Formation à l'utilisation de TEI pour l'édition critique de manuscrits (École nationale des chartes, mars 2011) — Techniques d'analyse et d'alignement avec TEI : aperçu rapide

Florence Clavaud (École nationale des chartes)

1. Introduction

Ces quelques lignes ont uniquement pour objectif de donner un aperçu rapide des dispositifs offerts par TEI pour parvenir à :

- poser sur le texte transcrit une grille d'interprétation
- aligner des segments de texte (autrement dit établir une correspondance entre plusieurs segments de texte) ou un segment de texte et une autre section du fichier TEI

On peut en effet avoir besoin de tels dispositifs pour aller plus loin dans l'analyse d'un texte, par ex. pour isoler (pour les afficher spécifiquement, les indexer et les rendre cherchables à part) des segments du discours narratif ou diplomatique, de reconnaître des registres littéraires ou des thématiques. On peut aussi avoir besoin d'établir des correspondances entre un segment de la transcription d'un manuscrit et une zone de l'image numérique qui le reproduit, ou entre un segment de la transcription et le segment contenant la traduction de cette transcription.

A la base, ces problèmes sont traités avec TEI à l'aide des mêmes éléments ou types d'éléments. Nous allons en présenter quelques-uns par l'exemple. Ces éléments font partie des modules 'analysis' (simple analytic mechanisms) (c'est le cas de `<interp>`, `<interpGrp>`, `<span>` et `<spanGrp>`) et 'linking' (c'est le cas de `<link>`, `<linkGrp>`, `<seg>`, `<anchor>`).

2. Création de grilles d'analyse ou d'interprétation du texte

Un peu comme dans d'autres situations, créer de telles grilles suppose de faire trois choses différentes :

- définir (intellectuellement) puis déclarer (dans le fichier TEI) son vocabulaire d'interprétation
- segmenter le texte, ou encore y poser des ancres qui permettront de borner les segments de texte
- établir la correspondance entre le vocabulaire d'interprétation et les segments du texte

.

2.1. Créer son vocabulaire d'interprétation

L'élément `<interpGrp>` permet de déclarer un tel vocabulaire. Il sera normalement encodé au sein de l'en-tête TEI, dans l'élément `<encodingDesc>`. Bien sûr, on pourra créer autant

de vocabulaires d'interprétation que nécessaire, si l'on a plusieurs points de vue différents sur le texte, si l'étude est transdisciplinaire, etc. Dans ce cas notamment, il sera important d'affecter un attribut **`@type`** à `<interpGrp>`.

`<interpGrp>` contient un ou plusieurs éléments `<interp>` ; chacun de ces éléments `<interp>` contiendra la dénomination, éventuellement la description, d'un des éléments du vocabulaire. Chacun d'eux recevra aussi un identifiant, au moyen d'un attribut **`@xml:id`**.

Exemple (extrait de l'édition en cours d'un corpus de textes juridiques médiévaux) :

```
        <interpGrp type="parties-du-discours"
>
         <interp xml:id="discours1"
>partie 1</interp>
         <interp xml:id="discours2"
>partie 2</interp>
         <interp xml:id="discours3"
>partie 3</interp>
         <interp xml:id="discours4"
>partie 4</interp>
         <interp xml:id="discours5"
>partie 5</interp>
         <!--  etc.  -->
      </interpGrp>
```

2.2. Segmenter le texte

Une fois son vocabulaire défini, il faut isoler dans le texte édité les segments relevant des termes du vocabulaire. Pour cela, on a le choix entre :

- utiliser des éléments contenant le texte, comme `<div>`, `<p>`, `<s>` ou encore `<seg>`, et leur attribuer un identifiant à l'aide de l'attribut **`@xml:id`** ; c'est ce qu'on fait le plus souvent
- utiliser des éléments vides de type milestone, comme `<anchor>`, qui marque un point dans le texte, en leur affectant aussi des identifiants. C'est ce qu'on fait notamment lorsque les portions de texte à isoler se chevauchent - ce qui peut par ex. arriver lorsqu'un même segment intéressant du discours commence au milieu d'un paragraphe et se termine au milieu d'un autre paragraphe, ou lorsqu'on doit poser plusieurs grilles d'interprétation

Exemple 1 (extrait simplifié de la même édition), avec `<seg>` :

```
        <p>
         <seg xml:id="pd001"
>Constitutis in nostra parlamenti curia, Petro Piquelin,
           alias Herbelot, magistro alutanorum seu cordubannarum
           <lb n="2"
/>et subulariorum civitatis seu ville suburbiorum
           et banleuce Aurelianense,
           appellante, ex una parte, et <lb n="3"
/>procuratore carissimi fratris et
           consanguinei nostri ducis Aurelianensi appellati, ac [...]</seg>,
         <seg xml:id="pd002"
> dictus Piquelin proponi fecit, quod in dicta villa
           Aurelianense quam plurimis honoribus, prerogativis
           <lb n="10"
/>et statutis policiam rei publice
           ejusdem civitatis concernentibus decorata,
           unumquodque ministeriorum <lb n="11"
/> dicte ville suum dinoscebatur
           habere magistrum dictorumque cordubannariorum et subulariorum dictus <lb n="12"
/>Piquelin fuerat et erat magister, [...]</seg>
```

```
            <!--  etc.  -->
          </p>
```

Exemple 2 (avec <anchor>) :

```
          <p>
           <anchor xml:id="a001"
/>Constitutis in nostra parlamenti curia, Petro
           Piquelin, alias Herbelot, magistro alutanorum seu cordubannarum
           <lb n="2"
/>et subulariorum civitatis seu ville suburbiorum
           et banleuce Aurelianense,
           appellante, ex una parte, et <lb n="3"
/>procuratore carissimi fratris et
           consanguinei nostri ducis Aurelianensi appellati, ac [...]<anchor xml:id="a002"
/>, <anchor xml:id="a003"
/> dictus Piquelin proponi fecit,
           quod in dicta villa Aurelianense quam plurimis honoribus, prerogativis
           <lb n="10"
/>et statutis policiam rei publice ejusdem civitatis
           concernentibus decorata, unumquodque ministeriorum <lb n="11"
/>
           dicte ville suum
           dinoscebatur habere magistrum dictorumque cordubannariorum et subulariorum
           dictus <lb n="12"
/>Piquelin fuerat et erat magister, [...]<anchor xml:id="a004"
/>
           <!--  etc.  -->
         </p>
```

2.3. Etablir la relation entre les éléments d'un vocabulaire d'analyse et les segments du texte

On pourra utiliser pour ce faire, dans le cas où on a choisi de segmenter le texte avec <seg> ou un élément comparable :

- soit un attribut **@ana**, que l'on ajoutera à l'élément <seg> préalablement posé, et qui renverra à l'élément <interp> concerné ;
- soit, hors du texte, dans un élément <link> lui-même contenu dans <linkGrp>, un attribut **@targets** dans lequel on saisira des pointeurs désignant chacun des éléments du document TEI entre lesquels existe une relation.

Exemple 1 (utilisation de <linkGrp> et <link> pour relier des segments <seg> de texte et des éléments du vocabulaire <interpGrp>) :

```
          <p>
           <seg xml:id="pd001"
>Constitutis in nostra parlamenti curia, Petro Piquelin,
              alias Herbelot, magistro alutanorum seu cordubannarum <lb n="2"
/>et
              subulariorum civitatis seu ville suburbiorum et banleuce Aurelianense,
              appellante, ex una parte, et <lb n="3"
/>procuratore carissimi fratris et
              consanguinei nostri ducis Aurelianensi appellati, ac [...]</seg>, <seg xml:id="pd
> dictus Piquelin proponi fecit, quod in dicta villa
              Aurelianense quam plurimis honoribus, prerogativis <lb n="10"
/>et
              statutis policiam rei publice ejusdem civitatis concernentibus decorata,
              unumquodque ministeriorum <lb n="11"
/> dicte ville suum dinoscebatur
              habere magistrum dictorumque cordubannariorum et subulariorum dictus <lb n="12"
```

```
/>Piquelin fuerat et erat magister, [...]</seg>
        <!--  etc.  -->
     </p>
     <!--  ailleurs dans le fichier TEI  -->
     <linkGrp targFunc="texte parties-du-discours"
>
        <link targets="#pd001 #discours1"
/>
        <link targets="#pd002 #discours4"
/>
        <!--  etc.  -->
     </linkGrp>
```

On aurait pu procéder autrement, en utilisant **@ana**, comme suit :

```
       <p>
        <seg xml:id="pd009"
 ana="#discours1"
>Constitutis in nostra parlamenti curia,
           Petro Piquelin, alias Herbelot, magistro alutanorum seu cordubannarum <lb n="2"
/>et subulariorum civitatis seu ville suburbiorum et banleuce
           Aurelianense, appellante, ex una parte, et <lb n="3"
/>procuratore
           carissimi fratris et consanguinei nostri ducis Aurelianensi appellati, ac
           [...]</seg>, <seg xml:id="pd010"
 ana="#discours4"
> dictus Piquelin
           proponi fecit, quod in dicta villa Aurelianense quam plurimis honoribus,
           prerogativis <lb n="10"
/>et statutis policiam rei publice ejusdem
           civitatis concernentibus decorata, unumquodque ministeriorum <lb n="11"
/>
           dicte ville suum dinoscebatur habere magistrum dictorumque
           cordubannariorum et subulariorum dictus <lb n="12"
/>Piquelin fuerat et
           erat magister, de quibus sui magistratus juribus nullus in ministeriis
           cordubannariatus <lb n="13"
/>et subulariatus, nisi prius per eundem
           magistrum approbatus [...]</seg>
     </p>
```

Dans le cas où on a utilisé des éléments <anchor> pour délimiter des portions de texte (exemple 2), on pourra créer quelque part dans le fichier TEI un élément <spanGrp>, dans lequel on saisira autant d'éléments que nécessaire pour définir (abstraitement) des segments de texte, à l'aide des attributs **@from** et **@to**. Puis on ajoutera à cet élément un attribut **@ana**.

Exemple 2 (utilistion de <spanGrp> pour construire virtuellement des segments de texte à partir d'éléments <anchor> et les relier avec le vocabulaire) :

```
       <p>
        <anchor xml:id="a001"
/>Constitutis in nostra parlamenti curia, Petro
        Piquelin, alias Herbelot, magistro alutanorum seu cordubannarum <lb n="2"
/>et subulariorum civitatis seu ville suburbiorum et banleuce Aurelianense,
        appellante, ex una parte, et <lb n="3"
/>procuratore carissimi fratris et
        consanguinei nostri ducis Aurelianensi appellati, ac [...]<anchor xml:id="a002"
/>, <anchor xml:id="a003"
/> dictus Piquelin proponi fecit,
        quod in dicta villa Aurelianense quam plurimis honoribus, prerogativis <lb n="10"
/>et statutis policiam rei publice ejusdem civitatis concernentibus
        decorata, unumquodque ministeriorum <lb n="11"
/> dicte ville suum
```

```
        dinoscebatur habere magistrum dictorumque cordubannariorum et subulariorum
        dictus <lb n="12"
/>Piquelin fuerat et erat magister, [...]<anchor xml:id="a004"
/>
        <!--  etc.  -->
     </p>
     <!--  ailleurs dans le fichier TEI  -->
     <spanGrp>
        <span from="#a001"
 to="#a002"
 ana="#discours1"
/>
        <span from="#a003"
 to="#a004"
 ana="#discours4"
/>
     </spanGrp>
```

3. Etablissement d’une correspondance entre deux portions de texte

Nous nous contenterons ici de donner un exemple réel, pour montrer comment créer une relation entre deux segments de texte. Un des cas auxquels on pense est celui de la correspondance entre une portion de texte manuscrit écrite dans une langue A, et un texte (dans le même manuscrit, ou plus souvent un texte de l’éditeur) traduisant ce segment de texte dans une langue B.

Exemple adapté de l’édition d’une page de la Bible glosée d’Anselme de Laon (http://theleme.enc.sorbonne.fr/dossiers/notice99.php) :

```
    <div type="transcription"
 xml:lang="lat"
>
     <p>
        <seg xml:id="transcr-001"
>Colosenses</seg>
        <add place="margin-left"
 xml:id="add-001"
>
           <seg xml:id="transcr-add-001"
>Colosenses sunt Asiani quib<ex>us</ex>
                 n<ex>on</ex> ipse Ap<ex>osto</ex>l<ex>u</ex>s
              p<ex>re</ex>dica-</seg>
           <lb/>
           <seg xml:id="transcr-add-002"
>-vit, s<ex>ed</ex> ej<ex>us</ex>
                 discip<ex>u</ex>li Archipp<ex>us</ex>
              <ex>et</ex> Epafras. Archipp<ex>us</ex> v<ex>er</ex>o in eos</seg>
           <!--  etc.  -->
        </add>
        <lb/>
        <seg xml:id="transcr-002"
> et hi sicut Laodicenses sunt Asia-</seg>
        <!--  etc.  -->
     </p>
  </div>
  <div type="traduction"
 xml:lang="fre"
 resp="#FC"
>
     <p>
```

```
        <seg xml:id="trad-001"
>Les Colossiens</seg>
        <add xml:id="add-003"
>
           <seg xml:id="trad-add-001"
>Les Colossiens sont des Asiatiques auxquels
              l'Apôtre lui-même n'a pas prêché,</seg>
           <seg xml:id="trad-add-002"
>mais ses disciples Archippe et Epafras.
              Archippe avait reçu sur eux</seg>
           <!--  etc.  -->
        </add>
        <lb/>
        <seg xml:id="trad-002"
> et ceux-ci comme les habitants de Laodicée sont des
           Asiatiques</seg>
        <!--  etc.  -->
     </p>
  </div>
  <!--  plus loin dans le fichier TEI  -->
  <linkGrp>
     <link targets="#transcr-001 #trad-001"
/>
     <link targets="#transcr-002 #trad-002"
/>
     <link targets="#transcr-add-001 #trad-add-001"
/>
     <link targets="#transcr-add-002 #trad-add-002"
/>
  </linkGrp>
```

4. Etablissement de la correspondance entre un segment de texte et une zone d'image numérique

Encore une présentation par l'exemple (exemple adapté de l'édition d'une page du Didascalicon d'Hugues de Saint-Victor (Paris, Bibliothèque Mazarine, 717, 93v° ; voir http://theleme.enc.sorbonne.fr/dossiers/notice100.php) :

```
  <TEI>
     <teiHeader/>
     <facsimile>
        <surface>
           <graphic url="fax.jpg"
 width="693px"
 height="1100px"
 xml:id="fax"
/>
           <zone xml:id="zone_1"
 ulx="213"
 uly="3"
 lrx="342"
 lry="67"
/>
           <zone xml:id="zone_2"
 ulx="29"
 uly="31"
 lrx="214"
 lry="67"
```

```
/>
           <zone xml:id="zone_3"
 ulx="81"
 uly="65"
 lrx="335"
 lry="91"
/>
           <zone xml:id="zone_4"
 ulx="80"
 uly="89"
 lrx="342"
 lry="117"
/>
        </surface>
     </facsimile>
     <text>
        <body>
           <div type="transcription"
 facs="#fax"
>
              <p>
                 <seg facs="#zone_1"
>De tribus generibus lectorum. </seg>
                 <seg facs="#zone_2"
>Satis ut puto aperte </seg>
                 <seg facs="#zone_3"
>demonstratum e<expan>st</expan> provectis et
                    aliquid </seg>
                 <seg facs="#zone_4"
>amplius de se
                       p<expan>ro</expan>mittentib<expan>us</expan>
                       n<expan>on</expan> idem e<expan>ss</expan>e </seg>
                 <!--  etc.  -->
              </p>
           </div>
        </body>
     </text>
  </TEI>
```

Noter qu’on aurait aussi pu écrire :

```
    <TEI>
     <teiHeader/>
     <facsimile>
        <surface>
           <graphic url="fax.jpg"
 width="693px"
 height="1100px"
 xml:id="fax"
/>
           <zone xml:id="zone_1"
 ulx="213"
 uly="3"
 lrx="342"
 lry="67"
/>
           <zone xml:id="zone_2"
 ulx="29"
 uly="31"
 lrx="214"
 lry="67"
/>
           <zone xml:id="zone_3"
 ulx="81"
 uly="65"
```

```
 lrx="335"
 lry="91"
/>
          <zone xml:id="zone_4"
 ulx="80"
 uly="89"
 lrx="342"
 lry="117"
/>
        </surface>
     </facsimile>
     <text>
        <body>
           <div type="transcription"
 facs="#fax"
>
              <p>
                 <seg xml:id="seg_001"
>De tribus generibus lectorum. </seg>
                 <seg xml:id="seg_002"
>Satis ut puto aperte </seg>
                 <seg xml:id="seg_003"
>demonstratum e<expan>st</expan> provectis et
                    aliquid </seg>
                 <seg xml:id="seg_004"
> amplius de se
                       p<expan>ro</expan>mittentib<expan>us</expan>
                       n<expan>on</expan> idem e<expan>ss</expan>e </seg>
                 <!--  etc.  -->
              </p>
           </div>
           <!--  plus loin  -->
           <linkGrp>
              <link targets="#zone_1 #seg_001"
/>
              <link targets="#zone_2 #seg_002"
/>
              <link targets="#zone_3 #seg_003"
/>
              <link targets="#zone_4 #seg_004"
/>
           </linkGrp>
        </body>
     </text>
  </TEI>
```