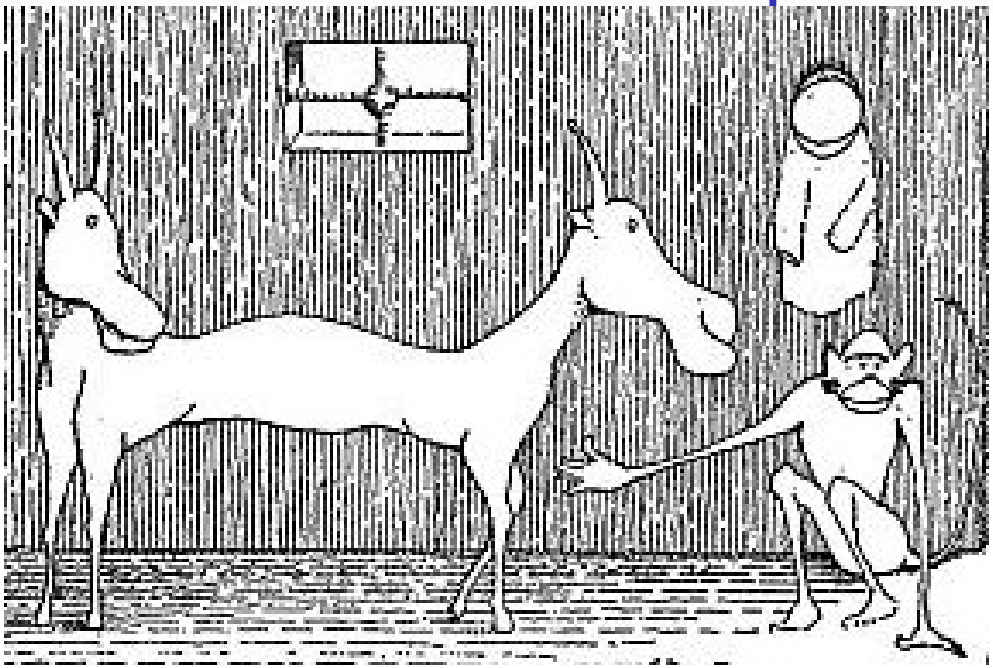
Metadata Matters: l'importance des metadonnées

Lou Burnard

avril 2011

Metadata Matters: l'importance des métadonnées



L'entête TEI: un pouchmi-poulia?

Tout texte TEI doit avoir un entête, une exigence qui répond à deux besoins distincts :

- ceux qui sont propres aux bibliothécaires : comment identifier et décrire un 'livre électronique'
- ceux qui sont propres aux utilisateurs : comment manipuler et comprendre l'encodage d'un texte numérisé

Ce qui introduit une tension

A quoi servent les métadonnées ?

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

- 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

Comment normaliser tout cela ?

Métadonnées numériques

- Les standards de métadonnées ont été conçus pour la description des ressources physiques, en particulier les livres, les documents, etc.
- Les standards Web pour la plupart répondent au besoin pressant d'identifier et de retrouver sur le web les ressources numerisés
- Maintenant qu'on passe du "Web des documents" au "Web des données", tous les deux ont besoin d' évoluer ...

La bonne soupe d'acronymes (1)

Dans le monde de métadonnées il y en a plein...

DCMI: Dublin Core Metadata Initiative Système très simple pour spécifier les métadonnées pour les ressources Web: 15 'lowest common denominator' champs

RDF: Resource Description Framework Standard W3C pour la représentation de n'importe quel système de métadonnées sous la forme d'objets

OAIS: Open Archival Information System Modèle abstrait mais très élaboré pour tout système d'archivage : norme ISO

OAI-PMH: Open Archives Initiative-Protocol for Metadata Harvesting Protocole ouvert pour la dissémination de métadonnées (fournies au format DCMI ou autre)

La bonne soupe d'acronymes (2)

Z39-50: protocole standard ANSI (et ISO 23950) pour la recherche des informations bibliographiques, surtout dans les grandes bibliothèques

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.

Où se positionne la TEI?

L'entête TEI

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

1. <fileDesc>: description bibliographique de la ressource et de ses origines
2. <encodingDesc>: fournit une description du rapport entre la ressource et la source (ou les sources) dont elle dérive
3. <profileDesc>: fournit des informations supplémentaires (non bibliographiques) sur la ressource, par ex. les langues, les participants, les thèmes...
4. <revisionDesc>: résume l'historique des modifications de la ressource

tous facultatifs, sauf le premier

L'entête 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>
```

File Description

- Obligatoire :
  - <titleStmt>: comprenant un titre pour la ressource et les mentions de responsabilité
  - <sourceDesc>: renseignements sur les sources dont dérive le fichier
  - <publicationStmt>: explicite le mode de distribution
- Facultatif:
  - <editionStmt>: pour noter la version de la ressource
  - <extent>: la taille du fichier, tel que stocké sur un support spécifique
  - <seriesStmt>: si la ressource fait partie d'une série d'éditions électroniques
  - <notesStmt>: notes génériques

Identification...

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

Fiabilité de l'identification

...peut être obtenue par

- un balisage plus fin:

```
<principal>
  <persName>
    <surname>Geffin</surname>
    <forename>Alexandre</forename>
  </persName>
</principal>
```

- un appel à un fichier de référence

```
<respStmt>
  <resp>transcription</resp>
  <orgName key="SEPE">Service éditorial et Publications
électroniques</orgName>
</respStmt>
<title ref="urn:isbn:0-395-36341-1">Le Grand Cyrus</title>
```

Description des sources

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

La TEI fournit une gamme riche d'éléments bibliographiques, structurés — ou non:

- <bibl>, <biblStruct>
- (pour un texte déjà informatisé) : <biblFull> (même contenu que <fileDesc>)
- <listBibl> une liste des éléments ci-dessus
- description en prose
- en plus, des éléments spécialisés pour les transcriptions de discours oraux ou les manuscrits.

Source classique (1)

```
<sourceDesc>
  <biblStruct xml:lang="fr">
    <monogr>
      <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 when="1846">MDCCCXLVI</date>
      </imprint>
    </monogr>
  </biblStruct>
</sourceDesc>
```

Source classique (2)

```
<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 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>
```

Pour les participants... voir plus tard

Source née numérique

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

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

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>
```

Description relative au codage (1)

<encodingDesc> regroupe des informations sur les méthodes ayant régi la création du texte numérisé, soit en texte libre, soit en utilisant des éléments spécifiques, tous membres de la classe model.encodingDescPart, y compris:

- <projectDesc> : buts du projet qui a conduit à la création de la ressource
- <samplingDecl> : critères et méthodes de sélection du texte
- <editorialDecl> : informations sur les principes éditoriaux, p.e. <correction>, <normalization>, <quotation>, <hyphenation>, <segmentation>, <interpretation>

Par exemple...

```
<encodingDesc>
  <projectDesc>
    <p>Textes réunis pour être utilisés dans Claremont Shakespeare Clinic,
Juin
       1990.</p>
  </projectDesc>
  <projectDesc xml:lang="zh">
    <p>□□□□□1990□7□□□□□□□□□□□□□</p>
  </projectDesc>
  <samplingDecl>
    <p>Échantillon de 2 000 mots pris à partir du début du texte.</p>
  </samplingDecl>
  <editorialDecl>
    <normalization>
      <p>Certains mots coupés par accident typographique en fin de ligne ont
été
          réassemblés sans commentaire.</p>
    </normalization>
    <quotation marks="all" form="std">
      <p>Les "guillements français" ont été remplacés par des "guillemets
droits" (sans
          symétrie)</p>
    </quotation>
  </editorialDecl>
</encodingDesc>
```

Description relative au codage (2)

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.

En gros, peut remplacer le manuel Mode d'emploi, et faciliter une gestion semi-automatique des documents.

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>
```

On peut fournir une taxinomie "maison"

```
<encodingDesc>
<!-- ... -->
  <classDecl>
    <taxonomy  xml:id="size">
      <category  xml:id="large">
        <catDesc>story occupies more than half a page</catDesc>
      </category>
      <category  xml:id="medium">
        <catDesc>story occupies between quarter and a half
                 page</catDesc>
      </category>
      <category  xml:id="small">
        <catDesc>story occupies less than a quarter
                 page</catDesc>
      </category>
<!-- etc -->
    </taxonomy>
    <taxonomy  xml:id="topic">
      <category  xml:id="politics-domestic">
        <catDesc>CRefers to domestic political
                 events</catDesc>
      </category>
      <category  xml:id="politics-foreign">
        <catDesc>Refers to foreign political
                 events</catDesc>
      </category>
      <category  xml:id="social-women">
        <catDesc>refers to role of women in
                 society</catDesc>
      </category>
      <category  xml:id="social-servants">
        <catDesc>refers to role of servants in
                 society</catDesc>
```

Definition des styles identifies 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>
```

Description du profil

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...

- <creation> : informations sur la création de la ressource, par ex. endroit, 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, par ex. les locuteurs d'un discours oral, les caractères d'un roman
- <settingDesc> : informations sur l'endroit d'une interaction linguistique par ex. le lieu d'enregistrement d'un discours oral, la scène d'un drame.

L'élément <creation>

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 langue du texte en se servant des codes normatifs d'ISO.
L'élément <language> (et son attribut associé, `xml:lang`) peut comprendre un langage, son écriture, et sa région:

```
<langUsage>
  <language ident="en">English</language>
  <language ident="fr-ca" xml:lang="fr">québécois</language>
  <language ident="zh-Latn" xml:lang="fr">Chinese using latin
script</language>
</langUsage>
```

Classification des textes

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

avec <catRef> référence directe à une catégorie définie localement

avec <classCode> référence à une catégorie communément admise et définie à l'externe (par ex. la classification décimale universelle ou CDU)

avec <keywords> assigne des termes descriptifs tirés d'un vocabulaire bibliographique contrôlé ou d'un nuage de tags

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>
```

This categorization applies to the whole text. For more fine grained classification, use @decls on e.g. a element.

Données

L'information sur une personne, un lieu, ou une organisation peut être consignée dans un élément structuré tel que <person>, <place>, ou <org>, contenant

- des éléments génériques
  - <trait>: des caractéristiques plutot stables, par ex. couleur des yeux, ethnicité, climat
  - <state>: des caractéristiques définies à un moment donné par ex. l'adresse, la fonction, la population
  - <event>: événements provoquant généralement un changement d'état, par ex. naissance, mariage, mort, conquête
- un petit nombre de spécialisations de ceux-ci
- un ensemble des attributs temporels (@when, @notBefore, @notAfter, @from, @to)
- possibilité de représenter les liens par un balisage déporté ("standoff") avec l'élément <relation>

Données personnelles (1)

```
<particDesc>
  <listPerson>
    <person  xml:id="P-1234"  sex="1">
      <p>informateur de bonne éducation, né à Shropshire UK, 12 Jan 1950, parle français
            couramment, statut socio-économique (SSE): commerçant.</p>
    </person>
    <person  xml:id="P-4332"  sex="2">
      <persName>
        <surname>Delaunay</surname>
        <forename>Liliane</forename>
        <forename>Alberte</forename>
      </persName>
      <residence  notAfter="1959">
        <address>
          <street>rue de Falaise</street>
          <settlement>la Guérinière, Caen</settlement>
        </address>
      </residence>
      <occupation>serveuse</occupation>
    </person>
    <relationGrp>
      <relation  type="personal"  name="spouse"  mutual="#P-1234 #P-4332"/>
    </relationGrp>
  </listPerson>
</particDesc>
```

```
<u  who="#P-1234">
  <s  n="311">on mange ou on mange pas</s>
</u>
<u  who="#P4332">
  <s  n="312">j'arrive</s>
</u>
```

Description des révisions

Et finalement, on met un <revisionDesc> pour fournir une liste des modifications apportées à une ressource.

```
<revisionDesc>
  <change when="2010-06-04">entièrement revisé pour Mutec</change>
  <change when="2007-08-14">en route vers Montréal</change>
  <change when="2004-11-15">élaboré pour AUF INRIA CARI tutoriel</change>
  <change when="2003-01-1">addition d'entête</change>
</revisionDesc>
```

Entête du corpus, entête du texte

Les métadonnées s'attachent à l'un des trois niveaux:

- globalement à un corpus de textes
- à la totalité d'un seul texte
- à une (ou plusieurs) partie(s) d'un texte

Donc, en TEI on trouve (facultativement):

- un entête de corpus
- un entête par texte
- la possibilité de faire des liens entre éléments att.declaring (p.e.), et att.declarable (p.e. composants de <encodingDesc>)

L'avenir

- L'entête TEI fut conçu il y a 15 ans, comme système utilisable par les non spécialistes
- Donc, il faut y ajouter des règles d'usage spécifiques à son projet pour s'en servir
- Il a été pris en main par des bibliothécaires professionnels : voir, par ex. `http://wiki.tei-c.org/index.php/Best_Practices_for_TEI_in_Libraries`
- Comme 'source des informations primaires' il reste essentiel.