

Initiation XML-TEI
Initiation à l'encodage des textes patrimoniaux
Stage XML/TEI
CESR Tours – 18 janvier 2011



Initiation à l'encodage des textes patrimoniaux

Formation XML/TEI

Lou Burnard

lou.burnard@tge-adonis.fr

Objectifs de cette presentation

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

La numérisation nous apporte de nouveaux défis!

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

- construire une base de données mutualisée, des instruments de recherche (*finding aid*)
- intégrer de tels instruments avec les textes qu'ils signalent
- intégrer de tels instruments dans une espèce de mère porteuse numérique, (*edition numérique*)
- donner support aux outils d'analyse complexe ('text-mining') distribués

La TEI peut nous aider...

Elle représente un modèle conceptuelle bien établie et consensuelle qui facilite alors

- la conversion des données existantes
- la création des données nouvelles
- l'intégration des données déjà existantes mais répandues dans plusieurs sources

Elle est basée sur des formats ouverts et des technologies ouvertes

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

Est-ce que ceux-ci represente la meme chose ?

A MONSEI-

GNEVR LE REVE-
rendissime Cardinal
du Bellay.

S.



EV le Personnage,
que tu ioues au Spectacle
de toute l'Europe,
voyre de tout le Monde
en ce grand Theatre
Romain, veu tant
d'affaires, & telz, que
seul quasi tu soutiens: ô
l'Honneur du sacré Col-

lege! pecheroy'-ie pas (comme dit le Pindare
Latin) contre le bien publicq', si par longues
paroles t'empeschoy' le tens, que tu donnes au
seruice de ton Prince, au profit de la Patrie, &
à l'accroissement de ton immortelle renommée?
Epiant donques quelque heure de ce peu de re-
laisz, que tu prens pour respirer soubz le pesant
faiz des affaires Francoyses (charge urayement
digne de si robustes epaules, non moins que le
Ciel de celles du grand Hercule) ma Muse a pris
la hardiesse d'ètrer au sacré Cabinet de tes sain-
ctes, & studieuses occupations: & la entre tant

a ij de

Joachim du Bellay
Défense et illustration de la
langue françoise (1549)

La Deffence, et illustration de la Langue françoise

L'auteur prie les lecteurs différer leur Jugement Jusques à la fin du livre, et ne le
condemner sans avoir premierement bien vu, et examiné ses raisons.

Epître à Monseigneur le révérendissime cardinal du Bellay S.

Veü le personnage que tu joues au spectacle de toute l'Europe, voire de tout le monde, en ce grand
Théâtre Romain, veü tant d'affaires, et telz que seul quasi tu soutiens, ô l'honneur du sacré Collège,
pecheroy-je pas (comme dit le Pindare Latin) contre le bien public, et par longues paroles t'empeschois le
service que tu donnes au service de ton Prince, au profit de la patrie et à l'accroissement de ton
immortelle renommée? Epiant doncques quelque heure de ce peu de relai que tu prens pour respirer
sous le pesant fais des affaires françoises (charge vraiment digne de si robustes epaules, non moins
que le ciel de celles du grand Hercule), ma Muse a pris la hardiesse d'entrer au sacré cabinet de tes
sainctes et studieuses occupations: et là, entre tant de riches et excellentes veues de jour en jour dédiés
à l'honneur de ta grandeur, nostre le tien humble et naïf, mais te'eloier bien heurieux, est venu entre

A MONSEIGNEUR

Le Révérendissime Cardinal du Bellay, S.

Veü le personnage que tu joues au spectacle de
toute l'Europe, voire de tout le monde, en ce grand
theatre romain; veü tant d'affaires et telz, que seul
quasi tu soutiens: ô l'honneur du sacré Collège! pe-
cheroy'-je pas (comme dit le Pindare latin) contre le
bien publicq', si par longues paroles t'empeschoy' le
tens que tu donnes au service de ton Prince, au profit
de la patrie, et à l'accroissement de ton immortelle
renommée? Epiant donques quelque heure de ce peu
de relai, que tu prens pour respirer soubz le pesant
faiz des affaires francoyses (charge vraiment digne
de si robustes epaules, non moins que le ciel de celle
du grand Hercule), ma Muse a pris la hardiesse d'en-
trer au sacré cabinet de tes saintes et studieuses oc-



Un texte n'est pas un document...

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'une communauté de lecteurs.

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

Un texte n'est pas un document...

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'une communauté de lecteurs.

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

Un texte n'est pas un document...

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'une communauté de lecteurs.

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

Un texte n'est pas un document...

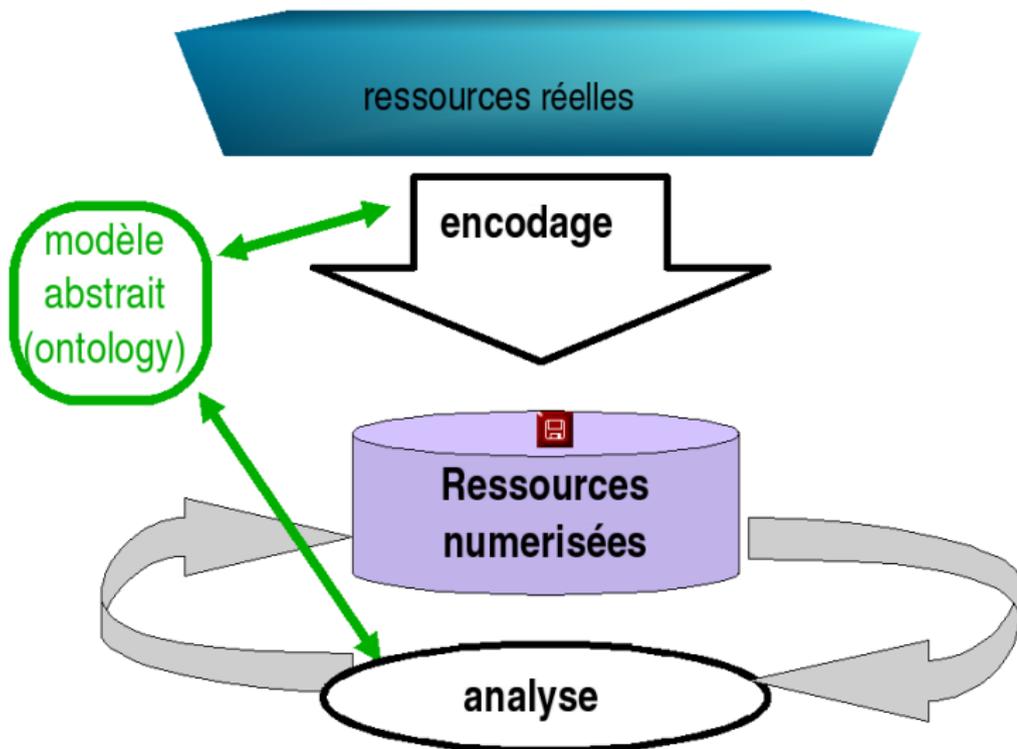
En quoi consiste l'essentiel d'un texte ?

- 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'une 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!
- Un text est plus qu'une séquence de formes lexicaux!
 - 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...

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

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

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
```

- Bonne nouvelle: il existe des logiciels capables de traduire entre 500 formats divers
- Mauvaise nouvelle: on en a besoin

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
```

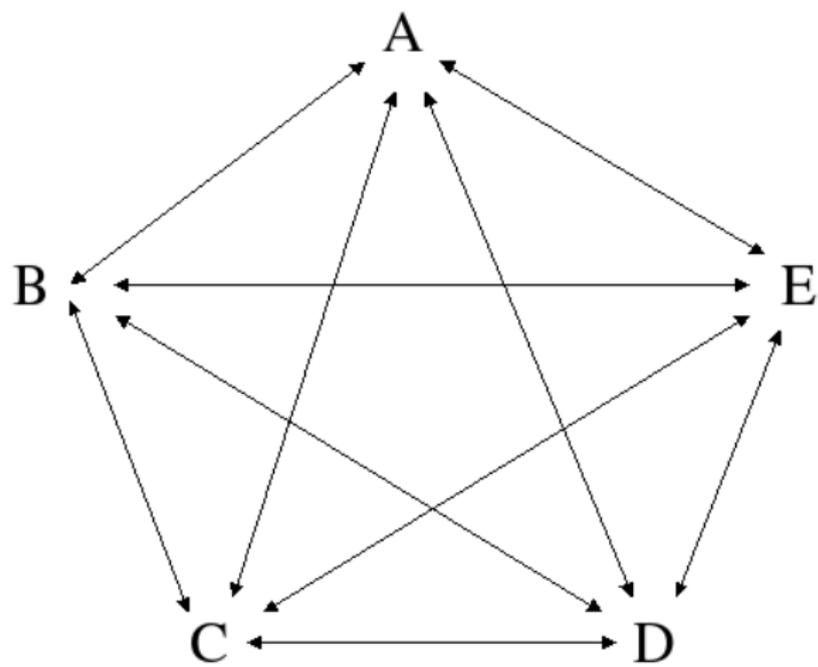
- Bonne nouvelle: il existe des logiciels capables de traduire entre 500 formats divers
- Mauvaise nouvelle: on en a besoin

Encodage ou babel?

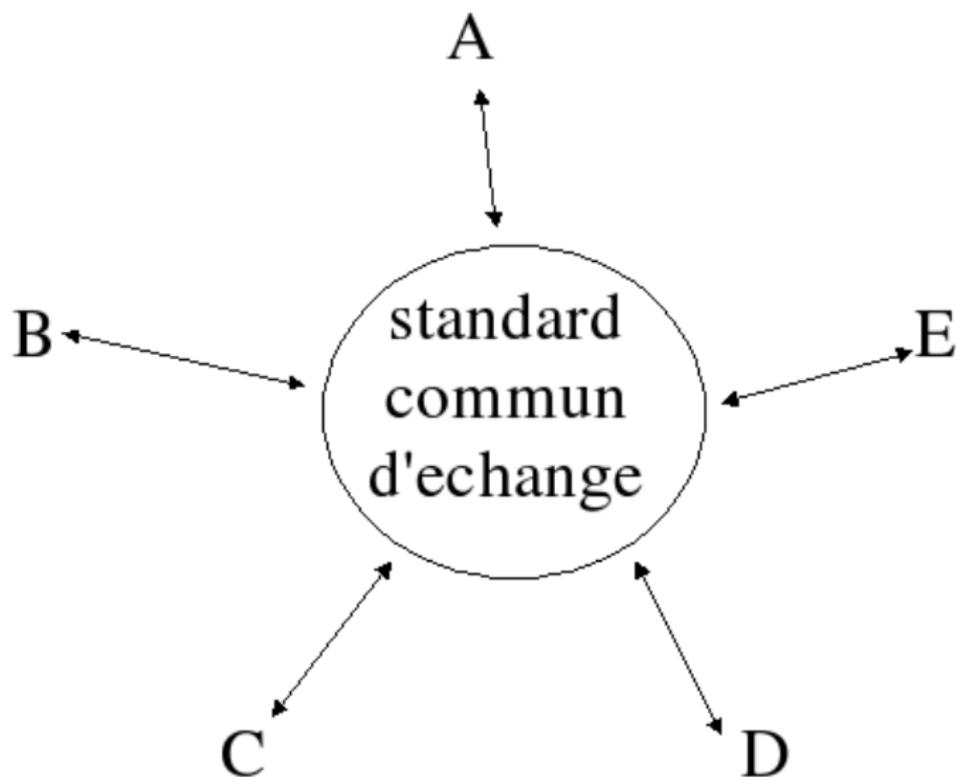
```
|chap1
<C 1> Loomings
\chapter[1]{Loomings}
:h1.1. Loomings
MOBY001001LOOMINGS
|C1
.chapter Loomings
.cp;.sp 6 a;.ce .bd 1. Loomings
<h1>Loomings</h1>
<p class="h1">Loomings
~x
```

- Bonne nouvelle: il existe des logiciels capables de traduire entre 500 formats divers
- Mauvaise nouvelle: on en a besoin

Echange d'informations (1)



Echange d'informations (2)



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 en forme
- 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 en forme
- 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 en forme
- 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 en forme
- 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 Personnage,  
<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 Personnage, que tu joues au  
    Spectacle de toute l'Europe...  
  </p>...  
</div>
```

... et avec

```
<pb n="4"/>
<s>
  <w pos="PPJ" lemma="voir">VEU</w>
  <w pos="ART" lemma="le">le</w>
  <w pos="SBC" lemma="personnage">Personnage</w>
  <pc>,</pc>
  <w pos="COO" lemma="que">que</w>
  ...
</s>
```

ou bien

```
<s>
  <choice>
    <reg>Vu</reg>
    <orig>Veu</orig>
  </choice>
  le <choice>
    <reg>Personnage</reg>
    <orig>Personnage</orig>
  </choice>,
  que tu joues au Spectacle...
</s>
```

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 connaître

- XML est une manière de représenter 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 seul transparent

- Un document XML contient au moins un *élément*
- Un élément possède une *balise d'ouverture*, facultativement de *contenu* et une *balise de fermeture*
- Un élément peut d'ailleurs porter des *attributs*, chacun portant un *nom* et une *valeur*
- Un document XML est *obligatoirement* 'well formed' (bien-formé) i.e. il doit suivre la syntaxe XML
- Un document bien-formé peut *facultativement* être *valide* i.e. il est conforme aux règles d'une *schéma* 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 traitement*
- 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 seule racine contienne le document entier
- 2 chaque arborescence soit proprement imbriquée
- 3 tous 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>`

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



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

- un labyrinthe 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 labyrinthe 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 labyrinthe 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 labyrinthe 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 labyrinthe 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 labyrinthe 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 numérisées
 - 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

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/>

... un encyclopédie du balisage

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 -- un vocabulaire XML aussi défini par la TEI -- qui permet 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

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

"Garage" services web pour conversion TEI/docx/html etc.

xaira moteur de recherche xml

TEI publisher systeme d'édition web

Open Text systeme d'édition de revues.org

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

Des références...

Site web du Consortium TEI <http://www.tei-c.org>

Depot sourceforge <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
- <http://www.mutec-shs.fr/la-tei-en-france-pratiques-et-perspectives>