidemo (LECTAUREP project pipeline for transforming PAGE XML files to TEI, and publishing them with TEI Publisher)