Standards basés sur XML : présentation de XCES, *XML Corpus Encoding Standard*

Andrei Popescu-Belis
TIM / ETI, Université de Genève

Cours n°10a

Rappel : comprendre les standards

- Sens de l'acronyme / titre
- Objectifs généraux
  - décrits dans un document introductif ou sur Internet
- Auteurs, organisateurs
  - utile pour comprendre les motivations et la portée du standard
- Exemples de documents respectant le standard
  - essentiels pour un aperçu introductif
- Définition du standard
  - une DTD et/ou un schéma XML (XSD)
  - un document expliquant la DTD
- Principales applications

Standards pour l'encodage de corpus / banques de textes

1. Première tentative: TEI (définie en SGML)
   - Text Encoding Initiative
2. Simplification de la TEI et adoption des directives du projet EAGLES: CES
   - Corpus Encoding Standard
3. Définition de CES en XML: XCES

« Mais que sont les corpus ? »

- Corpus (parfois : banques de textes)
  - collections de textes
    - au format électronique (en principe !)
    - rassemblés selon certains critères (langue, thème, forme, auteur, etc.)
    - en vue d'effectuer certaines opérations
- Applications
  - accès numérique aux textes, études littéraires, recherches documentaires, lexicographie, traduction (corpus multilingues), création d'outils de traitement de la langue
- Critères essentiels
  - taille, représentativité, réutilisabilité

Objectifs de XCES

- Standard pour baliser la structure des corpus textuels / application XML
- Deux parties
  - Annotation des méta-données = information sur le texte, sa version électronique, l'annotation
  - Annotation du texte = structures sur plusieurs niveaux
    - Niveau de la section / chapitre
    - Niveau du paragraphe
    - Niveau de la phrase
- *XCES : discussion autour de la documentation*
  http://www.cs.vassar.edu/XCES/

Aperçu simplifié de XCES

- Deux classes de balises : entête / corps de texte

```
<cesDoc version="4.3" type="text">
    <cesHeader version="2.0">
     ……
    </cesHeader>
    <text lang="fr">
        <body>
        ……
        </body>
    </text>
</cesDoc>
```

- Nécessité de **définitions** pour les balises

En-tête ou méta-données (1/2)

```
<cesHeader version="2.0">
        <fileDesc>
            <titleStmt>
               <h.title></h.title>
            </titleStmt>
            <publicationStmt>
               <distributor></distributor>
               <pubAddress></pubAddress>
               <availability status="restricted"></availability>
               <pubDate></pubDate>
            </publicationStmt>
            <sourceDesc>
               <biblStruct>
                  <analytic>
                     <h.title></h.title>
                     <h.author></h.author>
                  </analytic>
                  ……
```


En-tête, *suite* (2/2)

```
                  <monogr>
                     <h.title></h.title>
                     <h.author></h.author>
                     <imprint>
                        <pubPlace></pubPlace>
                        <publisher></publisher>
                        <pubDate value="ISO8601">AAAA-MM-JJ</pubDate>
                     </imprint>
                  </monogr>
               </biblStruct>
            </sourceDesc>
         </fileDesc>
         <profileDesc>
            <langUsage>
               <language id="fr" iso639="fr">French</language>
               <language id="en" iso639="en">English</language>
               <language id="la" iso639="la">Latin</language>
            </langUsage>
         </profileDesc>
      </cesHeader>
```


Définitions des balises XCES de l'entête (1/3)

- **<fileDesc>** Contient une description bibliographique complète d'un fichier électronique. [TEI]
- **<titleStmt>** Regroupe des informations concernant le corpus électronique (et non pas le document sur papier). Le titre doit être différent de celui du document imprimé original.
- **<h.title>** Le titre et l'éventuel sous-titre du corpus rassemblé dans le fichier.
- **<publicationStmt>** Regroupe des informations concernant la publication ou la diffusion d'un texte qu'il soit électronique ou non. [TEI] Sert à préciser la disponibilité publique du corpus
- **<distributor>** Fournit le nom d'une personne ou d'une institution responsable de la diffusion d'un texte. [TEI]
- **<pubAddress>** Adresse postale du distributeur.
- **<availability>** Fournit des informations concernant la disponibilité d'un texte, par exemple toute restriction sur son emploi ou diffusion, son statut en matière de droits d'auteur, etc. [TEI]Trois statuts sont possibles : free, restricted, unknown. Ils doivent être décrits sous la forme suivantes :<availability status="restricted">

Définitions des balises XCES de l'entête (2/3)

- **<pubDate>** Date de constitution du corpus présent dans ce fichier.
- **<sourceDesc>** Fournit une description bibliographique du texte original à partir duquel un texte électronique a été dérivé ou généré. [TEI] Il s'agit ici de la description du document imprimé proprement dit.
- **<biblStruct>** *Se compose des sous-éléments suivants :*
- **<monogr>** Utilisé pour les monographies et les publications en série.
- **<h.title>** Titre du document.
- **<h.author>** Auteur du document.
- **<imprint>** Indications relatives à la publication contenant les sous-éléments suivants :
- **<pubPlace>** Lieu d'édition.
- **<publisher>** Éditeur. Il peut être de trois types :
- **<pubDate>** Date de publication. À préciser sous la forme année-mois-jour :<pubDate value="ISO8601">AAAA-MM-JJ</pubDate>Si la mention du jour ou celle du mois sont inconnues, elles peuvent être omises : <pubDate value="ISO8601">2003</pubDate>

Définitions des balises XCES de l'entête (3/3)

- **<analytic>** Pour les parties de monographies (contributions) ou de publications en série (articles).Cette balise précède <monogr> et doit contenir une balise <h.author> (auteur de la contribution ou de l'article) et une balise <h.title> (titre de la contribution ou de l'article).
- **<profileDesc>** Fournit une description détaillée des aspects non bibliographiques d'un texte, spécifiquement les langues et le sous-langues employées, les circonstances de sa production, les participants, et leur environnement. [TEI]
- **<langUsage>** Décrit les langues, les sous-langues, les registres, les dialectes, etc., représentés à l'intérieur un texte. [TEI] À décrire sous la forme suivante :

```
– <langUsage>
    <language id="fr" iso639="fr">French</language>
    <language id="en" iso639="en">English</language>
    <language id="la" iso639="la">Latin</language>
  <langUsage>
```

- Pour spécifier la langue dans le texte, utiliser l'attribut *lang*.

Balises XCES pour le contenu (textes écrits)

```
<text lang="fr">
   <body>
      <div>
         <head>…</head>
         <p>...</p>
            <foreign lang="en">…</foreign>
         <sp>
            <speaker>...</speaker>
            <stage>...</stage>
         </sp>
         <poem>
            <lg >...</lg>
            <l>...</l>
         </poem>
         <list>
            <item>...</item>
         </list>
         ……
         ……
         <figure>
            <head>...</head>
            <figDesc>...</figDesc>
         </figure>
         <table>
            <row>...</row>
            <cell>...</cell>
         </table>
         <bibl>…</bibl>
         <caption>…</caption>
         <quote>…</quote>
         <note>...</note>
      </div>
   </body>
</text>
```

Balises du niveau du texte

- **<text>**
  Balise qui marque le début et la fin du texte et doit contenir la balise <body>
- **<body>**
  Contient le corps entier d'un texte unitaire unique, à l'exclusion de toute pièce liminaire ou annexe [TEI]
- ** **
  Contient une subdivision du texte. Le *type* permet de catégoriser la subdivision par une liste d'attributs (chapitre, section…)
  Le *n* ou *id* permettent de préciser la numérotation de la subdivision.

- **<head>**
  Contient tout type de titre, comme par exemple, le titre d'une section, ou l'en-tête d'une liste ou d'un glossaire. [TEI]

13

Balises du niveau du paragraphe (1)

- **<p>** = Un paragraphe d'un texte écrit
- **<sp>** = Passage parlé à l'écrit (dialogue, interview)
- **<speaker>** = Identification de la personne qui parle
- **<stage>** = Didascalie
- **<poem>** = Poème ou passage versifié
- **<lg>** = Toute subdivision du poème pertinente.
  Exemple <lg type="strophe" n="7">
- **<l>** = Vers
- **<list>** = Liste d'éléments (avec tirets ou puces)
  - **<item>** = Tout élément d'une liste
- **<figure>** = Figure, graphique, illustration.
  - **<head>** = Permet de préciser un titre éventuel.
  - **<figDesc>** = Description de la figure (si elle n'est pas au format texte)

14

Balises du niveau du paragraphe (2)

- **<table>** = Tout texte disposé sous forme de lignes et de colonnes.
  Exemple : <table cols="2" rows="2">
  - **<row>** = ligne où se trouve l'information
  - **<cell>** = colonne où se trouve l'information
- **<bibl>** = Référence bibliographique
- **<caption>** = Légende d'une image, figure, tableau…
- **<quote>** = Citation d'un autre auteur (paragraphe séparé)
- **<note>** = Note ou une annotation, avec des attributs pour indiquer le type, l'emplacement et la source de la note.
  Exemple : <note id="N1" place="foot">.
  Appel de note : <ptr target="N1"/>

15

Attributs

- ***lang*** = indique la langue, reprenant par exemple l'élément **<language>** du *header*
- Exemples : <poem lang="fr"> ou <foreign lang="en">
- ***id*** = identifiant unique d'un élément ; doit commencer avec une lettre, peut contenir des lettres, des chiffres, des tirets ou des points
- ***n*** = nom ou nombre de cet élément ; peut comporter toute chaîne de caractères ; souvent employé pour enregistrer des systèmes de référence traditionnels (p.ex. notes, etc.)

16

Conclusion sur XCES

- Standard répandu pour l'encodage de textes
- Permet d'avoir un format commun de structure
  - sépare le contenu (XCES) de la forme (HTML)
- Mécanisme de feuilles de style assez complexe
  - ne fonctionne pas avec tous les processeurs XSLT
  - peut être simplifié selon les besoins de chacun

17

Exercices

Découverte et affichage de textes au format XCES

Exercices

- Téléchargez (de préférence dans un dossier séparé)
  - les quatre textes encodés en XCES
    - xces-exemple-hugo.xml, xces-multext-1984-header.xml, xces-multext-joc-body.xml, xces-cahiers-rifal.xml
  - la DTD et son fichier auxiliaire
    - xcesDoc.dtd, xheader.elt
  - les feuilles de style
    - xces-vers-html.xsl (feuille simplifiée)
    - le paquet xces-rifal-xsl.zip
- Décompressez le paquet ZIP dans le même dossier que les autres fichiers (bouton droit de la souris, commande 'Extraire ici')

19

Questions

- Pour les quatre documents XML téléchargés
  - sont-ils valides ? Bien formés ? (Utilisez XMLSpy)
  - que contiennent-ils ?
- En utilisant chacune des deux feuilles de style fournies, convertissez ces documents en HTML
  - pour xces-vers-html.xsl indiquer ce fichier à XML Spy
  - pour appliquer le paquet, il faut indiquer à XML la feuille xces-rifal-xsl\html\cesDoc.xsl
    - les autres feuilles du paquet seront appelées automatiquement
- Noter pour chaque fichier laquelle des deux feuilles de style fournit un meilleur formatage HTML
  - laquelle des feuilles vous paraît préférable ?
  - laquelle des feuilles vous paraît la plus intelligible ?

20

Création d'un document XCES

- Sur le modèle du poème d'Hugo, encodez en XCES un court poème ou texte de votre choix
- Vérifiez-en la bonne formation et la validité
- Transformez-le en HTML et vérifiez l'affichage
- **Devoir** : mettez-le dans votre dossier Web personnel sous le nom poeme.xml

21