L'en-tête TEI : des métadonnées pour le fichier TEI

Florence Clavaud

février 2016

TEI et les métadonnées

Les métadonnées peuvent être définies comme des données sur d'autres données ou sources d'information. Elles servent à identifier, retrouver, utiliser et gérer, préserver ces informations.

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 de n'importe quelle ressource à l'aide de concepts orientés objet

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 numériques, focalisé sur les aspects administratifs, la structuration physique, etc.

Tout fichier 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.)

L'en-tête TEI (élément <teiHeader> contient quatre éléments principaux :

1. <fileDesc> : fournit une description bibliographique complète du fichier et de ses sources
2. <encodingDesc> : documente les rapports entre le fichier 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.

Voici un en-tête TEI minimal :

```
<teiHeader>
 <fileDesc>
  <titleStmt>
   <title>Titre de la ressource ?</title>
  </titleStmt>
  <publicationStmt>
   <p>Qui la diffuse ?</p>
  </publicationStmt>
  <sourceDesc>
   <p>De quelle autre ressource dérive-t-elle ?</p>
  </sourceDesc>
 </fileDesc>
</teiHeader>
```

Les éléments constitutifs de l'en-tête peuvent être :

- en prose libre (des séries de paragraphes ou du texte directement)
- des éléments englobant des informations relatives au même sujet et fortement structurées
- des déclarations (éléments dont le nom se termine par "Decl" (comme refsDecl) donnant des informations sur certaines pratiques d'encodage appliquées au texte contenu dans le fichier
- des descriptions (éléments dont le nom se termine par "Desc" (comme msDesc ou settingDesc) décrivant certaines caractéristiques du fichier, soit en prose, soit à l'aide de sous-éléments

La description du fichier (<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

La mention de titre et de responsabilités (<titleStmt>)

Une ressource électronique a au moins un titre, et une ou plusieurs mentions de responsabilité, stockées soit dans des éléments spécifiques soit dans l'élément générique <respStmt> répétable :

```
<title>Manuscrits d'André-Marie Ampère</title>
<title type="subtitle">Une édition numérique</title>
<author>André-Marie Ampère</author>
<principal>Christine Blondel</principal>
<principal>Marco Segala</principal>
<funder>ANR Corpus 2012</funder>
<respStmt>
 <name>Delphine Usal, <affiliation>Centre Alexandre
Koyré</affiliation>
 </name>
 <resp>Informatique</resp>
</respStmt>
<respStmt>
 <name>Aude Coustumer, <affiliation>Centre Alexandre
Koyré</affiliation>
 </name>
 <resp>Édition web</resp>
</respStmt>
<respStmt>
 <name>Philippe Pons, <affiliation>Centre Alexandre
Koyré</affiliation>
 </name>
```

<editionStmt> et <extent>

La mention d'édition est optionnelle lorsqu'il s'agit de la première édition, obligatoire pour les éditions suivantes....

```
<editionStmt>
 <edition>1<hi rend="sup">ère</hi> édition</edition>
</editionStmt>
```

```
<extent>Taille du fichier (texte plus balises) : ~1Mo ;
texte complet : 567 571
caractères, 89 219 mots ; texte source (sans les notes et le
paratexte) : 407 746
caractères, 65 009 mots.</extent>
```

L'adresse bibliographique (<publicationStmt>)

L'élément <publicationSmt> peut contenir, soit du texte, soit un ou plusieurs éléments <publisher>, <distributor>, <authority>, associés à des éléments <pubPlace>, <address>, <availability>, <idno>, <date>.

```
<publicationStmt>
 <date>2012</date>
 <publisher>École nationale des chartes</publisher>
 <address>
  <addrLine>19, rue de la Sorbonne</addrLine>
  <addrLine>75005 Paris</addrLine>
  <addrLine>tél.: +33 (0)1 55 42 75 00</addrLine>
  <addrLine>http://www.enc.sorbonne.fr/</addrLine>
 </address>
 <availability>
  <p>L'École nationale des chartes met à disposition cette
ressource électronique
     structurée, protégée par le code de la propriété
intellectuelle sur les bases de
     données (<ref tar-
get="http://www.legifrance.gouv.fr/affichCodeArticle.do?idArticle=
    type="externalLink">L341-1</ref>), selon les termes de
la licence Creative
     Commons : « <ref tar-
get="http://creativecommons.org/licenses/by-nc-nd/2.0/fr/"
```

L'élément <seriesStmt>

Cet élément est utilisé lorque 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, ou encore mentionner qu'ils sont nés numériques.

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

- description en prose dans un élément <p>
- une référence bibliographique dans <bibl>, la même chose sous une forme plus contrainte dans <biblStruct> ou complète dans <biblFull>
- une liste de références bibliographiques comme ci-dessus, dans <listBibl>
- en plus, des éléments spécialisés pour les transcriptions de discours oraux (<recordingStmt> ou les manuscrits (<msDesc>).

Exemple de description de source imprimée

```
<sourceDesc>
 <bibl>
  <title>Linguae vasconum primitiae</title> per dominum
<author>Bernardum
      Dechepare</author>, rectorem Sancti Michaelis veteris.
 <pubPlace>[Bourdeaulx]</pubPlace> : <publisher>[François
Morpain,
      imprimeur]</publisher>, <date>[1545]</date>.
<extent>28 ff. : fig. sur bois ;
      in-4</extent>
  <textLang mainLang="baq">basque</textLang>
 </bibl>
</sourceDesc>
```

Exemple de description d'une source manuscrite

```
<sourceDesc>
 <msDesc>
  <msIdentifier>
   <country>France</country>
   <institution>Archives nationales</institution>
   <repository>site de Paris</repository>
   <collection>Maison du roi</collection>
   <idno>O/1/284, n° 525</idno>
  </msIdentifier>
  <msContents>
   <summary>
    <date when="1763-05">1763, mai</date> [Versailles]. En
prévision d'un voyage de
        la cour à Marly, le gouverneur de Versailles soumet
au roi des questions
        relatives au logement d'un certain nombre de
courtisans et de serviteurs royaux
        qui accompagnent le souverain. Le château sera
éclairé comme à l'ordinaire et le
        roi fournira du bois de chauffage à certains
domestiques seulement, ainsi qu'une
        demi bougie et de l'argent au personnel
ecclésiastique.</summary>
   <textLang mainLang="fre">en français</textLang>
  </msContents>
  <physDesc>
   <objectDesc form="feuillet">
```

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>
```

Source née numérique

```
<sourceDesc>
 <p>Document né numérique</p>
</sourceDesc>
```

Description de l'encodage (<encodingDesc>)

<encodingDesc> 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, i.e. <correction>, <normalization>, <quotation>, <hyphenation>, <segmentation>, <interpretation>
- <classDecl> : les systèmes de classification utilisés
- <tagsDecl> : les règles spécifiques applicables à certains éléments

Par exemple...

```
<encodingDesc>
 <projectDesc>
  <p>Édition électronique lancée en avril 2010 dans le
cadre du mémoire de recherche
      de 2<emph rend="super">e</emph> 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>
```

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.

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>
```

<classDecl> et <charDecl> : exemples d'utilisation

```
<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>
 </classDecl>
 <charDecl>
  <glyph xml:id="paraphe">
   <desc>le paraphe de Philippe de Noailles</desc>
   <graphic url="paraphe.jpg"/>
  </glyph>
 </charDecl>
 <geoDecl datum="WGS84"/>
</encodingDesc>
```

On peut décrire son propre balisage

```
<encodingDesc>
<!-- ... -->
 <tagsDecl>
  <namespace name="http://www.tei-c.org/ns/1.0">
   <tagUsage gi="cit" occurs="410"/>
   <tagUsage gi="div" occurs="115"/>
   <tagUsage gi="gap" occurs="3"/>
   <tagUsage gi="head" occurs="156"/>
   <tagUsage gi="hi" occurs="147"/>
   <tagUsage gi="l" occurs="2"/>
   <tagUsage gi="lg" occurs="1"/>
   <tagUsage gi="p" occurs="680"/>
   <tagUsage gi="quote" occurs="3"/>
   <tagUsage gi="s" occurs="2415"/>
   <tagUsage gi="w" occurs="41799"/>
  </namespace>
 </tagsDecl>
</encodingDesc>
```

On peut definir des styles

- <rendition> : donne des informations structurées sur l'apparence de certains composants du document source

```
<tagsDecl>
 <rendition xml:id="r-center"
  scheme="css">text-align: center;</rendition>
 <rendition xml:id="r-small"
  scheme="css">font-size: small;</rendition>
 <rendition xml:id="r-large"
  scheme="css">font-size: large;</rendition>
</tagsDecl>
```

L'élément <appInfo>

- <appInfo> : donne des informations structurées sur le logiciel qui a servi à produire le fichier

```
<appInfo>
 <application version="1.7"
  ident="ImageMarkupTool" notAfter="2008-06-01">
  <label>Image Markup Tool</label>
  <ptr target="#P1"/>
  <ptr target="#P2"/>
 </application>
</appInfo>
```

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. Les éléments disponibles (membres de la classe *model.profileDescPart*) comprennent:

- <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

Création

On peut donner des informations directement dans un élément <p>, ou, dans le cadre d'une édition génétique, déclarer les différentes étapes d'écriture. Exemple :

```
<profileDesc>
 <creation>
  <listChange ordered="true">
   <change xml:id="G_10_1.0">primera campaña de escritura:
correcciones inmediatas
        o "de escritura" (supresiones lineares)</change>
   <change xml:id="G_10_1.1">segunda campaña de escritura:
correcciones no
        inmediatas o "de lectura" (substituciones, añadidos
interlineares, marcas de
        inversión…)</change>
   <change xml:id="G_10_1.2">tercera campaña de escritura:
nuevas correcciones no
        inmediatas o "de lectura" que invalidan o cambian
las correcciones anteriores
        (substituciones, añadidos interlineares, marcas de
inversión…)</change>
   <change xml:id="G_10_1.3">cuarta campaña de escritura:
nuevas correcciones no
        inmediatas o "de lectura" que invalidan o cambian
las correcciones anteriores
        (substituciones, añadidos interlineares, marcas de
```

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>
```

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>
```

La classification ci-dessus s'appliquer 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.

<settingDesc> sert à décrire 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.

```
<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>
 <p>Pierre Mendès France, Entretiens avec Jean Lacouture
```

<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. Une série d'éléments <change> pourra alors être encodé, dans l'ordre chronologique inverse des révisions du texte.

```
<revisionDesc>
 <change when="2012-10" who="#OGJ #FC">Quelques
corrections</change>
 <change when="2012-06" who="#FC">Publication</change>
 <change when="2012-05"
  who="#FC #OGJ #LG">Relecture</change>
 <change when="2012-01" who="#LG">Première
révision</change>
 <change when="2011-11" who="#LG">Premier encodage</change>
</revisionDesc>
```