Initiation XML-TEI

Initiation à l'encodage des textes patrimoniaux

Stage XML/TEI

(Niveau 1)

janvier 2010

Objectifs de cette presentation

1. Préciser ce que c'est que l'encodage textuel
2. Présenter les concepts fondamentaux de TEI-XML

La numerisation nous apporte de nouveaux defis!

De plus en plus, on veut faire des choses nouvelles avec nos objets numériques:

- construire une base de donnees mutualisée, des instruments de recherche (*finding aid*)
- integrer de tels instruments avec les textes qu'ils signallent
- integrer de tels instruments dans une espece de mère porteuse numérique, (*edition numerique*)
- donner support aux outils d'analyse complexe (‘text-mining’) distribués

La TEI peut nous aider...

Elle represente une modele conceptuelle bien établie et consensuelle qui facilite alors

- la conversion des donnees existantes
- la création des données nouvelles
- l'intégration des données deja existantes mais répandues dans plusieures sources

Elle est basée sur des formats ouverts et des technologies ouvertes

Elle s'appuye sur une théorie explicite de l'ontologie textuel

Qu'est-ce que c'est qu'un texte (1)?

A MONSEI-

GNEVR LE REVE-
rendiſſime Cardinal
du Bellay.

S.

EV le Perſonnaige, que tu ioues au Spectacle de toute l'Europe, uoyre de tout le Monde en ce grand Theatre Romain, ueu tant d'affaires, & telz, que ſeul quaſi tu ſoutiens: ô l'Hõneur du ſacré College! pecheroy'-ie pas (comme dit le Pindare Latin) contre le bien publicq', ſi par longues paroles i'empeſchoy' le tens, que tu donnes au

Qu'est-ce que c'est qu'un texte (2)?

File Edit View History Bookmarks Tools Help

http://www.tlfq.ulaval.ca/axl/Francophonie/Du_Bellay.htm

Most Visited Latest Headlines Google Maps Adonis-related Travel TextAnalysis Shopping

Joachim du Bellay

Défense et illustration de la langue françoyse (1549)

La Deffence, et Illustration de la Langue Francoyse

L'auteur prie les lecteurs différer leur jugement jusques à la fin du livre, et ne le condamner sans avoir premièrement bien vu, et examiné ses raisons.

Épître à Monseigneur le révérendissime cardinal du Bellay S.

Vu le personnage que tu joues au spectacle de toute l'Europe, voire de tout le monde, en ce grand Théâtre Romain, vu tant d'affaires, et tels que seul quasi tu soutiens, ô l'honneur du sacré Collège, pécherai-je pas (comme dit le Pindare Latin) contre le bien public, si par longues paroles j'empêchais le temps que tu donnes au service de ton prince, au profit de la patrie et à l'accroissement de ton immortelle renommée ? Épiant donc quelques heures de ce peu de relais que tu prends pour respirer sous le pesant faix des affaires françaises (charge vraiment digne de si robustes épaules, non moins que le ciel de celles du grand Hercule), ma Muse a pris la hardiesse d'entrer au sacré cabinet de tes

L'ontologie textuel

En quoi consiste l'essentiel d'un texte ?

- en l'apparence des lettres et leur mise-en-page?
- en la version originelle (pretendue) de cette copie?
- en les interpretations/lectures apportées ou trouvées? en les intentions (supposées) de son auteur?

Un "texte" est quelque chose d'abstrait: la construction d'un communauté de lecteurs.

L'encodage explicite cette abstraction à fin de la mieux gérer

L'ontologie textuel

En quoi consiste l'essentiel d'un texte ?

- en l'apparence des lettres et leur mise-en-page?
- en la version originelle (pretendue) de cette copie?
- en les interpretations/lectures apportées ou trouvées? en les intentions (supposées) de son auteur?

Un "texte" est quelque chose d'abstrait: la construction d'un communauté de lecteurs.

L'encodage explicite cette abstraction à fin de la mieux gérer

L'ontologie textuel

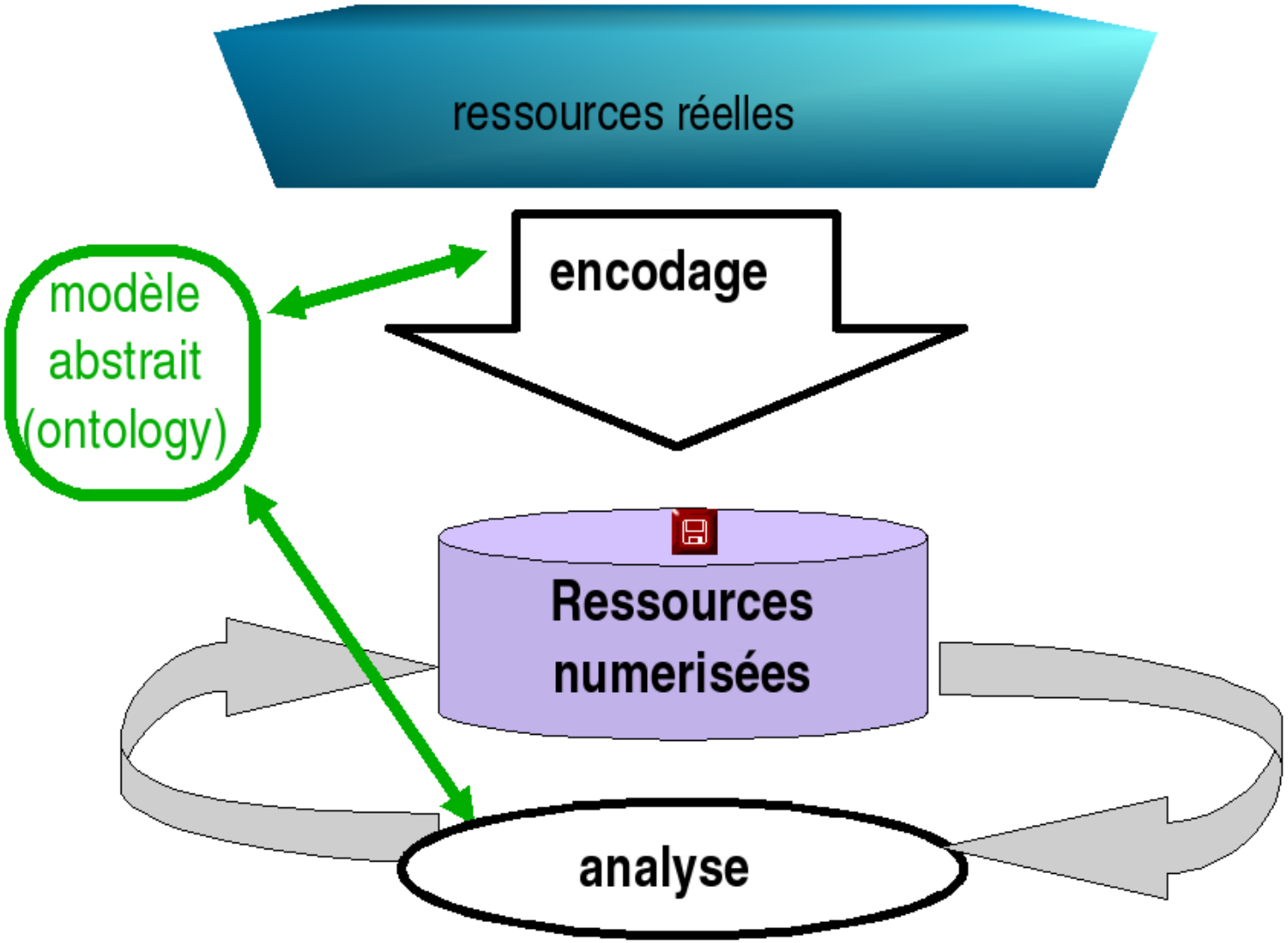
En quoi consiste l'essentiel d'un texte ?

- en l'apparence des lettres et leur mise-en-page?
- en la version originelle (pretendue) de cette copie?
- en les interpretations/lectures apportées ou trouvées? en les intentions (supposées) de son auteur?

Un "texte" est quelque chose d'abstrait: la construction d'un communauté de lecteurs.

L'encodage explicite cette abstraction à fin de la mieux gérer

Qu'est-ce qu'on fait en numérisant un texte?

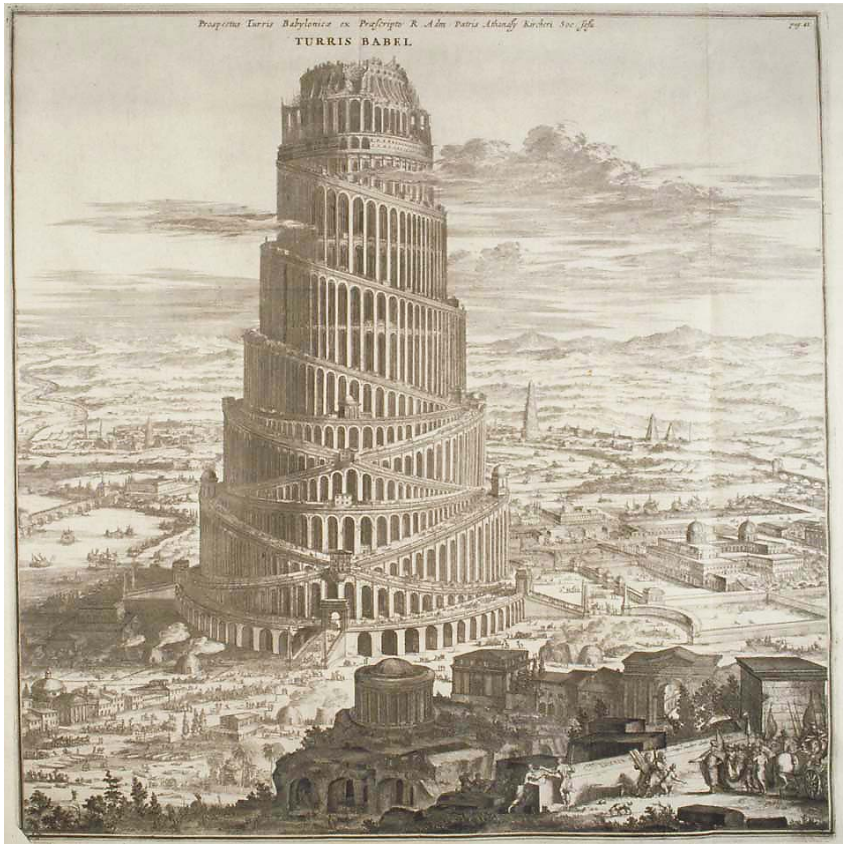


L'encodage

- Un texte est plus qu'une séquence de caractères encodés
  - Il a une structure et une signification
  - Un texte peut avoir plusieurs lectures variantes
  - La portée d'un texte peut être enrichie par des annotations
- L'encodage explicite les lectures
- Sans explicitation, on ne peut rien traiter

L'effet Babel

Bien sûr il existe plusieurs lectures possibles pour la plupart des textes...



... et (malheureseument) plusieurs manières d'expression pour ces lectures!

Par exemple...

I

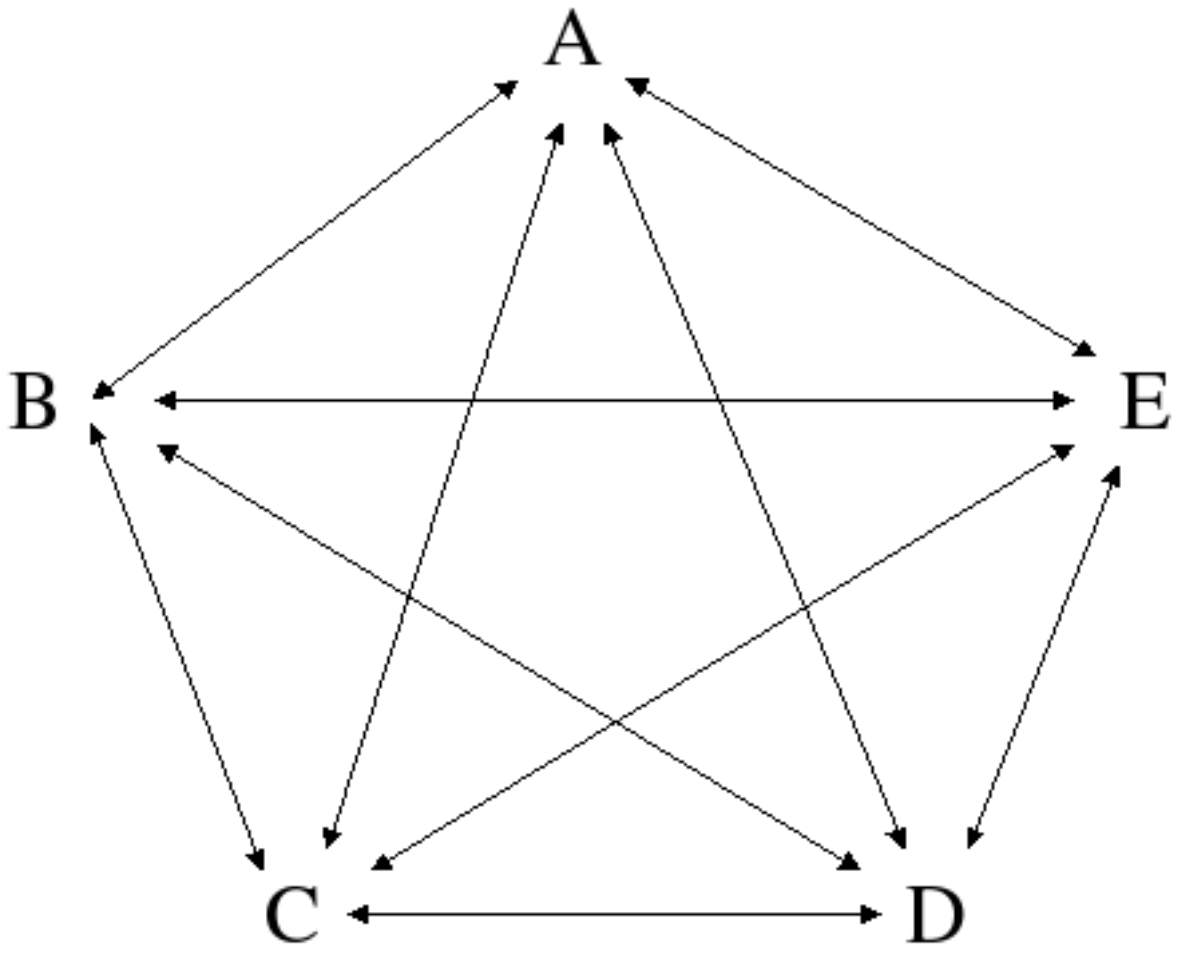
Loomings

Call me Ishmael. Some years ago – never mind how long precisely – having little or no money in my purse, and nothing particular to interest me on shore, I thought I would sail about a little and see the watery part of the world. It is a way I have of driving off the

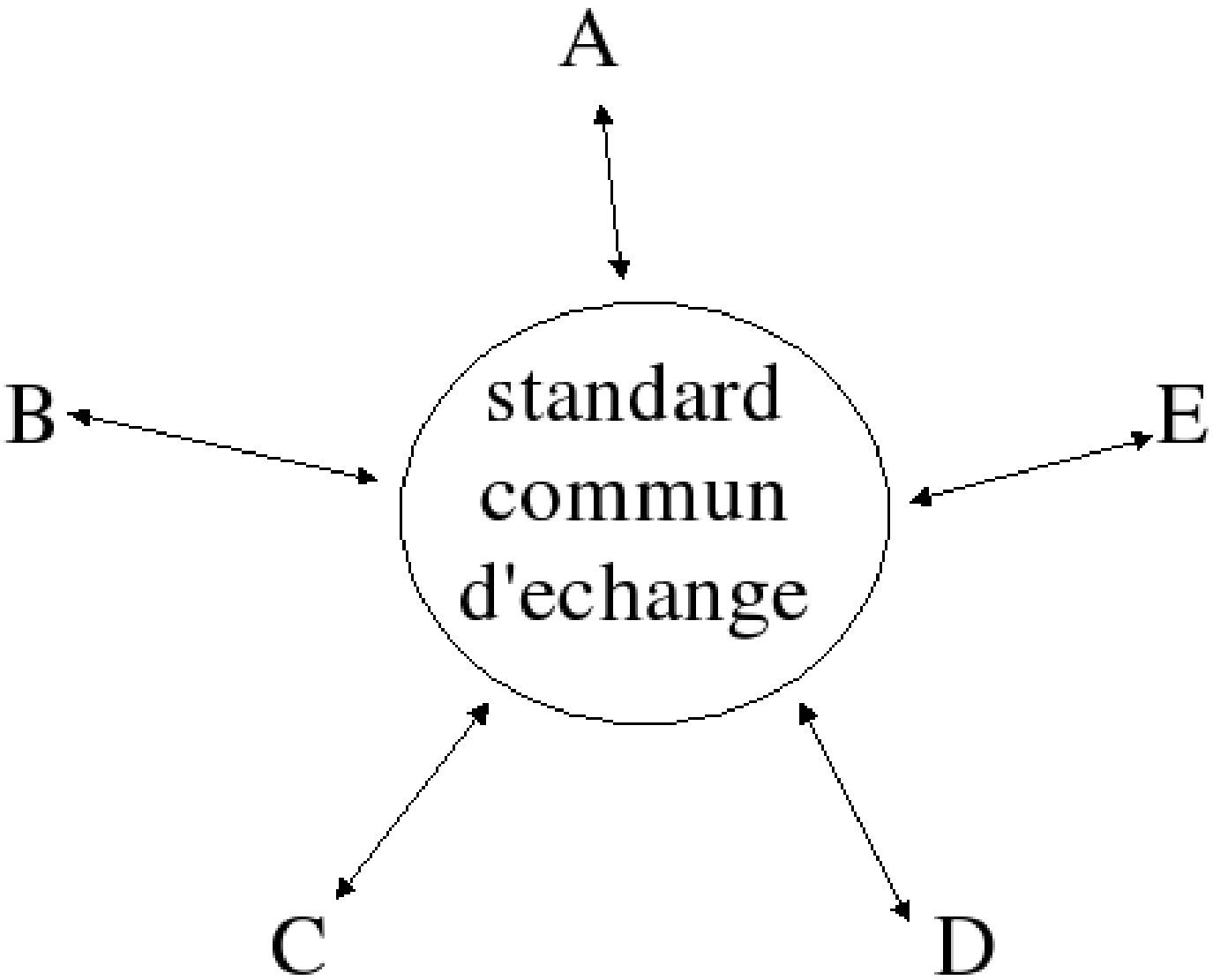
Encodage ou babel?

```
|chap1
<C 1> Loomings
\chapter[1]{Loomings}
:h1.1.  Loomings
MOBY001001LOOMINGS
|C1
.chapter Loomings
.cp;.sp 6 a;.ce .bd 1. Loomings
<h1>Loomings</h1>
<p class="h1">Loomings
~x
```

Echange d'informations (1)



Echange d'informations (2)



Genres de balisage

- Au debut, il y avait le balisage *procédural*

```
ENCRE ROUGE; imprimer balance; ENCRE NOIRE
```

- qui, avec généralisation, est devenu le balisage *descriptif*

```
<balance type='overdrawn'>des chiffres</balance>
```

Un balisage descriptif facilite

- la ré-utilisation des données
- une évolution modulaire et progressive de software

Définitions

- Un balisage explicite les distinctions qu'on désire faire en traitant une chaîne de caractères
- Le balisage est une manière de nommer et de caractériser les composants d'une structure textuelle, d'une manière quasiment formelle
- Quel genre de composants? les objets ou leur apparences?

Séparation de forme et contenu

- Un balisage descriptif s'intéresse plus au contenu qu'à sa mise à jour
- cette séparation facilite la ré-utilisation
- et augmente la flexibilité

Séparation de forme et contenu

- Un balisage descriptif s'intéresse plus au contenu qu'à sa mise à jour
- cette séparation facilite la ré-utilisation
- et augmente la flexibilité

Séparation de forme et contenu

- Un balisage descriptif s'intéresse plus au contenu qu'à sa mise à jour
- cette séparation facilite la ré-utilisation
- et augmente la flexibilité

Qu'est ce qu'on balisera?

Comparer:

```
<pb n="4"/>A MONSEI-
<lb/>GNEUR LE REVE-
<lb/>rendissime Cardinal
<lb/>du Bellay.
<lb/>S

<lb/>
<c rend="lettrine">V</c>EU le Personnaige,
<lb/>que tu joues au Spec-
<lb/>tacle de toute l'Europe...
```

avec

```
<div type="dedicace">
 <head>A MONSEIGNEUR LE REVERENDISSIME CARDINAL DU BELLAY</head>
 <salute>S<ex>alut</ex>
 </salute>
 <p>
  <c rend="lettrine">V</c>EU le Personnaige, que tu joues au
   Spectacle de toute l'Europe...
 </p>...
</div>
```

Un langage d'encodage sert à...

- spécifier les caractères d'un texte
- expliciter la/les structures aperçue/s dans un texte
- linéariser le texte
- spécifier les méta-informations, renseignements contextuels etc.

Mais il faut choisir... selon les buts du projet

La bonne soupe d'acronymes

SGML	Standard Generalized Markup Language
HTML	Hypertext Markup Language
W3C	World Wide Web Consortium
XML	eXtensible Markup Language
DTD	Document Type Definition (or Declaration)
CSS	Cascading Style Sheet
Xpath	XML Path Language
XSLT	eXtensible Stylesheet Language - Transformations
RelaxNG	Regular Expression Language for XML (New Generation)

à ne pas oublier TEI, la *Text Encoding Initiative*

XML: ce que c'est et pourquoi on devrait le connaitre

- XML est une manière de representer les données structurées en forme de chaîne de caractères
- un document XML ressemble à un document HTML, sauf que:-
  - XML est extensible
  - un document XML doit être bien formé
  - un document XML peut être valide
- XML est indépendant de l'application, de la plateforme et du vendeur
- XML rend le pouvoir aux fournisseurs de données, et facilite l'intégration des ressources diverses et polyglottes

(Presque) tout ce qu'il faut savoir au sujet de l'XML, sur un transparent

- Un document XML document contient au moin un *element*
- Un element possede une *balise d'ouverture*, facultativement de *contenu* et une *balise de fermeture*
- Un element peut d'ailleurs porter des *attributs*, chacun portant un *nom* et une *valeur*
- Un document XML est *obligatoirement* 'well formed' (bien-forme) i.e. il doit suivre la syntaxe XML
- Un document bien-forme peut *facultativement* etre *valide* i.e. il est conforme aux regles d'une *schema* quelconque

Un petit document XML

```
<?xml version="1.0" encoding="utf-8" ?>
  <cookBook>
  <recipe n="1">
    <head>Soupe de pierre</head>
    <ingredientList>
      <ingredient>un oignon</ingredient>
      <ingredient>deux carottes</ingredient>
      <ingredient>de l'eau</ingredient>
         ...
      <ingredient>une pierre</ingredient>
      <ingredient>des paysans naïfs</ingredient>
    </ingredientList>
    <procedure>
      <step>mettre l'eau à bouillir dans un grande chaudron</step>
       ....
      <step>enlever la pierre et servir</step>
    </procedure>
   </recipe>
   <recipe n="2">
          <!-- deuxieme recette ici -->
  </recipe>
        <!-- hic desunt multa -->
</cookBook>
```

Syntaxe XML

Un document XML contient:-

- des *éléments*, qui portent (facultativement) des *attributs*, marqués par *balises*
- des *commentaires*
- des *instructions de traîtement*
- des *references à entité* (interne ou externe)
- des sections CDATA
- ...et des caractères Unicode

C'est tout!

XML: règles du jeu

- Un document XML représente une arborescence composée de noeuds
- il y a un seul noeud racine qui contient tous les autres
- chaque noeud peut être
  - une arborescence
  - un élément (qui porte facultativement des attributs)
  - une chaîne de caractères
- Chaque élément porte un nom ou identification générique
- Chaque attribut porte un nom et une valeur
- les noms sont liés avec un namespace (espace de noms)

Representation d'une arborescence XML

- Un document XML linéarisé commence par une instruction de traitement special
- Les occurrences d'élément sont marqués entre balises ouvrantes et balises fermantes
- Les caractères < et & sont Magiques et doivent être cachés au moyen de références entité (< et & respectivement)
- Les paires nom/valeurs qui constituent les attributs d'un élément peuvent apparaître sans ordre à l'intérieur d'une balise ouvrante
- L'espace de noms auquel appartient un élément peut être signalé par un namespace-prefix (p.e. xml:) prédéfini

Syntaxe XML: le "fine print"

Pour qu'un document soit *bien formé*, il faut que:

1. une seul racine contienne le document entier
2. chaque arborescence soit proprement imbriquée
3. tout les noms soient sensibles à la casse
4. chaque balise ouvrante ait sa balise fermante (sauf qu'on peut combiner les deux, le noeud étant vide)
5. les valeurs d'attribut soient présentées correctement entre guillemets

Bien formé? Oui ou non?

- `<seg>some text</seg>`
- `<seg><foo>some</foo> <bar>text</bar></seg>`
- `<seg><foo>some <bar></foo> text</bar></seg>`
- `<seg type="text">some text</seg>`
- `<seg type='text'>some text</seg>`
- `<seg type=text>some text</seg>`
- `<seg type = "text">some text</seg>`
- `<seg type="text">some text<seg/>`
- `<seg type="text">some text<gap/></seg>`
- `<seg type="text">some text< /seg>`
- `<seg type="text">some text</Seg>`

Bien formé? Oui ou non?

- `<seg>some text</seg>`
- `<seg><foo>some</foo> <bar>text</bar></seg>`
- `<seg><foo>some <bar></foo> text</bar></seg>`
- `<seg type="text">some text</seg>`
- `<seg type='text'>some text</seg>`
- `<seg type=text>some text</seg>`
- `<seg type = "text">some text</seg>`
- `<seg type="text">some text<seg/>`
- `<seg type="text">some text<gap/></seg>`
- `<seg type="text">some text< /seg>`
- `<seg type="text">some text</Seg>`

Bien formé? Oui ou non?

- `<seg>some text</seg>`
- `<seg><foo>some</foo> <bar>text</bar></seg>`
- `<seg><foo>some <bar></foo> text</bar></seg>`
- `<seg type="text">some text</seg>`
- `<seg type='text'>some text</seg>`
- `<seg type=text>some text</seg>`
- `<seg type = "text">some text</seg>`
- `<seg type="text">some text<seg/>`
- `<seg type="text">some text<gap/></seg>`
- `<seg type="text">some text< /seg>`
- `<seg type="text">some text</Seg>`

Bien formé? Oui ou non?

- `<seg>some text</seg>`
- `<seg><foo>some</foo> <bar>text</bar></seg>`
- `<seg><foo>some <bar></foo> text</bar></seg>`
- `<seg type="text">some text</seg>`
- `<seg type='text'>some text</seg>`
- `<seg type=text>some text</seg>`
- `<seg type = "text">some text</seg>`
- `<seg type="text">some text<seg/>`
- `<seg type="text">some text<gap/></seg>`
- `<seg type="text">some text< /seg>`
- `<seg type="text">some text</Seg>`

Bien formé? Oui ou non?

- `<seg>some text</seg>`
- `<seg><foo>some</foo> <bar>text</bar></seg>`
- `<seg><foo>some <bar></foo> text</bar></seg>`
- `<seg type="text">some text</seg>`
- `<seg type='text'>some text</seg>`
- `<seg type=text>some text</seg>`
- `<seg type = "text">some text</seg>`
- `<seg type="text">some text<seg/>`
- `<seg type="text">some text<gap/></seg>`
- `<seg type="text">some text< /seg>`
- `<seg type="text">some text</Seg>`

Bien formé? Oui ou non?

- `<seg>some text</seg>`
- `<seg><foo>some</foo> <bar>text</bar></seg>`
- `<seg><foo>some <bar></foo> text</bar></seg>`
- `<seg type="text">some text</seg>`
- `<seg type='text'>some text</seg>`
- `<seg type=text>some text</seg>`
- `<seg type = "text">some text</seg>`
- `<seg type="text">some text<seg/>`
- `<seg type="text">some text<gap/></seg>`
- `<seg type="text">some text< /seg>`
- `<seg type="text">some text</Seg>`

Bien formé? Oui ou non?

- `<seg>some text</seg>`
- `<seg><foo>some</foo> <bar>text</bar></seg>`
- `<seg><foo>some <bar></foo> text</bar></seg>`
- `<seg type="text">some text</seg>`
- `<seg type='text'>some text</seg>`
- `<seg type=text>some text</seg>`
- `<seg type = "text">some text</seg>`
- `<seg type="text">some text<seg/>`
- `<seg type="text">some text<gap/></seg>`
- `<seg type="text">some text< /seg>`
- `<seg type="text">some text</Seg>`

Bien formé? Oui ou non?

- `<seg>some text</seg>`
- `<seg><foo>some</foo> <bar>text</bar></seg>`
- `<seg><foo>some <bar></foo> text</bar></seg>`
- `<seg type="text">some text</seg>`
- `<seg type='text'>some text</seg>`
- `<seg type=text>some text</seg>`
- `<seg type = "text">some text</seg>`
- `<seg type="text">some text<seg/>`
- `<seg type="text">some text<gap/></seg>`
- `<seg type="text">some text< /seg>`
- `<seg type="text">some text</Seg>`

Bien formé? Oui ou non?

- `<seg>some text</seg>`
- `<seg><foo>some</foo> <bar>text</bar></seg>`
- `<seg><foo>some <bar></foo> text</bar></seg>`
- `<seg type="text">some text</seg>`
- `<seg type='text'>some text</seg>`
- `<seg type=text>some text</seg>`
- `<seg type = "text">some text</seg>`
- `<seg type="text">some text<seg/>`
- `<seg type="text">some text<gap/></seg>`
- `<seg type="text">some text< /seg>`
- `<seg type="text">some text</Seg>`

Bien formé? Oui ou non?

- `<seg>some text</seg>`
- `<seg><foo>some</foo> <bar>text</bar></seg>`
- `<seg><foo>some <bar></foo> text</bar></seg>`
- `<seg type="text">some text</seg>`
- `<seg type='text'>some text</seg>`
- `<seg type=text>some text</seg>`
- `<seg type = "text">some text</seg>`
- `<seg type="text">some text<seg/>`
- `<seg type="text">some text<gap/></seg>`
- `<seg type="text">some text< /seg>`
- `<seg type="text">some text</Seg>`

Bien formé? Oui ou non?

- `<seg>some text</seg>`
- `<seg><foo>some</foo> <bar>text</bar></seg>`
- `<seg><foo>some <bar></foo> text</bar></seg>`
- `<seg type="text">some text</seg>`
- `<seg type='text'>some text</seg>`
- `<seg type=text>some text</seg>`
- `<seg type = "text">some text</seg>`
- `<seg type="text">some text<seg/>`
- `<seg type="text">some text<gap/></seg>`
- `<seg type="text">some text< /seg>`
- `<seg type="text">some text</Seg>`

Validation XML

Un document XML *valide* est (bien sûr) bien formé, et en plus conforme à des règles supplémentaires, qui constituent un *schéma*

Un schéma peut spécifier:

- le nom de l'élément racine
- les noms de tous les éléments légaux
- les noms et les types des attributs
- des règles concernant l'imbrication et le contenu des éléments
- et quelques autres menus propos...

n.b. Un schéma ne spécifie point la signification sémantique des éléments

Langues de schéma

Un schéma peut être exprimé en :

- WSD: langage schéma du W3C
- RNG: norme ISO "Relax NG"
- DTD: norme ISO

La TEI se sert de Relax NG

Par exemple...


N° 6 — Première année

CINQ CENTIMES

Lundi 6 août 1883

LE PETIT COMTOIS

JOURNAL RÉPUBLICAIN DÉMOCRATIQUE QUOTIDIEN

RÉDACTEUR EN CHEF, JULES GROS

ABONNEMENTS

RÉDACTION ET ADMINISTRATION

BESANÇON — 7, Square Saint-Amour, 7 — BESANÇON

INSERTIONS

CHEMIN DE FER. — Service d'été.

Départ de Besançon

Arrivée à Besançon

LA POLITIQUE COLONIALE

JEAN DES CHAPRAIS.

DÉPÊCHES DE NUIT

Service de notre correspondant spécial.

INFORMATIONS GÉNÉRALES

Les magistrats des commissions mixtes.

Incident à Philippeville

Un petit exercice intellectuel...

Imaginez vous: on a des milliers des pages à baliser....

- Quels traits existent dans ces materiaux?
- Quels traits faut-il baliser? lequels seraient (in)utiles?
- Comment justifier l'étiquettage choisi?
- Comment garantir et la consistence et l'étendue du balisage?

Maintenant, on réduit le budget de 50%. Répétez l'exercice!

Par exemple...

On commence par noter la structuration:

- la page contient des entêtes, des colonnes, des lignes, etc.
- le journal contient des titres, des regroupements de notices, une episode de feuilleton, des notices de publicité, etc.
- dans le texte du journal, il y a partout des noms d'autre journaux, de personnages, des lieux, etc.
- le texte est (principalement) en français, d'un style d'interêt historique
- on note également des références aux evênements historiques
- et on a bien-sûr des infos supplementaires regardant la production, la dissemination, la bibliographie etc. de cette source...

Focalisons...

con-
alle-
ir les
écon-
ibles
com-
s na-

stion
'agito-
poli-
avec
térêts
a dis-
dence
alle-

tarum, recteur de l'académie d'Alger, a demandé la révocation de ce professeur imprudent.

L'organisation de l'artillerie de forteresse.

Le général Tricoche, directeur de l'artillerie au ministère de la guerre, sera nommé général de division et chargé de l'organisation des batteries d'artillerie de forteresse, avec le titre d'inspecteur général.

Il sera remplacé dans ses fonctions au ministère par le général Lavocat.

Tremblement de terre en Grèce.

Aujourd'hui, à deux heures du matin, un fort tremblement de terre a été ressenti au Pirée.

---

ifique

cour a M. de Sauvelat, et lui demande la

Balisage "sans balises"

```
Un journal algérien annonce que M. Constantin, recteur de l'académie
d'Alger, a demandé la révocation de ce professeur imprudent.
L'organisation de l'artillerie de forteresse Le général Tricoche,
directeur de l'artillerie au ministère de la guerre, sera nommé
général de division et chargé des l'organisation des batteries
d'artillerie de forteresse, avec le titre d'inspecteur général.  Il
sera remplacé dans ses fonctions au ministère par le général Lavocat.
Tremblement de terre en Grèce. Aujourd'hui, à deux heures du matin, un
fort tremblement de terre a été ressenti au Pirée.  0n télégraphie
d'Athènes qu'aucun accident de personnes n'est à déplorer.
```

Un balisage TEI

```
<div type="groupe" decls="DDN" n="4">
 <head type="RUBRIQUE">DEPECHES DE NUIT</head>
 <head type="SIGNATURE">Service de notre correspondant
spécial.</head>
 <head type="RUBRIQUE">INFORMATIONS GENERALES</head>
 <div type="NOTICE" rend="EMP.4M LG.1" n="4">
  <head type="TITRE">L'organisation de l'artillerie de
forteresse</head>
   <p>Le général Tricoche, directeur de l'artillerie
      au ministère de la guerre, sera nommé général de
      division et chargé des l'organisation des batteries
      d'artillerie de forteresse, avec le titre d'inspecteur
      général.  </p>
   <p>Il sera remplacé dans ses fonctions au ministère
      par le général Lavocat.</p>
 </div>
 <div type="NOTICE" rend="EMP.4M LG.1" n="5">
  <head>Tremblement de terre en Grèce.</head>
  <p>Aujourd'hui, à deux heures du matin, un fort
      tremblement de terre a été ressenti au Pirée.</p>
  <p>On télégraphie d'Athènes qu'aucun accident de
      personnes n'est à déplorer.</p>
 </div>
</div>
```

Encore un balisage TEI...

```
<p>
 <s n="42">
  <w type="NOM" lemma="XXXX">0n</w>
  <w type="VER:pres" lemma="télégraphier">télégraphie</w>
  <w type="PRP" lemma="de">d'</w>
  <w type="NOM" lemma="XXXX">Athènes</w>
  <w type="KON" lemma="que">qu'</w>
  <w type="PRO:IND" lemma="aucun">aucun</w>
  <w type="NOM" lemma="accident">accident</w>
  <w type="PRP" lemma="de">de</w>
  <w type="NOM" lemma="personne">personnes</w>
  <w type="ADV" lemma="ne">n'</w>
  <w type="VER:pres" lemma="être">est</w>
  <w type="PRP" lemma="à">à</w>
  <w type="VER:infi" lemma="déplorer">déplorer</w>
  <c type="SENT">.</c>
 </s>
</p>
```

.. et encore un...

```
<p>
 <persName ref="#TRIC01">Le général Tricoche</persName>, directeur
de l'artillerie au <orgName key="MinG">ministère de la
guerre</orgName>, sera nommé général de
division et chargé des l'organisation des batteries
d'artillerie de forteresse, avec le titre d'inspecteur
général.
</p>
<p>Il sera remplacé dans ses fonctions au
<rs key="MinG">ministère</rs> par <persName ref="#LAVOC32">le
général
   Lavocat</persName>.</p>
<!-- ... -->
<person xml:id="LAVOC32">
 <persName>
  <forename>Jean-Louis</forename>
  <surname>Lavocat</surname>
 </persName>
 <birth when="1823-02-09">ne le 9 fevrier 1823 a Besancon</birth>
<!-- ...-->
</person>
<relation type="remplacement" active="#LAVOC32" pas-
sive="#TRIC01"/>
```

...et encore ...

```
<teiHeader>
 <fileDesc>
  <titleStmt>
   <title>Le Petit Comtois:  TEI-XML Edition</title>
  </titleStmt>
  <publicationStmt>
   <p>Prepared for MISAT</p>
  </publicationStmt>
  <sourceDesc>
   <bibl>
    <title>Le Petit Comtois:  Journal Républicain
          démocratique quotidien </title>
   </bibl>
  </sourceDesc>
  <encodingDesc>
   <classDecl>
    <taxonomy xml:id="notice-class">
     <category xml:id="DDN">
      <catDesc>Dépêches de nuit</catDesc>
     </category>
     <category xml:id="CR">
      <catDesc>Chroniques régionales</catDesc>
     </category>
<!-- ...-->
    </taxonomy>
   </classDecl>
```

Qu'est-ce que c'est que la TEI?

- un labyrinth trop complexe pour les humanistes?
- un système fasciste imposant des normes informatiques à ceux qui n'en ont pas besoin?
- un machin des bibliothécaires numérisés?
- un bibelot inutile qui sort du TAL ?
- une manière de gérer les trucs informatisés vraiment démodée quoi, puisqu'on a tout sur google...

Qu'est-ce que c'est que la TEI?

- un labyrinth trop complexe pour les humanistes?
- un système fasciste imposant des normes informatiques à ceux qui n'en ont pas besoin?
- un machin des bibliothécaires numérisés?
- un bibelot inutile qui sort du TAL ?
- une manière de gérer les trucs informatisés vraiment démodée quoi, puisqu'on a tout sur google...

Qu'est-ce que c'est que la TEI?

- un labyrinth trop complexe pour les humanistes?
- un système fasciste imposant des normes informatiques à ceux qui n'en ont pas besoin?
- un machin des bibliothécaires numérisés?
- un bibelot inutile qui sort du TAL ?
- une manière de gérer les trucs informatisés vraiment démodée quoi, puisqu'on a tout sur google...

Qu'est-ce que c'est que la TEI?

- un labyrinth trop complexe pour les humanistes?
- un système fasciste imposant des normes informatiques à ceux qui n'en ont pas besoin?
- un machin des bibliothécaires numérisés?
- un bibelot inutile qui sort du TAL ?
- une manière de gérer les trucs informatisés vraiment démodée quoi, puisqu'on a tout sur google...

Qu'est-ce que c'est que la TEI?

- un labyrinth trop complexe pour les humanistes?
- un système fasciste imposant des normes informatiques à ceux qui n'en ont pas besoin?
- un machin des bibliothécaires numérisés?
- un bibelot inutile qui sort du TAL ?
- une manière de gérer les trucs informatisés vraiment démodée quoi, puisqu'on a tout sur google...

Les enjeux de la TEI

"Text Encoding for Interchange"

- faciliter la création, l'échange, et l'intégration des données textuelles informatisées
  - toute sorte de texte
  - toutes les langues
  - toute origine temporelle ou culturelle
- La TEI s'adresse également ...
  - aux débutants, cherchant des solutions bien connues et consensuelles
  - aux experts, cherchant à créer de nouvelles solutions

Les buts de la TEI

- faire des recommandations qui se basent sur un consensus existant
- préférer les solutions générales à celles spécifiques à une discipline
- en même temps permettant la spécialisation et l'extension

D'où est sorti la TEI?

- Dans les années 90, c'était un projet de recherche en "digital humanities"...
  - Parainé par trois associations érudites
  - Financé 1990-1994 par NEH, EU LE Programme et SSHRC au Canada
- Influences majeures
  - bibliothèques et archives numerisées
  - ingénierie linguistique
  - édition des sources littéraires ou historiques
- Consortium international établi en 1999 (voir http://www.tei-c.org/)

Qu'est-ce que c'est que la TEI aujourd'hui?

- Un ensemble des *Guidelines* (lignes directrices): peu prescriptives
- représentant un consensus au sujet des distinctions significatives dans un vaste ensemble de matériaux textuels
- qui s'expriment en deux gros volumes de prose et un ensemble de définitions formelles
- ces définitions peuvent être exprimées en plusieurs langues schéma:
  - TEI P1-P3 (1991-1999) : en SGML DTD
  - TEI P4 (2000) : soit en SGML soit en XML DTD
  - TEI P5 (2005-) soit en XML DTD, en W3C Schema, ou en RelaxNG
- derrière cet ensemble se trouve un modèle formel ("conceptual schema") de plus en plus élaboré en système de classes, voire en ontologie

et une communauté internationale active...

Il n'y a pas de “TEI dtd”

- TEI est un système *modulaire*. On s'en sert pour créer un système d'encodage selon ses propres besoins, en sélectionnant des *modules* spécifiques
- Chaque module définit un groupe d'éléments (et leurs attributs)
- on peut sélectionner les éléments voulus, et même en changer des propriétés
- on peut y mélanger des éléments nouveaux, ou bien natifs ou bien d'autres standards

Les Guidelines augmentent le syntaxe avec une semantique

vocabulaire 521 éléments, regroupés en 146 classes

règles d'usage 22 modules, avec 7185 lignes de règles en Relax NG

données contraintes 21 datatypes, plusieurs règles formalisées en schematron

règles de sélection pas formalisées, mais documentées

règles d'utilisation *beaucoup* de prose

règles maison à construire intégralement.

L'envergure TEI

- Structuration basique des textes continus
- Transcription diplomatique, images, multimédia, annotations...
- Données formelles : dates, noms de lieux ou de personnes...
- Données paratextuelles et "meta"
- Analyses linguistiques à tout niveau
- Documentation de balisage
- Et cetera: voir `http://www.tei-c.org/P5/Guidelines/`

Est-il possible d'encyclopédier le balisage?

La TEI est un système modulaire

Pour construire un schéma TEI, il faut spécifier:

- les modules (ensembles de déclarations) dont on a besoin
- des modifications éventuelles
- et cela se fait avec un document TEI.

Modules

- Il existe des modules TEI "core" (noyau):
  - pour l'entête (metadonnées)
  - pour l'infrastructure
  - pour une structuration basique des textes
- qui sont complétés par des modules plus spécialisés p.e.
  - les textes oraux
  - la transcription des sources originales
  - l'édition critique
  - les dictionnaires
  - l'annotation linguistique
  - noms de lieux et de personnes
  - description de manuscrits
  - liens et analyses hypertextuels
  - documentation des systèmes de balisage
  - etc etc

"One Document Does it all (ODD)"

- Les Guidelines et ses schémas sont tous produits à partir d'une même ressource XML qui contient:
  - de la prose descriptive (une grande quantité)
  - des exemples d'utilisations (plusieurs)
  - des déclarations formelles pour les constituants du modèle abstrait de TEI
  - les éléments et leurs attributs
  - les modules
  - les classes, et les macros
- On appelle cette ressource un ODD (bien qu'elle consiste en des centaines de petits fichiers)

Et alors?

- Vu ses visées ambitieuses, on ne peut se servir du système TEI qu'en le personnalisant
- Les personnalisations s'expriment en langue ODD

```
<schemaSpec ident="TEIlite">
<desc>This is TEI Lite with a small change</desc>
 <moduleRef name="tei"/>
 <moduleRef name="linking"/>
 <moduleRef name="structure"/>
 <moduleRef name="teiheader"/>
 <élémentSpec ident="head" mode="change">
 <content><rng:ref name="text"/></content>
```

La personnalisation

- Une balise représente un élément
- Les éléments sont regroupés
  - en modules
  - en classes
- Un système de balisage se fait en combinant
  - des références à module
  - des déclarations de balises additionnelles ou modifiées

L'importance de la personnalisation

- La système TEI comprend 20 modules, quelques centaines d'éléments, et des douzaines de classes
- Un projet typique aura besoin de
  - faire un choix des modules
  - affiner l'ensemble d'éléments obtenus des modules choisis
  - ajouter des contraintes spécifiques aux datatypes
  - (peut-être) ajouter des éléments spécifiques
  - (peut-être) faire des localisations des balises
  - (assurément) produire de la doc spécifique au projet
- Tout cela s'effectue avec un ODD

Workflow: outils

- les outils ne se font pas tout seuls!
- la production des outils
  - évolue d'une manière très darwinienne
  - nécessite un professionnalisme et des compétences informatiques
- d'où l'importance de la standardisation

sciences humaines: les puces sur l'éléphant...

L'outillage TEI

Une des raisons fortes pour lesquelles se servir de la TEI est l'existence des outils TEI p.e.

roma	http://www.tei-c.org/Roma permet de construire des schemas TEI
tei-vesta	conversion TEI/docx/html etc.
XAIRA	moteur de recherche xml

Une autre est la possibilite de se servir de n'importe quelle outillage XML -- parce que la TEI est 100% standard XML!

La boite à outils TEI-XML

Pour maîtriser les documents XML, on a besoin de les

- créer, modifier, valider...
- transformer, visualiser...
- rechercher, analyser...
- stocker, gérer, trouver ...

Création, modification, validation

peut s'effectuer avec

- un éditeur ascii classique (pe emacs)
- un éditeur XML spécialisé (pe Oxygen)
- un système bureautique (pe OpenOffice)

Pour la validation, il faut un schéma. Roma est un outil pour construire facilement des schémas TEI.

Transformation et visualisation

Opération fondamentale, s'appuyant sur des standards:

- XSLT (Extensible Stylesheet Language) et XPath
- CSS (Cascading Stylesheets)

Fonctionnement intégré dans les navigateurs, ou autres logiciels

Recherches et analyses

- peuvent bien s'effectuer avec des outils de transformation
- ou bien avec des logiciels spécialisés qui s'y appuient
- XQuery: standard langue de requête, complémentaire de XPath

L'analyse textuelle pose des problèmes intéressants:

- les données non-structurées
- une richesse de métadonnées
- des structures poly-arborescentes

XAIRA est un outil libre permettant l'indexation et recherche des corpus XML

Stockage, gestion, mise a jour

A quel niveau ? On pourrait:

- stocker les éléments (ou autres fragments) dans un SBDX et reconstituer les documents sur demande, comme p.e. la plupart des CMS, wiki, etc.
- stocker les documents (ou fragments de document) eux-mêmes

Systèmes courants pour gérer et distribuer les documents XML:

- eXist
- Fedora
- XTF

L'importance des métadonnées...

Open TEI

- La TEI n'est plus un projet de recherche
- Ses outputs sont distribués sous licence GNU
- Sa gestion et son évolution se font en public, sur Sourceforge à http://tei.sf.net
- Tous ses composants sont disponibles sous Debian Linux (etc.)
- Pourtant, la mention "TEI" reste protegée, et les évolutions techniques sont toujours faites avec l'autorisation d'un Conseil Technique, élu par les membres du Consortium.

Open TEI: ca veut dire quoi?

- La TEI reste une initiative communautaire, gérée par et pour les membres de cette communauté.
- Comme le Open Source Software (libre), elle s'ouvre à la communauté mondiale des développeurs
- L'inscription à la TEI reste et restera facultative – ce qu'il nous faut, c'est de la participation.
- This means you!

Consortium web site `http://www.tei-c.org`

Sourceforge repository `http://tei.sf.net`

Quelques références francophones:

- tei-fr@listserv.inist.fr
- http://lespetitescases.net/index102/
- http://www.culture.gouv.fr/culture/dglf/riofi/tei.htm"
- http://artist.inist.fr/article.php3?id_article=122"/>