

# Initiation à l'encodage XML-TEI

Lou Burnard

# Objectif de cette session

Répondre à ces questions :

- ❶ Qu'est-ce un texte? qu'est-ce un document ?
- ❷ Qu'est-ce le balisage ?
- ❸ C'est quoi le balisage XML? Pourquoi est-il devenu si important?

## La tournée numérique

Les sciences humaines et sociales s'occupent surtout au texte ...

- (majoritairement non-numérique) les livres, les manuscrits, les fonds d'archive ...
- ainsi que d'autres manifestations culturelles/communicatives (de plus en plus numérisées) par ex., les sons, les images, les cahiers de recherches, les tweets

Les 'digital humanities' se préoccupent avec les outils et les techniques qui permettent de manipuler d'une manière intégrée toutes ces manifestations, et donc de gérer cette patrimoine de plus en plus signifiante.

Le balisage (markup, encodage) est une composant incontournable de ces manipulations

## Texte et texte numérique

Un texte peut être considéré selon trois axes :

- Un texte a une existence physique, ayant des **traits visuels** qu'on peut (plus ou moins) transférer automatiquement d'une instance à une autre
- Un texte possède des propriétés linguistiques et structurelles, qu'on ne peut transcrire, traduire, ou transmettre qu'avec une compréhension humaine
- Un texte présente des informations sur le monde réel, qu'on peut comprendre (ou non) ou annoter, et qui nous permet de générer de nouveaux textes

Un balisage effectif devrait donc opérer dans tous ses trois axes.

## Par exemple ...

Diary of Robert Graves 1935-39 and ancillary material

Copyright St John's College Robert Graves Trust

July 27 Saturday

Gr1-156

July 27 Saturday

Got up at 11.30. Rosa came. ~~then~~  
~~then~~ Worked at inserts of Richards's  
first chapter. Laura had a talk with  
Carl about Department.  
She slept until 5 (I working on  
Richards). I went to ~~John~~ Marguerite  
to order my grey americans, & to  
Pascada to open windows (shutting shutters)  
turn out lights, take out away  
friskettes.

Then worked at Jordan's life, after L.  
to 1.5. it Carl brought melon,

# Exemple continue ...

## Diary of Robert Graves 1935-39 and ancillary material

Copyright St John's College Robert Graves Trust

New Search

Diary Scans

[« Return to Search Results](#)

### July 1935

[« June](#)   [« Abstract »](#)   [August »](#)

SUN	MON	TUE	WED	THU	FRI	SAT
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

#### DISPLAYED DIARY SCAN(S)



July 27 Saturday

[» Annotated markup](#)

[» Full-sized Image](#)

[» Gallery Scan](#)

### July 27 Saturday

Got up at 11.30. **Rosa** came. ~~xxxxx~~[crossed out]

~~xxxxx~~[crossed out] Worked at inserts of **Richards'** first chapter. **Laura** had a talk with **Carl** about deployment.

She slept until 5 (I working on **Richards**<sup>1</sup>). I went to **Fábrica**<sup>[RG]</sup> **Margarita** to order my grey *americano* <sup>2</sup>, & to **Posada** to open windows (shutting shutters) turn out lights, take <sup>[RG]</sup> away perishables.

Then worked at **Gordon's Life**<sup>3</sup>, after **L.** went over it. **Carl** brought melon, & we had coffee ice. **Laura's** stomach bad. I went to **Fábrica** & recovered her parasol & fan from *camión* <sup>4</sup>. More work on **Gordon**<sup>5</sup>. Bed at 12.

**Gelat** expects no result of law suit for two months. **Concordia** ceiling finished: tiles green & yellow, being laid diagonally.

#### EDITORIAL NOTES

<sup>1</sup> **Old Soldier Sahib**, eds.

<sup>2</sup> Spanish [slang?] for "jacket" KG; KG also replaces the final "o" with "a". eds.

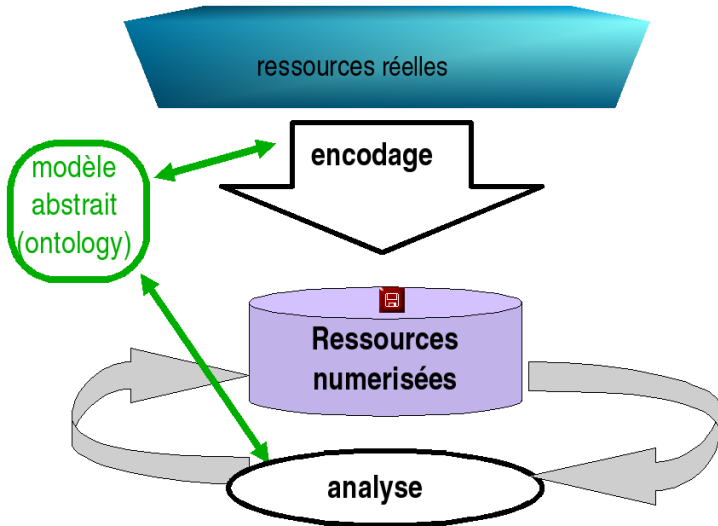
<sup>3</sup> See **A Mistake Somewhere**, eds

<sup>4</sup> bus. KG

<sup>5</sup> i.e. Gordon's autobiography: see above. eds.

une re-présentation du contenu linguistique, et structure, avec annotations sur sa portée, son contexte..

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



## La TEI pourrait nous aider...

Elle représente une modèle conceptuelle de texte bien établie qui facilite :

- 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 représente la même chose ?

A MONSEI<sup>R</sup>.

GNEVR LE REVE-  
rendissime Cardinal  
du Bellay.

S.



EV le Personnage,  
que tu ioues au Spectacle  
de toute l'Europe,  
uoire de tout le Monde  
en ce grand Theatre  
Romain, ueu 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 j'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-  
laiz, 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'entrer au sacré Cabinet de tes sain-  
tes, & studieuses occupations: & la entre tant

a ij de

A MONSEIGNEUR

*Le Reverendissime Cardinal du Bellay, S.*

Veu le personnage que tu joues au spectacle de  
toute l'Europe, uoie de tout le monde, en ce grand  
theatre romain; veu tant d'affaires de telz, que seul  
quasi tu soutiens: ô l'honneur du sacré College! pe-  
cheroy'-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'en-  
trer au sacré cabinet de tes saintes et studieuses oc-

## Et ceux-ci ?

# A MONSEI-

NEVR LE REVE-  
rendissime Cardinal  
du Bellay.

S.



EV le Personnage,  
que tu ioues au Specta-  
cle de toute l'Europe,  
uoire de tout le Mon-  
de en ce grand Thea-  
tre Romain, veu tant  
d'affaires, & telz, que  
seul quasi tu soutiens: ô  
l'honneur du sacré Collè-

lege! pecheroy-je 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 doncques quelque heure de ce peu de re-  
lais, que tu prens pour respirer soubz le pesant  
faix 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'entrer au sacré Cabinet de tes saintes,  
& studieuses occupations: & la entre tant  
a ij de

A MONSIEUR

*Le Reverendissime Cardinal du Bellay, S.*

Veux le personnage que tu joues au spectacle de  
toute l'Europe, voire de tout le monde, en ce grand  
theatre romain: veu tant d'affaires et telz. que seul



**Joachim du Bellay**

**Défense et illustration de la  
langue françoise (1549)**



**La Défense, et illustration de la Langue Francoise**

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, pêcherai-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 saintes et studieuses occupations : et là, entre tant  
de riches et excellents vœux de jour en jour dédiés à l'image de ta grandeur, pendre le sien humble et petit,  
mais toutefois bien heureux s'il rencontre quelque faveur devant les yeux de ta bonté, semblable à celle des  
Dieux immortels, qui n'ont moins agréables les pauvres présents d'un riche vouloir que les superbes et  
ambitieuses offrandes.

# 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 (supposée) de cette copie?
- en les interprétations/lectures apportées ou trouvées?
- en les intentions (supposées) de son auteur?

Un "document" est une chose physique, que nous pouvons numériser.

Un "texte" est une abstraction, construite par ou pour une communauté de lecteurs, que nous pouvons encoder.

# 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 (supposée) de cette copie?
- en les interprétations/lectures apportées ou trouvées?
- en les intentions (supposées) de son auteur?

Un "document" est une chose physique, que nous pouvons numériser.

Un "texte" est une abstraction, construite par ou pour une communauté de lecteurs, que nous pouvons encoder.

# 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 (supposée) de cette copie?
- en les interprétations/lectures apportées ou trouvées?
- en les intentions (supposées) de son auteur?

Un "document" est une chose physique, que nous pouvons numériser.

Un "texte" est une abstraction, construite par ou pour une communauté de lecteurs, que nous pouvons encoder.

# 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 (supposée) de cette copie?
- en les interprétations/lectures apportées ou trouvées?
- en les intentions (supposées) de son auteur?

Un "document" est une chose physique, que nous pouvons numériser.

Un "texte" est une abstraction, construite par ou pour une communauté de lecteurs, que nous pouvons encoder.

## 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 (supposée) de cette copie?
- en les interprétations/lectures apportées ou trouvées?
- en les intentions (supposées) de son auteur?

Un "document" est une chose physique, que nous pouvons numériser.

Un "texte" est une abstraction, construite par ou pour une communauté de lecteurs, que nous pouvons encoder.

# 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
```

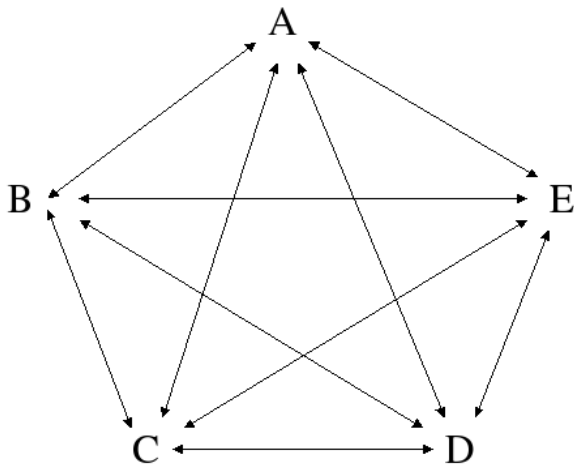
- Bonne nouvelle: il existe des logiciels capables de traduire entre 500 formats divers
- Mauvaise nouvelle: on en a besoin

## 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
```

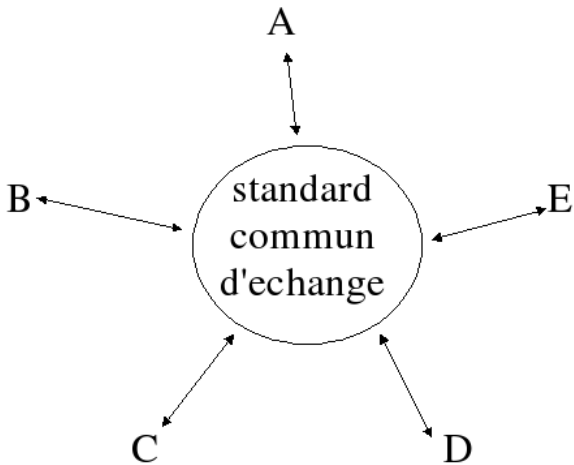
- Bonne nouvelle: il existe des logiciels capables de traduire entre 500 formats divers
- Mauvaise nouvelle: on en a besoin

## Echange d'informations (1)



sans format pivot: 20 passerelles requises ( $n \times n - n$ )

## Echange d'informations (2)



avec format pivot: 10 passerelles requises ( $2n$ )

# 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? La réponse dépend des usages prévus...

# 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
- Un balisage presentationnel par contre s'intéresse plus à l'affichage qu'à sa sémantique



# 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
- Un balisage presentationnel par contre s'intéresse plus à l'affichage qu'à sa sémantique

# 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
- Un balisage presentationnel par contre s'intéresse plus à l'affichage qu'à sa sémantique

# 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
- Un balisage presentationnel par contre s'intéresse plus à l'affichage qu'à sa sémantique

# 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
- Un balisage presentationnel par contre s'intéresse plus à l'affichage qu'à sa sémantique

## 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">Personnaige</w>
  <pc>,</pc>
  <w pos="C00" 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>
```

## Notre systeme d'encodage devrait etre capable de..

- 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.

Jusqu'a present, XML semble une bonne solution ...

# 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'un schéma quelconque

```
<?xml version="1.0" ?>
<root>
<element attribute="value"> content </element>
<!-- comment -->
</root>
```

## 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">
```

# 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 paires nom/valeurs qui constituent les attributs d'un élément peuvent apparaître sans ordre à l'intérieur d'une balise ouvrante
- Les caractères < et & sont Magiques et doivent être cachés au moyen de références entité (&lt; et &amp; respectivement)
- L'espace de noms auquel appartient un élément peut être signalé par un **namespace-prefix** (p.e. xml:) prédéfini
- Les commentaires sont délimités par <- - et - ->
- Les références entité sont délimités par & et ;
- Les sections CDATA sont délimités par <![CDATA[ et ]]>

## 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>`

## XML est un standard international

- Un document XML doit respecter le 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 contenu avec l'attribut prédéfini @xml:lang

Un 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

Avec un schéma, on peut spécifier :

- les nom des éléments qui constituer la racine d'un document
- les noms de tous les éléments légaux
- les noms, les types, et les valeurs par défaut de tous les attributs
- des règles concernant l'imbrication et le contenu des éléments
- et quelques autres menus propos...

Un schéma donc vous permet de contrôler par exemple que 'tout chapitre ait son titre', que 'tout recette comporte une liste des ingrédients', que 'la valeur de tout attribut @when soit conforme au standard ISO' ... etc.

Un espace de noms, par contraste, ne vous permet que de labelliser le vocabulaire d'où est dérivé un ensemble d'éléments.

# Langues de schéma

Un schéma est exprimé dans une langue formelle. Actuellement, on peut choisir entre :

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

La TEI se sert de Relax NG



## Exercice 1

D'abord, nous allons expérimenter un logiciel spécialisé pour créer et modifier des fichiers XML...