L'en-tête TEI : l'importance des métadonnées

Lou Burnard et Florence Clavaud

novembre 2017

L'importance des métadonnées

Les métadonnées – les données sur d'autres données – servent à identifier, retrouver, utiliser et gérer, préserver les ressources numériques.

Quelques standards de métadonnées :

DCMI: Dublin Core Metadata Initiative Système très simple pour décrire les ressources : 15 éléments de données

RDF: Resource Description Framework Standard W3C pour la représentation d'informations concernant n'importe quelle ressource et des relations entre ressources

EAD: Encoded Archival Description Standard international pour la description des fonds d'archives

METS: Metadata Encoding and Transmission Standard Standard international pour la description des ressources, focalisé sur les aspects administratifs, la structuration physique, etc.

L'en-tête TEI

Tout document TEI doit avoir un en-tête, qui est utilisé pour stocker deux types de métadonnées :

- celles qui serviront pour identifier et décrire le fichier, comme on le ferait d'une ressource électronique dans un catalogue de bibliothèque (mentions de titre et de responsabilité, mention d'édition et de collection, adresse bibliographique, collation, etc.)
- celles qui serviront plus globalement à l'utilisateur du fichier et lui permettront de comprendre comment le texte a été encodé (description de la source éventuelle du fichier, présentation du projet, des règles éditoriales, des propriétés des composants du texte, etc.)

A quoi servent les métadonnées ?

Pour le bibliothécaire (ou le gestionnaire de ressources électroniques), il faut, entre autres...

- identifier d'une manière définitive la ressource
- documenter ses composants, ses supports, son organisation
- déclarer ses propriétés juridiques (droits d'auteur etc.)

Pour l'utilisateur, il faut, entre autres...

- résumer sa structure logique
- spécifier les utilisations prévues voire possibles
- décrire son schéma analytique ("codebook") s'il y en a
- résumer ses propriétés et son contenu pour les moteurs de recherche

L'entete TEI est conçu pour repondre a ces deux besoins

L'élément <teiHeader>

Inspiré de la pratique AACR2, il contient quatre éléments principaux: :

1. <fileDesc> : fournit une description bibliographique complète du document TEI et de ses sources
2. <encodingDesc> : documente les rapports entre le document et la source (ou les sources) dont il dérive (contexte général et motivations, règles éditoriales...)
3. <profileDesc> : fournit des informations supplémentaires (non bibliographiques) sur le fichier, telles que les langues utilisées, les modalités de production du fichier, les participants, les thèmes...
4. <revisionDesc>: fournit l'historique des modifications du fichier.

<fileDesc> est obligatoire, tous les autres éléments sont optionnels.

L'en-tête TEI minimal

```
<teiHeader>
 <fileDesc>
  <titleStmt>
   <title>Mon titre</title>
  </titleStmt>
  <publicationStmt>
   <p>Mon agence de distribution</p>
  </publicationStmt>
  <sourceDesc>
   <p>Ma provenance</p>
  </sourceDesc>
 </fileDesc>
</teiHeader>
```

Contenu des éléments constitutifs de l'en-tête

Les informations peuvent être fournies :

- en prose libre (des séries de paragraphes ou du texte directement)
- en utilisant des éléments spécialisés et bien structurés

Par convention,

- les éléments dont le nom se termine par "Decl" (comme refsDecl) donnent des informations sur certaines pratiques d'encodage appliquées au texte contenu dans le fichier
- les éléments dont le nom se termine par "Desc" (comme msDesc ou settingDesc) décrivent certaines caractéristiques du fichier, soit en prose, soit à l'aide de sous-éléments

La TEI est très libérale ...

La déscription du document (<fileDesc>)

- Éléments obligatoires :
  - <titleStmt>: le titre de la ressource électronique et la ou les mentions du ou des responsables de sa création
  - <sourceDesc>: renseignements sur la ou les sources dont dérive le fichier
  - <publicationStmt>: précise les modalités de diffusion de la ressource
- Éléments facultatifs :
  - <editionStmt> : la mention d'édition
  - <extent> : la taille du fichier, le nombre de supports de stockage, le nombre de fichiers (si le document TEI est composé de plusieurs fichiers)
  - <seriesStmt> : la mention de collection, si la ressource fait partie d'une série d'éditions électroniques
  - <notesStmt>: des notes complémentaires

Identification de la ressource

Une ressource peut avoir plusieurs titres (ou aucun) :

```
<title>Artamène</title>
<title type="alt">Le Grand Cyrus</title>
<title type="sub">Edition numérisée</title>
<title type="generic">Feuille de manuscrit</title>
```

On peut nommer plusieurs responsables:

```
<author>Scudéry, Madeleine de</author>
<principal>Gefen, Alexandre</principal>
<funder>Fonds National Suisse de la Recherche
Scientifique</funder>
<respStmt>
 <resp>transcription</resp>
 <orgName>SEPE, IRHT, Orléans</orgName>
</respStmt>
```

L'élément <editionStmt>

Cet élément est utilisé lorsque la ressource a été édité sous forme numérique plus qu'une fois et sert à distinguer de quelle édition il s'agit.

```
<editionStmt>
  <edition>Deuxième edition</edition>
</editionStmt>
```

Distribution de la ressource (<publicationStmt>)

L'agence responsable de la mise à disposition d'une ressource peut être considéré comme un <publisher>, un <distributor>, ou un <authority> (ou plusieurs d'entre eux)

Avec chaque agence on peut associer un <pubPlace>, <address>, <availability>, <idno>, et/ou <date>.

(L'ordre est signifiant)

```
<publicationStmt>
 <date>2012</date>
 <publisher>École nationale des chartes</publisher>
 <address>
  <addrLine>19, rue de la Sorbonne</addrLine>
  <addrLine>75005 Paris</addrLine>
 </address>
 <availability>
  <licence target="http://creativecommons.org/licenses/by/3.0/deed.fr">
      Creative Commons Attribution 3.0 non transposé (CC BY 3.0) </licence>
 </availability>
 <distributor>Lou Burnard Consulting</distributor>
</publicationStmt>
```

L'élément <seriesStmt>

Cet élément est utilisé lorsque la ressource fait partie d'un ensemble partageant un même titre (collection), ou bien constitue un ou plusieurs volumes d'un item, ou encore est un numéro distinct d'une publication en série (périodique).

```
<seriesStmt>
 <title>Éditions en ligne de l'École des chartes</title>
 <idno type="URI">http://elec.enc.sorbonne.fr</idno>
 <idno type="vol">3</idno>
</seriesStmt>
```

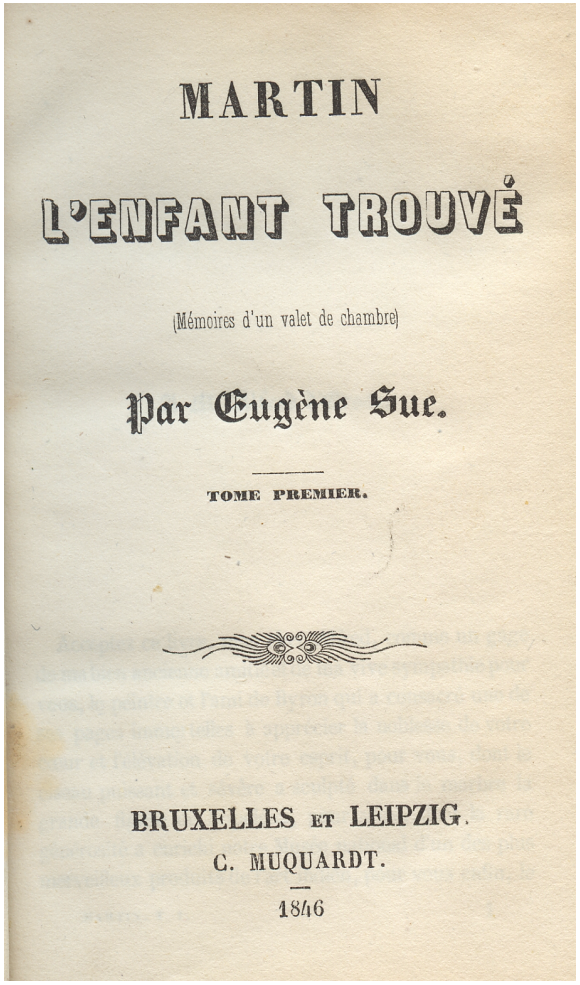
La description des sources (<sourceDesc>)

La plupart des textes encodés en TEI n'ont pas été créés sous forme numérique... il faut donc décrire leurs sources.

La TEI offre plusieurs solutions pour ce faire, plus ou moins structurées :

- déscription en prose dans un élément <p>
- une référence bibliographique dans <bibl>, <biblStruct>, ou <biblFull>
- une déscription structurée pour les transcriptions de manuscrit
- une déscription structurée pour les transcriptions de discours oraux
- une liste de références bibliographiques comme ci-dessus, dans <listBibl>

Source imprimée



MARTIN

L'ENFANT TROUVÉ

(Mémoires d'un valet de chambre)

Par Eugène Sue.

TOME PREMIER.

BRUXELLES ET LEIPZIG.

C. MUQUARDT.

1846

```
<sourceDesc>
  <bibl xml:lang="fr">
    <author>Sue, Eugène</author>
    <title>Martin, l'enfant
trouvé</title>
    <title type="sub">(Mémoires d'un
valet de chambre)</title>
    <imprint>
      <pubPlace>Bruxelles et
Leipzig</pubPlace>
      <publisher>C. Muquardt</publisher>
      <date>1846</date>
    </imprint>
  </bibl>
</sourceDesc>
```

Encore une source imprimée

```
<bibl type="book" subtype="monograph"
  xml:id="brief_discours_1614">
  <title level="m">Brief Discours pour la reformation des mariages</title>.
<pubPlace>Paris</pubPlace>, de l'imprimerie d'<publisher>Anthoine du
    Brueil</publisher>, rue Saint-Jacques, au dessus de Saint-Benoist, à la
Couronne,
<date when="1614">1614</date>, <biblScope type="pp">pp 3-16</biblScope> dans
<title level="m">Variétés Historiques et Littéraires. Recueil de pièces
volantes rares et
    curieuses en prose et en vers</title>, Revues et annotés par M.
<editor>
   <name>
    <forename>Édouard</forename>
    <surname>Fournier</surname>
   </name>
  </editor>, <biblScope type="vol">Tome
    IV</biblScope>. A <pubPlace>Paris</pubPlace>, Chez <publisher>P.
Jannet</publisher>.
<date when="1856">MDCCCLVI</date>.
</bibl>
```

Source manuscrite

```
<sourceDesc>
  <msDesc>
    <msIdentifier>
      <country>France</country>
      <settlement>Paris</settlement>
      <repository>Archives nationales</repository>
      <collection>Commerce et Industrie</collection>
      <idno>F/12/5080</idno>
    </msIdentifier>
    <msContents>
      <p>Minute d'un rapport de proposition à la Légion d'honneur fait, en 1850, par le
            ministre du Commerce et de l'Agriculture et président de la Société de
géographie,
            Jean-Baptiste Dumas, au Président de la République, en faveur des frères
            d'Abbadie, Antoine (1810-1897) et Arnaud (1815-1893), auteurs d'un voyage en
            Abyssinie.</p>
    </msContents>
    <physDesc>
      <p>Deux feuilles de papier 24 x 12 cm ; écriture à l'encre noire.</p>
      <handDesc>
        <handNote  xml:id="AA"
          scope="major">Antoine d'Abbadie</handNote>
        <handNote  xml:id="DJB"
          scope="minor">Jean-Baptiste Dumas</handNote>
        <handNote  xml:id="EPR"
          scope="minor">membre inconnu du cabinet du
                ministre</handNote>
      </handDesc>
    </physDesc>
  </msDesc>
</sourceDesc>
```

Nous revenons sur cela plus tard...

Source orale

```
<sourceDesc>
  <recordingStmt>
    <recording type="audio" dur="P30M">
      <respStmt>
        <resp>Location recording by</resp>
        <name>Sound Services Ltd.</name>
      </respStmt>
      <equipment>
        <p>Multiple close microphones mixed down to stereo Digital Audio Tape,
standard
              play, 44.1 KHz sampling frequency</p>
      </equipment>
      <date>12 Jan 1987</date>
    </recording>
  </recordingStmt>
</sourceDesc>
```

```
<sourceDesc>
  <recordingStmt>
    <recording type="video"
      when="1989-06-24" dur="P60M">
      <p>
        <title>24 Heures</title>: émission télévisée <date>24
juin 1989</date>
      </p>
    </recording>
  </recordingStmt>
</sourceDesc>
```

Source née numérique

```
<sourceDesc>
 <bibl>
  <title>Manifeste des Digital humanities</title>
  <author>Marin Dacos</author>
  <idno type="URL"> http://tcp.hypotheses.org/318</idno>
  <date when="2010-05-21"/>
 </bibl>
</sourceDesc>
```

```
<sourceDesc>
 <p>Aucune source: ce document est né numérique</p>
</sourceDesc>
```

Description de l'encodage (<encodingDesc>)

Cet element facultatif regroupe des informations sur les méthodes qui ont régi la création du texte numérisé, soit en texte libre, soit en utilisant des éléments spécifiques, tels que :

- <projectDesc> : les objectifs du projet
- <samplingDecl> : critères et méthodes de sélection du texte
- <editorialDecl> : informations sur les principes éditoriaux, par ex <correction>, <normalization>, <quotation>, <hyphenation>, <segmentation>, <interpretation> ...
- <classDecl> : les systèmes de classification utilisés
- <tagsDecl> : règles spécifiques applicables à certains éléments

<encodingDesc> exemple simple

```
<encodingDesc>
 <projectDesc>
  <p>Édition électronique lancée en avril 2010 dans le
cadre du mémoire de recherche
      de 2<hi rend="super">e</hi> année du Master "Nouvelles
technologies appliquées
      à l'histoire" à l'École nationale des chartes de
Natalia Pashkeeva.</p>
  <p>Globalement, le présent projet vise l'édition
électronique des carnets de prison
      et d'exil d'Henri Delescluze ainsi que sa
correspondance écrite associée aux dits
      carnets conservés aux Archives nationales dans le
fonds de Charles et d'Henri
      Delescluze sous le cote 494 AP/1.</p>
 </projectDesc>
 <samplingDecl>
  <p>Seuls les carnets 1 et 2 sont édités. On leur a
adjoint la correspondance active
      et passive de Henri Delescluze pendant la période de
rédaction des 2 carnets.</p>
 </samplingDecl>
</encodingDesc>
```

Descriptions plus détaillées de l'encodage

Des balises plus formalisées sont également disponibles:

- <charDecl> : déclaration des glyphes ou caractères non-UNICODE, a référencer dans le texte par l'élément <g>
- <classDecl>: déclaration structurée du système de classification des textes d'un corpus, ou de schéma analytique, à référencer dans le texte par *@ana* ou *@decls*
- <refsDecl> ou <tagsDecl>: déclarations structurées du système de référence (p.e. `I.2.ii`) par rapport avec la structuration XML, et de l'usage (fréquence etc.) des balises XML dans le document même
- <geoDecl>, <metDecl>, <fsdDecl>, <variantEncoding> : fournissent des informations utiles pour comprendre et exploiter l'encodage de la géolocalisation, des analyses métriques ou linguistiques, et de la variation textuelle.
- etc.

On peut définir des caractères non-Unicode

```
<charDecl>
 <glyph xml:id="z103">
  <glyphName>LATIN LETTER Z WITH TWO STROKES</glyphName>
  <mapping type="standardized">Z</mapping>
  <mapping type="PUA">U+E304</mapping>
 </glyph>
</charDecl>
```

Dans une transcription, on peut encoder des caractères non-Unicode avec l'élément <g>:

```
<p> ... mulct<g ref="#z103">z</g> ... </p>
```

```
<charDecl>
 <glyph xml:id="PN">
  <desc>le paraphe de Philippe de Noailles</desc>
  <graphic url="paraphe.jpg"/>
 </glyph>
</charDecl>
<!-- et dans le document --> ... oui <g ref="#PN"/> ...
<!-- ou --> ... oui <g ref="#PN">#</g> ...
```

On peut fournir une taxinomie ”maison”

```
<encodingDesc>
  <classDecl>
    <taxonomy xml:id="types-documents">
      <category xml:id="typedoc-001">
        <catDesc>bon du roi</catDesc>
      </category>
      <category xml:id="typedoc-002">
        <catDesc>acte royal</catDesc>
      </category>
    </taxonomy>
    <taxonomy xml:id="topiques">
      <category xml:id="politique-domestique">
        <catDesc>évenements regionales et
                nationales</catDesc>
      </category>
      <category xml:id="politique-etranger">
        <catDesc>évenements à l'étranger</catDesc>
      </category>
      <category xml:id="socio-femmes">
        <catDesc>discussion des
                femmes</catDesc>
      </category>
      <category xml:id="socio-domestiques">
        <catDesc>discussion des
                domestiques</catDesc>
      </category>
<!-- etc -->
    </taxonomy>
  </classDecl>
<!-- et dans le profileDesc on dira ...-->
  <catRef target="#socio-femmes #typedoc-001"/>
</encodingDesc>
```

On peut definir des styles identifiés dans une source

```
<tagsDecl>
<!-- On se sert de CSS pour definir la mise en italique -->
  <rendition  xml:id="IT"  scheme="css">font-style: italic</rendition>
<!-- Definition d'une police -->
  <rendition  xml:id="FontRoman"
    scheme="css">font-family: serif</rendition>
<!-- Par defaut, les elements emph et hi sont en italiques -->
  <namespace  name="http://www.tei-c.org/ns/1.0">
    <tagUsage  gi="emph"  render="#IT"/>
    <tagUsage  gi="hi"  render="#IT"/>
<!-- par defaut l'element text se sert de la police FontRoman -->
    <tagUsage  gi="text"
      render="#FontRoman"/>
  </namespace>
</tagsDecl>
<!-- ... -->
<text>
  <body>
    <div>
      <p  rendition="#IT">
<!-- Cette para se sert de la police FontRoman, en italique-->
      </p>
      <p>
<!-- Cette para se sert de la meme police mais n'est pas en italique -->
      </p>
    </div>
  </body>
</text>
```

Description du profil (<profileDesc>)

Description détaillée des aspects non bibliographiques du texte, notamment les langues utilisées et leurs variantes, les circonstances de sa production, les parties prenantes et leur environnement par ex :

- <creation>: informations sur la création de la ressource, comme le lieu, la date
- <langUsage>: informations sur les langues, les registres, les dialectes etc. employés
- <textDesc> et <textClass> : classement(s) thématique ou typologique de la ressource selon une classification interne ou externe
- <particDesc> : informations sur les 'participants' d'une interaction linguistique, comme les locuteurs d'un discours oral, les caractères d'un roman
- <settingDesc> : informations sur l'endroit d'une interaction linguistique comme le lieu d'enregistrement d'un discours o la scène d'un drame.

Création

Au plus simple ne contient que des notes informelles sur la genèse d'un texte ou document, par exemple:

```
<creation> Première version finie en
<date value="1929-08">Août 1929</date> à Taos, Nouveau
Mexique</creation>
```

Une structuration plus complexe, adaptée aux besoins des éditions génétiques est aussi possible:

```
<creation>
 <listChange>
  <change xml:id="draft-0">notes d'auteur reunis dans
carnet rouge</change>
  <change xml:id="draft-1">partie tapuscrite de la premiere
version
      complete</change>
  <change xml:id="draft-1-n">annotations d'auteur sur le
tapuscrit</change>
 </listChange>
</creation>
```

Spécification des langues

Il faut spécifier la ou les langue(s) du texte en se servant des codes ISO.
L'élément <language> (et son attribut associé, `xml:lang`) peut comprendre un langage, son écriture, et sa région.

```
<langUsage>
 <language ident="oci">occitan (langue principale du
manuscrit)</language>
 <language ident="lat">latin (langue utilisée dans le
manuscrit)</language>
 <language ident="fre">français contemporain (langue de
l'édition)</language>
 <language ident="frm">français moyen méridional (langue
utilisée dans le
    manuscrit)</language>
</langUsage>
```

Le format des codes d'identifiants des langues est standardisé par l'IETF: BCP 47

Classification des textes

<textClass> fournit une classification (par sujet, medium, type...) pour un texte entier donné. Plusieurs méthodes sont disponibles :

avec <catRef> pour faire une référence directe à une catégorie définie localement (dans <classDecl>, voir plus haut)

avec <classCode> pour faire référence à un système descriptif faisant l'objet d'un consensus et défini à l'externe

avec <keywords> pour déclarer des mots-clés pris dans un vocabulaire bibliographique contrôlé ou dans un nuage de mots

Exemple

```
<profileDesc>
 <creation>
  <date when="1962"/>
 </creation>
 <textClass>
  <catRef target="#WRI #ALLTIM1 #ALLAVA2 #ALLTYP3 #WRIDOM5
#WRILEV2 #WRIMED1 #WRIPP5 #WRISAM3 #WRISTA2 #WRITAS0"/>
  <classCode scheme="DLEE">W nonAc: humanities
arts</classCode>
  <keywords scheme="COPAC">
   <term>History, Modern - 19th century</term>
   <term>Capitalism - History - 19th century</term>
   <term>World, 1848-1875</term>
  </keywords>
 </textClass>
</profileDesc>
```

Cette classification s'applique au texte entier. On pourra aussi l'utiliser pour catégoriser une partie du texte, par ex. une division , à l'aide de l'attribut *@decls* de cette division, qui aura pour valeur un pointeur vers un des éléments de classification.

<textDesc>

<textDesc> fournit une description précise de la situation dans laquelle un texte a été produit, et le caractérise d'une manière relativement indépendante de toute théorie.

```
<textDesc n="novel">
 <channel mode="w">print; part issues</channel>
 <constitution type="single"/>
 <derivation type="original"/>
 <domain type="art"/>
 <factuality type="fiction"/>
 <interaction type="none"/>
 <preparedness type="prepared"/>
 <purpose type="entertain"
  degree="high"/>
 <purpose type="inform"
  degree="medium"/>
</textDesc>
```

<settingDesc>

Cet élément décrit le(s) contexte(s) dans lesquels se situe une interaction linguistique, soit sous la forme d'une description en prose, soit sous celle d'une série d'éléments décrivant le contexte.

```
<settingDesc>
 <p>Pierre Mendès France, Entretiens avec Jean Lacouture
(1980-1981 )</p>
</settingDesc>
```

```
<settingDesc>
 <setting>
  <name>Paris, France</name>
  <time>Fin 19e</time>
 </setting>
</settingDesc>
```

<particDesc>

Cet élément décrit les locuteurs ou autres participants à la production d'un texte.

```
<particDesc>
  <listPerson>
    <person xml:id="P-1234" sex="2"
      age="mid">
      <p>Female informant, well-educated, born in Shropshire UK, 12 Jan 1950,
of unknown
          occupation. Speaks French fluently. Socio-Economic status B2.</p>
    </person>
    <person xml:id="P-4332" sex="1">
      <persName>
        <surname>Hancock</surname>
        <forename>Antony</forename>
        <forename>Aloysius</forename>
        <forename>St John</forename>
      </persName>
      <residence notAfter="1959">
        <address>
          <street>Railway Cuttings</street>
          <settlement>East Cheam</settlement>
        </address>
      </residence>
      <occupation>comedian</occupation>
    </person>
    <listRelation>
      <relation type="personal"
        name="spouse" mutual="#P-1234 #P-4332"/>
    </listRelation>
  </listPerson>
</particDesc>
```

Description des révisions

Et finalement, on utilisera un élément <revisionDesc> pour fournir une liste des modifications apportées à une ressource. Il contient une série d'éléments <change> dans l'ordre chronologique inverse des révisions signifiantes du texte.

```
<revisionDesc>
 <change>entièrement revisé pour Mutec</change>
 <change>en route vers Montréal, aout 2012</change>
 <change>addition d'entête par LB</change>
</revisionDesc>
```

Les attributs @*when* ou @*who* peuvent être utilisés pour préciser les informations:

```
<revisionDesc>
 <change when="2015-06" who="#LB">Publication</change>
 <change when="2012-10">Entièrement revisé pour
Mutec</change>
 <change when="2011-09" who="#OGJ #FC">Plusieurs
corrections</change>
</revisionDesc>
```