Initiation à l'encodage XML-TEI

Lou Burnard

Objectifs de cette formation

1. Préciser ce que c'est que l'encodage textuel
2. Présenter les concepts fondamentaux de TEI-XML
3. Offrir beaucoup d' experimentation pratique avec les outils 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 données 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

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

A MONSEI-
GNEVR LE REVE-
rendiſſime Cardinal
du Bellay.
S.

VEV 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 ſeruice de ton Prince, au profit de la Patrie, & à l'accroiſſement de ton immortelle renommée? Epiant donques quelque heure de ce peu de relaiz, 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'ẽtrer au ſacré Cabinet de tes ſainctes, & ſtudieuſes occupations: & la entre tant

A MONSEIGNEUR

Le Reverendiſſime Cardinal du Bellay, S.

Veu le personnaige que tu joues au spectacle de toute l'Europe, voyre de tout le monde, en ce grand theatre romain; veu tant d'affaires et telz, que seul quasi tu soutiens : ô l'honneur du sacré College! pecheroy'-je pas (comme dit le Pindare latin) contre le bien publicq', si par longues paroles j'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 relaiz, que tu prens pour respirer soubz le pesant faiz des affaires francoyses (charge vrayement digne de si robustes epaules, non moins que le ciel de celle du grand Hercule), ma Muse a pris la hardiesse d'entrer 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'un 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'un 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'un communauté de lecteurs.

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

Un texte n'est pas un document...

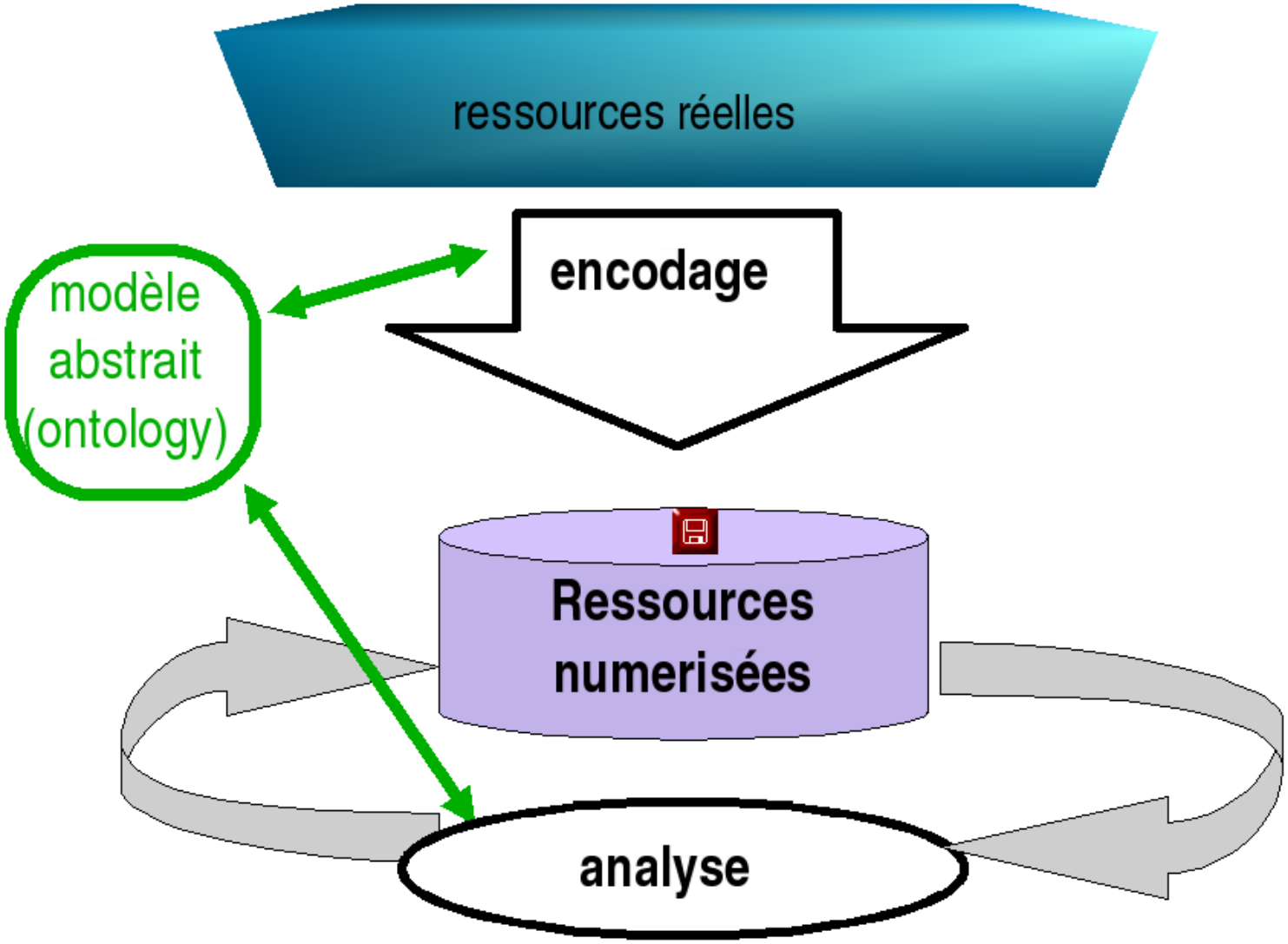
En quoi consiste l'essentiel d'un texte ?

- 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'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?

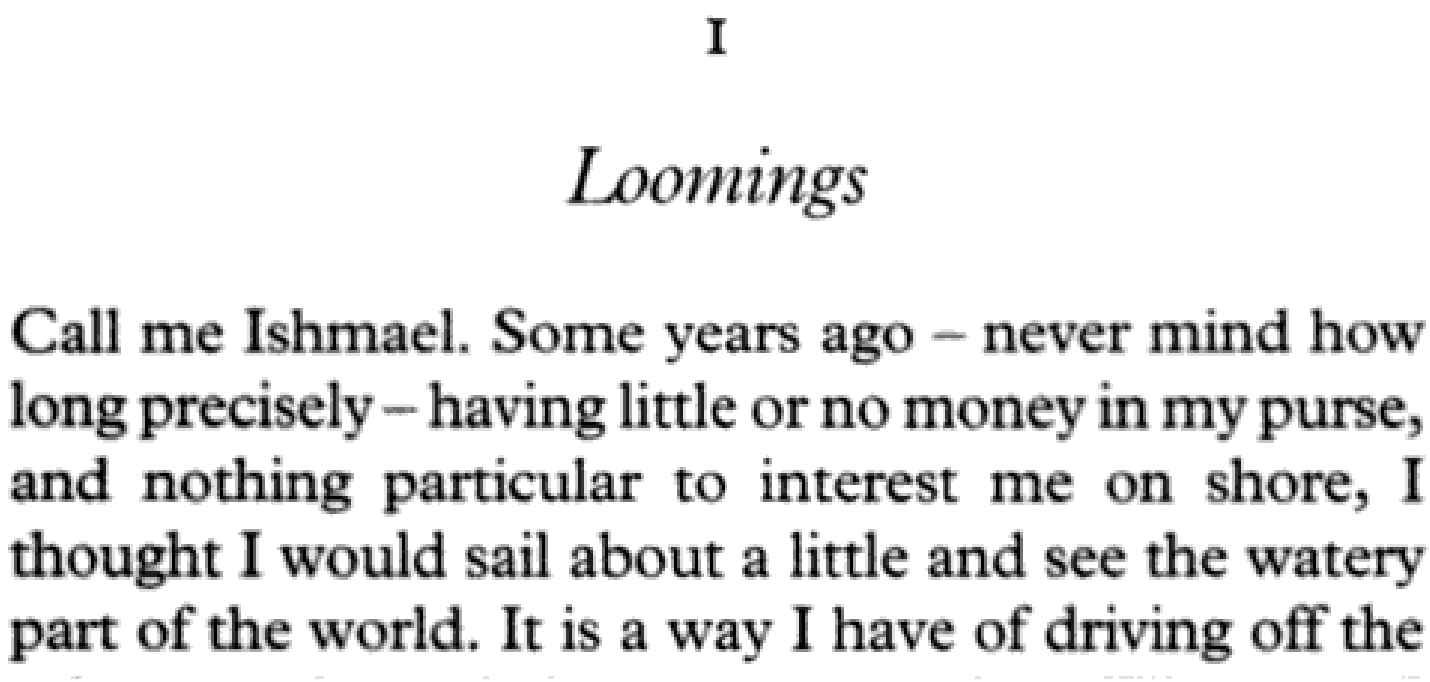


L'encodage

- Un texte est plus qu'une séquence de caractères encodés!
- 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 (malheureseument) 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
```

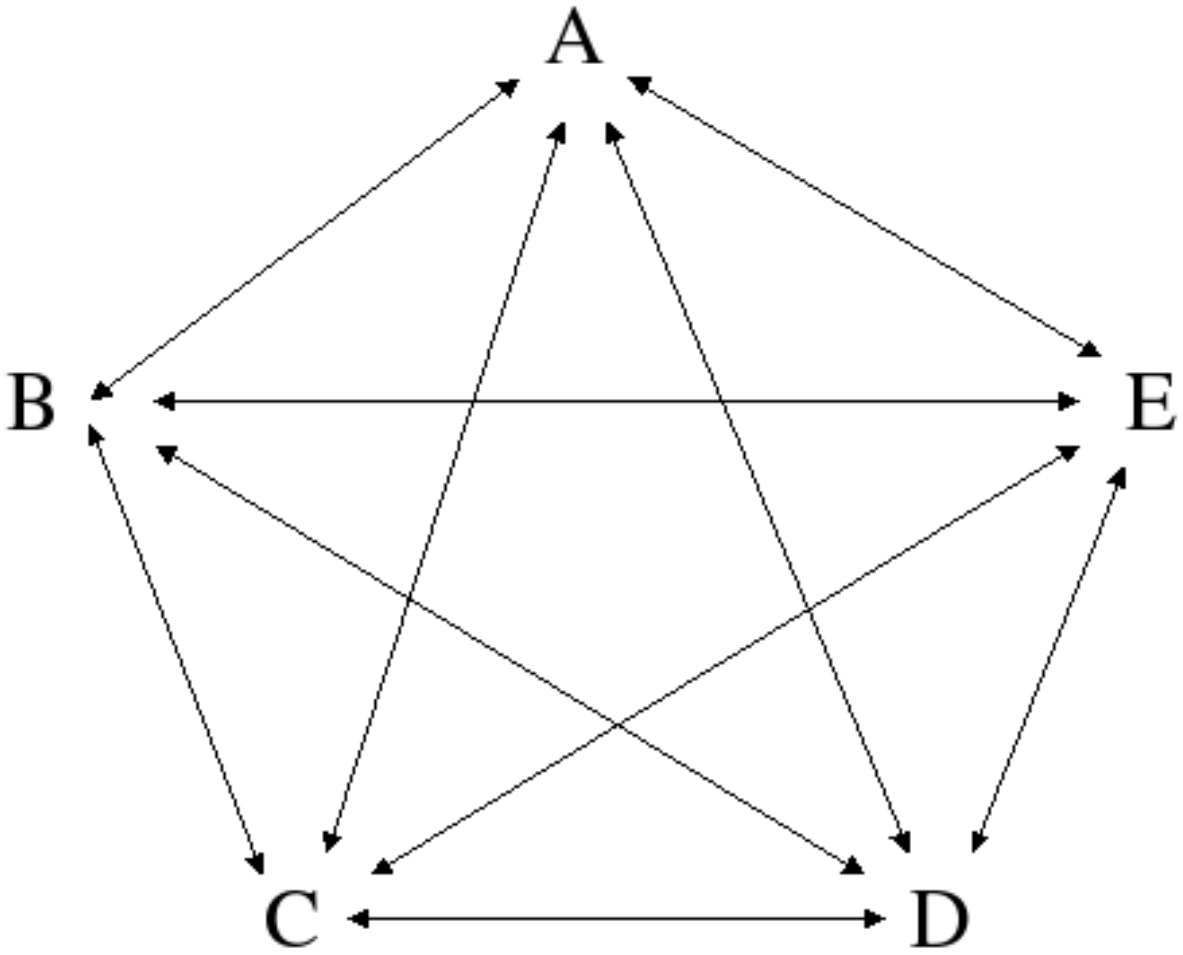
- Bonne nouvelle: il existe des logiciels capables de traduire entre 500 formats divers
- Mauvaise nouvelle: on en a besoin

Encodage ou babel?

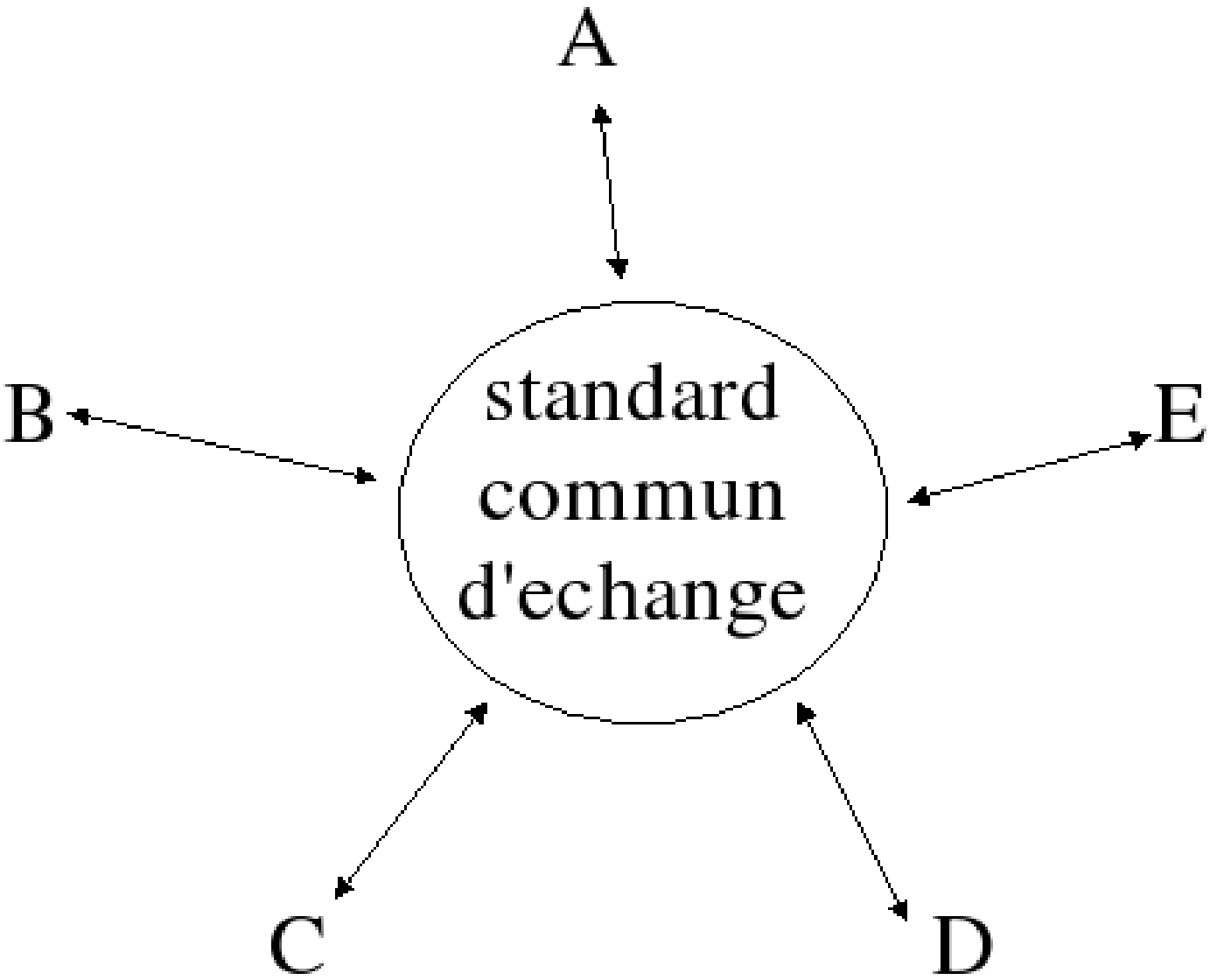
```
|chap1
<C 1> Loomings
\chapter[1]{Loomings}
:h1.1.  Loomings
MOBY001001LOOMINGS
|C1
.chapter Loomings
.cp;.sp 6 a;.ce .bd 1. Loomings
<h1>Loomings</h1>
<p class="h1">Loomings
~x
```

- 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 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>
```

... 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">Personnaige</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>Personnaige</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 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 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* etre *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 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>`

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>`

XML est un standard international

- Un document XML doit se servir du standard ISO 10646 (aka Unicode)
  - un répertoire de caractères 31-bit adéquate à la plupart des systèmes d'écriture humaine
  - encodé en deux formats UTF8 ou UTF16
- un document peut spécifier qu'il contient les mêmes caractères encodés d'une autre manière (notamment ISO 8859)
- un élément peut spécifier le langage de sa contenue avec l'attribut prédéfini *@xml:lang*

L'attribut *@xml:id* est également prédéfini par le W3C.

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