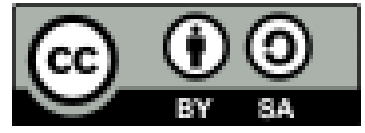
Formation Edition numérique

Les Métadonnées

Simon Gabay



Principes de base

Kézaco?

- Ce sont des données numériques qui servent à représenter ou décrire d'autres données (numériques ou non).
- Elles donnent des informations sur la source, la nature, le contenu, l'histoire, la localisation du document qu'elles décrivent.
- Elles peuvent (doivent?) être standardisées.

Utilité

- Elles fournissent un index qui permettent de faciliter et accélérer les recherches.
- La normalisation permet de simplifier l'échange de données (on parle d'interopérabilité).

Métadonnées et TEI

- Dans un document encodé en TEI, on trouve dans le `<teiHeader>` (cf. TEI) les métadonnées du document.
- Le `<teiHeader>` fournit une description structurée des données contenues dans le document XML.
- Certains éléments sont obligatoires, d'autres sont facultatifs.
- La hiérarchie des données est contrainte par le schéma.

Un document TEI minimal

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
      <fileDesc>
         <titleStmt>
            <title>Title</title>
         </titleStmt>
         <publicationStmt>
            <p>Publication Information</p>
         </publicationStmt>
         <sourceDesc>
            <p>Information about the source</p>
         </sourceDesc>
      </fileDesc>
  </teiHeader>
  <text>
      <body>
         <p>Some text here.</p>
      </body>
  </text>
</TEI>
```

La Sainte Trinité du `<teiHeader>`

1. `<titleStmt>` (cf. TEI) donne le nom du fichier

```
<titleStmt>
  <title>Exercice sur un poème de Lope de Vega</title>
</titleStmt>
```

2. `<publicationStmt>` (cf. TEI) donne des informations concernant la publication (licence, diffuseur...)

```
<publicationStmt>
  <p>Simon Gabay, UniNe. CC-BY.</p>
</publicationStmt>
```

3. `<sourceDesc>` (cf. TEI) Des informations concernant la source

```
<sourceDesc>
  <p>Un poème de Lope de Vega.</p>
</sourceDesc>
```

`<titleStmt>` vs `<sourceDesc>`

- `<titleStmt>` n'est pas le nom de l'œuvre (littéraire) *encodée*, mais de l'édition (critique) *produite*.
- Pensons à certains titres comme *Andromaque, édition de la version de 1668*. Le titre de l'édition contient le titre original, mais pas uniquement.
- `<titleStmt>` et `<sourceDesc>` se recoupent partiellement, mais restent fondamentalement différents.
- Cette distinction prend plus de sens dans le cas d'une monographie (oui, on peut écrire sa thèse en TEI, sans doute même devrait-on…), dont le titre est nécessairement différent des sources.

Pourquoi faire simple si on peut faire compliqué?

On peut encoder des choses très différentes en TEI, ce qui explique certaines incronguités apparentes. Par exemple, pourquoi encoder ainsi:

```
<titleStmt>
  <title>Exercice sur un poème de Lope de Vega</title>
</titleStmt>
```

Et non ainsi:

```
<titleStmt>Exercice sur un poème…</titleStmt>
```

Pourquoi? Parce que

Parce qu'il s'agit de la version minimale de `<titleStmt>`, dans lequel on peut ajouter d'autres informations que le simple `<title>` (cf. TEI).

```
<titleStmt>
  <title>Encodage d'un poème de Lope de Vega</title>
  <author>Lope de Vega</author>
  <editor>
    <persName>
      <forename>Simon</forename>
      <surname>Gabay</surname>
    </persName>
  </editor>
</titleStmt>
```

Notons que `<author>` (cf. TEI) pourrait être lui aussi encodé comme `<editor>` (cf. TEI) avec `<persName>` (cf. TEI), `<forename>` (cf. TEI) et `<surname>` (cf. TEI).

L'encodage emmental

(_N.B._ les français parlent de *gruyère*, mais nous savons bien qu'il n'y a pas de trou dans le gruyère...)

L'encodage en XML-TEI est un encodage qui prévoit des trous que l'on peut remplir par la suite selon nos besoins, et ce de manière simple. Ce n'est pas le cas de tous les langages (<- critique feutrée des informaticiens qui ne comprennent pas que les humanistes utilisent encore le XML).

Des bienfaits de la globalisation (en TEI)

Nous venons de voir apparaître la balise `<persName>`. Cette balise n'est pas propre au `<teiHeader>`, et on peut la retrouver un peu partout dans un document TEI, comme dans la balise `<l>` (cf. TEI) du `<body>` (cf. TEI) que nous avons précédemment vue.

```
<l>Elle a trouué <persName>Pyrrhus</persName>, porté fur
  des Soldats,</l>
```

Idem pour `<title>`, que l'on peut aussi retrouver à différents endroits d'un document TEI, comme `<bibl>` :

```
<bibl>
  <author>Molière</author>
  <title>Le Festin de Pierre</title>
  <editor>Joan DeJean</editor>
  <pubPlace>Genève</pubPlace>
  <publisher>Droz</publisher>
  <date>1999</date>
</bibl>
```

Des limites de la globalisation (même en TEI)

Il n'est cependant pas possible de recycler toutes les balises de la TEI (*e.g.* `<titleStmt>`) et dans le cas où c'est possible on ne peut pas le faire à n'importe quel endroit (*e.g.* `<author>`):

```
<p><author>Victor</author>, est l'auteur de <title>Notre-
  Dame de Paris</title></p>
```

L'élément `<title>` peut être mis dans un `<p>`, mais pas `<author>`.

Des limites de la globalisation (suite)

Parfois il est possible d'utiliser un élément sans pour autant que cela soit souhaitable, comme `<persName>` dans `<bibl>` (cf. TEI):

```
<bibl>
  <persName>Molière</persName>
  <title>Le Festin de Pierre</title>
  <persName>Joan DeJean</persName>
  <pubPlace>Genève</pubPlace>
  <publisher>Droz</publisher>
  <date>1999</date>
</bibl>
```

Même s'il est possible de préciser

```
<bibl>
  <persName type="auteur">Molière</persName>
  <title>Le Festin de Pierre</title>
  <persName type="éditrice">Joan DeJean</persName>
  …
</bibl>
```

Deuxième étage de la fusée

`<fileDesc>` XXL

`<fileDesc>` ne se limite pas à ces trois éléments:

```
<fileDesc>
  <titleStmt>…</titleStmt>
  <publicationStmt>…</publicationStmt>
  <sourceDesc>…</sourceDesc>
</fileDesc>
```

Voici une version plus développée:

```
<teiHeader>
  <fileDesc>
    <titleStmt>…</titleStmt>
    <editionStmt>…</editionStmt>
    <publicationStmt>…</publicationStmt>
    <seriesStmt>…</seriesStmt>
    <noteStmt>…</noteStmt>
    <sourceDesc>…</sourceDesc>
  </fileDesc>
</teiHeader>
```

`<editionStmt>`

`<editionStmt>` (cf. TEI) permet de donner des informations sur l'édition:

- est-ce la première version?
- une révision?

Il est possible de préciser les responsabilités des collaborateurs dans le processur éditorial avec `<resp>` :

```
<editionStmt>
  <edition ref="1">
    <date type="publication" n="1" when="YYYY-MM-DD"/>
  </edition>
  <respStmt>
    <persName ref="#pointeur">Nom</persName>
    <resp>
      <date>YYYY</date>Rôle/ce qui a été fait.
    </resp>
  </respStmt>
</editionStmt>
```

`<publicationStmt>` et `<noteStmt>`

`<publicationStmt>` (cf. TEI) permet de donner des informations sur la publication et la diffusion d'un texte:

- qui publie?
- Avec quels droits?

```
<publicationStmt>
  <authority>Institution</authority>
  <address>
    <addressLine>Ligne d'adresse postale</addressLine>
    <addressLine>Ligne d'adresse postale</addressLine>
  </address>
  <availability status="restricted">
    <licence  target="url">Nom de la licence</licence>
  </availability>
</publicationStmt>
<noteStmt>
  <note>Si je dois préciser quelque chose</note>
</noteStmt>
```

Remarque

On remarque que dans le premier exemple donné au début du cours `<publicationStmt>` contient un `<p>`, absent de l'exemple précédent:

```
<publicationStmt>
  <p>Publication Information</p>
</publicationStmt>
```

La TEI offre ici le choix entre un texte rédigé comme un paragraphe, ou des données encodées de manière plus rigide et systématique. Il faut donc choisir entre les deux méthodes: il est impossible de faire les deux en même temps, et donc de faire suivre `<p>` de `<authority>`, `<address>` ...
Il en va de même pour `<sourceDesc>`.

`<sourceDesc>` (livre)

`<sourceDesc>` (cf. TEI) permet de décrire la source à partir de laquelle un texte électronique a été dérivé ou produit. Son contenu diffère grandement si cette source est

- un livre
- un manuscrit

Pour un livre:

```
<sourceDesc>
  <bibl>
    <author>
      <forename>prénom</forename>
      <surname>nom de famille</surname>
    </author>
    <title>Titre</title>
    <publisher>Editeur(/"publieur")</publisher>
    <pubPlace>Lieu de publication</pubPlace>
    <date when="YYYY-MM-DD">date</date>
  </bibl>
</sourceDesc>
```

`<sourceDesc>` (manuscrit)

`<msDesc>` (cf. TEI) n'est pas aussi générique que `<bibl>`. Il a été développé par les médiévistes pour les manuscrits, mais se trouve désormais utilisé partout où la source est un document unique ou rare (imprimé ancien, épigraphie…).

```
<sourceDesc>
  <msDesc>
    <msIdentifier>
      <country>Pays</country>
      <settlement>Ville</settlement>
      <institution>Bibliothèque</institution>
      <repository>Dépot</repository>
      <collection>Collection</collection>
      <idno type="shelfmark">Cote</idno>
    </msIdentifier>
  </msDesc>
</sourceDesc>
```

Troisième étage de la fusée

Non à l'anarchie! (NON!)

Il est non seulement nécessaire de connaître le nom des balises et l'endroit où il est possible de les utiliser, mais aussi leur ordre d'enchaînement. Si cela est vrai dans tout le document TEI, c'est particulièrement vrai dans le `<teiHeader>`. Ainsi, ce `<teiHeader>` n'est pas valide

```
<teiHeader>
  <fileDesc>
    <sourceDesc>
      <p>Information à propos de la source</p>
    </sourceDesc>
    <titleStmt>
      <title>Titre</title>
    </titleStmt>
    <publicationStmt>
      <p>Informations relatives à la publication</p>
    </publicationStmt>
  </fileDesc>
</teiHeader>
```

Bigger than big

La structure

```
<teiHeader>
  <fileDesc>
    <sourceDesc>…</sourceDesc>
    <titleStmt>…</titleStmt>
    <publicationStmt>…</publicationStmt>
  </fileDesc>
</teiHeader>
```

signifie que `<fileDesc>` n'est pas la seule balise possible dans `<teiHeader>`. Peuvent ainsi être ajoutées (***dans cet ordre***):

```
<teiHeader>
  <fileDesc>…</fileDesc>
  <encodingDesc>…</encodingDesc>
  <profileDesc>…</profileDesc>
  <revisionDesc>…</revisionDesc>
</teiHeader>
```

`<encodingDesc>`

`<encodingDesc>` (cf. TEI) documente la relation d'un texte électronique avec sa ou ses sources. Il permet notamment de donner des informations sur l'encodage.

```
<encodingDesc>
  <projectDesc>
    <p>Description du projet</p>
  </projectDesc>
  <editorialDecl>
    <correction>
      <p>Le texte a-t-il subit des corrections?</p>
    </correction>
    <hyphenation>
      <p>Quid des césures/tirets de fin de ligne?</p>
    </hyphenation>
    <normalization>
      <p>Normalisation graphique ou non?</p>
    </normalization>
  </editorialDecl>
</encodingDesc>
```

`<profileDesc>`

`<profileDesc>` (cf. TEI) fournit une description détaillée des aspects non bibliographiques du texte: création, langues utilisées, mots-clefs, noms des personnes ou de lieux mentionnés...

```
<profileDesc>
  <creation>
    <date type="type" when="YYYY-MM-DD"/>
  </creation>
  <textClass>
    <keywords scheme="Nom_du_référentiel">
      <list>
        <item>Un mot clef</item>
        <item>Un autre</item>
      </list>
    </keywords>
    <languageUsage>
      <language ident="code_ISO">Langue</language>
    </languageUsage>
  </textClass>
</profileDesc>
```

`<revisionDesc>`

`<revisionDesc>` (cf. TEI) fournit un résumé de l'historique des révisions d'un fichier. Qui a changé quoi et à quelle date?

```
<revisionDesc>
  <change when="YYYY-MM-DD">
    <persName>nom</persName>
    Ce qui a été changé
  </change>
</revisionDesc>
```