

# Principes et enjeux du balisage

Lou Burnard

## Texte et texte numérique

Un texte peut être considéré selon trois axes :

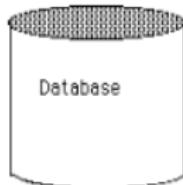
- Un texte a une existence physique, ayant des **traits visuels** qu'on peut (plus ou moins) transférer automatiquement d'une instance à une autre
- Un texte possède des **propriétés linguistiques et structurelles**, qu'on ne peut transcrire, traduire, ou transmettre qu'avec une compréhension humaine
- 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 ces trois axes.

# Traitements numériques du texte



- \* word processing
- \* indexing
- \* database



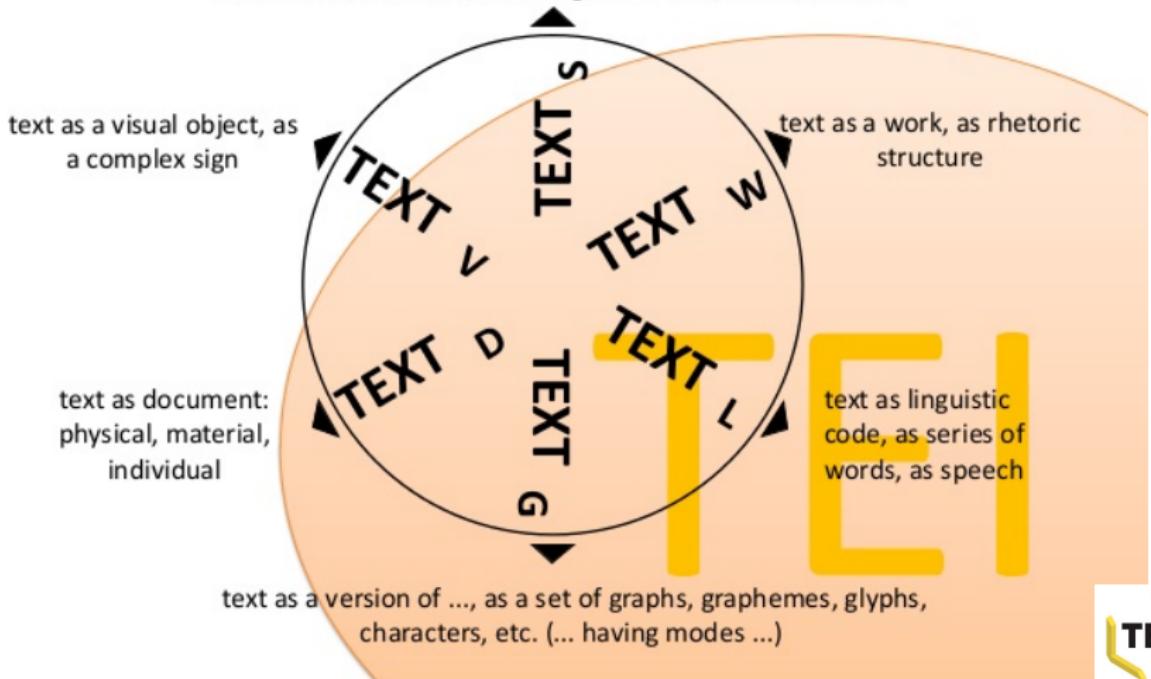
aardvark  
abacus  
abracadabra  
acidulation  
adumbrate  
aetherial  
affection  
agony  
aha  
ai  
ajacanthus

(Or maybe it's more than a trinity)

## The Textwheel

(Patrick Sahle: Digitale Editionsformen, 2013)

text as idea, intention, meaning, semantics, sense, content



## Software families

Existing software systems tend to specialize ...

- document management and production systems
- image management and production systems
- linguistic analysis and management
- database systems

# Convergence

But convergence is now on everyone's digital agenda. When you make a mashup combining

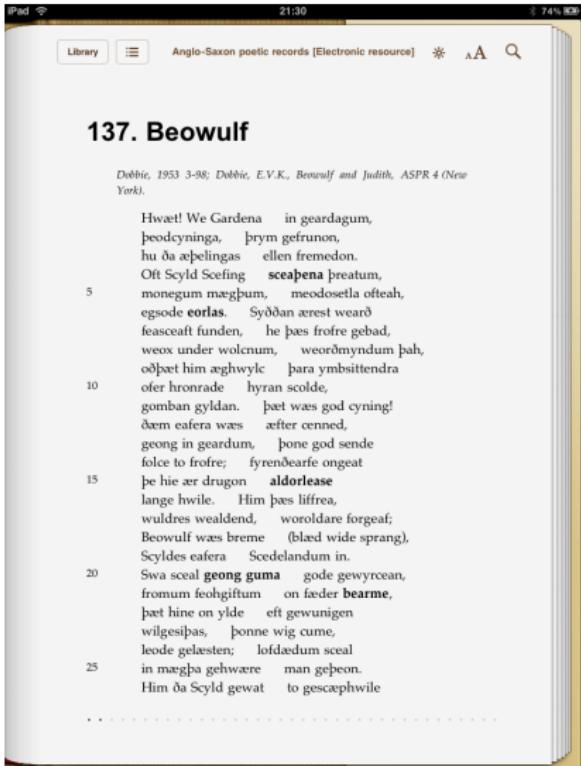
- a GIS database about places in the Aegean sea
- a historical gazeteer of placenames in the same area
- a corpus of texts mentioning those placenames

you need to combine the strengths of a database with tools for linguistic analysis, and with tools for rendering spatial information.

A few examples:

- <https://pleiades.stoa.org/places/109236>
- <http://www.mappingpaintings.org>
- <https://mapoflondon.uvic.ca/map.htm>

# The problem



- Today's digital library applications still focus on serving up virtual pages for the reader: the metaphor of the book is so pervasive that we can barely see it.
- Self-evidently, digitization makes it possible to offer cheaper and more accessible simulations of printed or written pages.
- But this is not enough... digital texts should aim to go 'beyond the page'

## What use is a digital text ?

Digital applications enable us to do more with a text, and especially with a collection of texts!

- more than simply read it from beginning to end
- more than attach annotations to it for others to read,
- more than perform brute-force “text mining” on it.

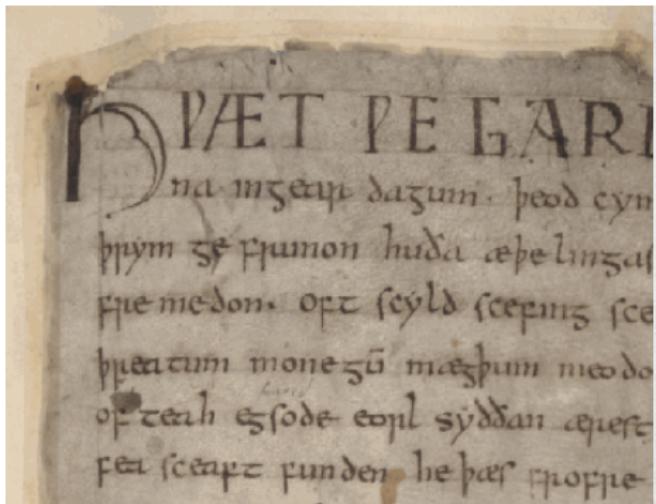
The content of the digital library must therefore be enriched, even if this requires the use of techniques which are not currently automatable.

# What's that noise in the digital library?



- A digital edition should capture the intentions and meaning of a text, not simply its appearance
- Otherwise, there can be no analysis beyond the documentary level, no 'conversation between books'

# Enrichment or Representation?



*When we go from this...*

Hwæt wē Gār-Dena  
þēod-cyninga brym <sup>glory</sup>  
hū ðā æþelingas ellen  
Oft Scyld Scēfing  
5 monegum mægþum  
egsode Eorl[e], - syðða  
fēasceaft funden; hē  
wēox under wolcnum,  
oðþæt him ēghwylc  
10 ofer hron-rāde hýran

*... to this, what is happening?*

## Editing

It's customary to distinguish (at least) these types or levels of interpretation:

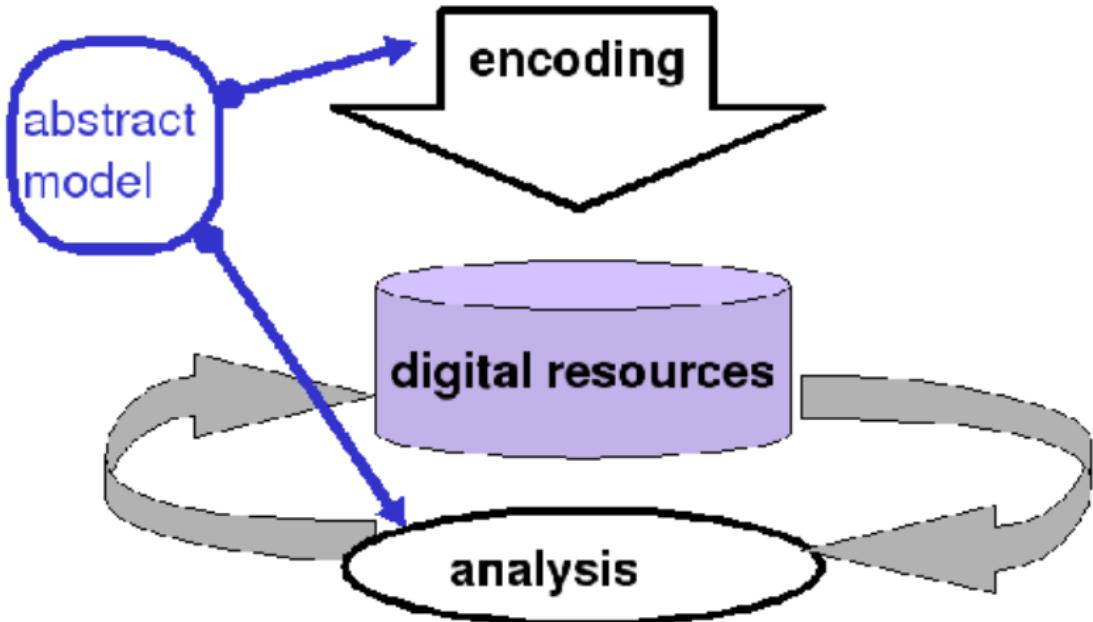
- paleographic level : identifying the characters and other graphemic components
- documentary or diplomatic level : determining what was originally written
- editorial or semantic level : determining how it ought to be read

Digitization provides an opportunity to make each step explicit, complex, and reversible



# The hermeneutic circle of digital enrichment

## Resources



## Enrichment

Adding markup to a document determines how it can be processed. It can concern many different aspects :

- the presentation of the document – its use of writing styles or typefaces, its rendering and layout
- the rhetorical organization of a document – its sections and subsections, its paragraphs and lists and headings and footnotes
- metatextual aspects of the document – its corrections and additions and deletions and errors and lacunae
- linguistic properties of a document – its syntax and morphology and semantics
- the document as an object – information about its origins and custodial history, its transmission and reception, its social function and category...

and many others.



# Un texte numérique peut être simplement ...

un 'substitut' (surrogate) représentant l'apparence d'un document existant

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.  
~~Worked~~ Worked at inserts of Richards' first chapter. Laura had a talk with Carl about deportment.

She slept until 5 (I working on Richards). I went to ~~the~~ Margarita to order my grey americans, & to Posada to open windows (shutting shutters) Turn out lights, take note away pencils.

Then worked at Gordon's life, after L. went over it. Carl brought melon, & we had coffee ice. Laura's stomach bad. I went to Fabrics & re-wound her

# ... ou peut aller plus loin

une représentation du contenu linguistique, de sa structure, avec des annotations sur sa portée, son contexte..

**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**

<a href="#">« June</a>	<a href="#">Abstract</a>	<a href="#">August »</a>				
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			

**July 27 Saturday**

Got up at 11.30. **Rosa** came.\*\*\*\*\*[crossed out]  
\*\*\*\*\*[crossed out] Worked at inserts of **Richards'** first chapter. **Laura** had a talk with **Carl** about deportment.

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

**DISPLAYED DIARY SCAN(S)**

  
[July 27 Saturday](#)  
[» Annotated markup](#)  
[» Full-sized Image](#)  
[» Gallery Scan](#)

**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.

© 2003 HCMC · University of Victoria · XML Markup · About this Publication

## .. et en dessous

....

-<div type="diaryentry" n="1935-07-27">

    <head> July 27 Saturday </head>

    -<p>

        Got up at 11.30.

        <rs type="person" key="Ro1">Rosa</rs>

        came.

        <unclear reason="crossed out"/>

    </p>

    -<p>

        <unclear reason="crossed out"/>

        Worked at inserts of

        <rs type="person" key="FR2">Richards</rs>

        ' first chapter.

        <rs type="person" key="LR1">Laura</rs>

        had a talk with

        <rs type="person" key="KG1">Carl</rs>

        about deportment.

    </p>

    -<p>

        She slept until 5 (I working on

        <rs type="person" key="FR2">Richards</rs>

    -<note>

        -<bibl>

            -<rs type="cita" key="OSS">

## La TEI nous propose un modèle conceptuel

- bien établi (depuis plus que 30 ans)
- adequat aux besoins pratiques
  - la conversion des données existantes
  - la création des données nouvelles
  - l'intégration des données déjà existantes mais répandues dans plusieurs sources
- adequat aux besoins scientifiques
  - derive d'un consensus et des pratiques consensuelles
  - mais aussi extensible
- exprimé en utilisant des formats ouverts et des technologies ouvertes

Est-ce que ceci représente la même chose ?

# A MONSEI.

GENEVRE REVÉ-

rendissime Cardinal  
du Bellay.

S.



E V le Personnage,  
que tu ioues au Specta-  
cle de toute l'Europe,  
uoyre de tout le Mon-  
de en ce grand Thea-  
tre 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 i'empeschoy' le tens, que tu donnes au  
scruiice de ton Prince, au profit de la Patrie, &  
& l'accroissement de ton immortelle renommée!  
Epant doncques quelque heure de ce peu de re-  
laiz, 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 celles du grand Hercule) ma Muse a pris  
la hardiesse d'êter au sacré Cabinet de tes sain-  
tes, & studicuses occupations: & la entre tant

4 ij de

A MONSEIGNEUR

*Le Reverendissime Cardinal du Bellay, S.*

Veu le personnage 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? Epant doncques 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-

Et ceci ?

# A MONSEI.

GN E V R L E R E V E -  
rendissime Cardinal  
du Bellay.

S.



E V le Personnage,  
que tu ioues au Specta-  
cle de toute l'Europe,  
uoire de tout le Mon-  
de en ce grand Thea-  
tre Romain, ueu tant  
d'affaires, & telz, que  
seul quasi tu soutiens: ô  
l'Honneur du Sacré Cola-  
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  
scrutice de ton Prince, au profit de la Patrie, &  
& l'accroissement de ton immortelle renommée?  
Epant doncques 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 épaules, 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, & studicuses occupations: & la entre tant

4 ij de

A MONSEIGNEUR

*Le Reverendissime Cardinal du Bellay, S.*

Veu le personnage 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



Joachim du Bellay

Défense et illustration de la  
langue françoysie (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 condamner sans avoir premièrement bien vu, et examiné ses raisons.

Épître à Monseigneur le réverendissime 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 honneur du sacré Collège, pecheroy-ie 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? Epant donc quelques heures de ce peu de relais que tu prends pour respirer sous le pesant faiz des affaires françoyses (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, entr de riches et excellents vœux de jour en jour dédiés à l'image de ta grandeur, prendre le sien humble et p mais toutefois bien heureux s'il rencontre quelque faveur devant les yeux de ta bonté, semblable à celle Dieux immortels, qui n'ont moins agréables les pauvres présents d'un bien riche vouloir que les superbe ambitieuses offrandes.

## Un texte n'est pas un document...

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

- l'apparence des lettres et leur mise-en-page
- la version originelle (supposée) de cette copie

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

- les interprétations/lectures apportées ou trouvées
- les intentions (supposées) de son auteur

# L'encodage

- Un texte est plus qu'une séquence de caractères encodés !
- Un texte est plus qu'une séquence de formes lexicales !
  - 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
```

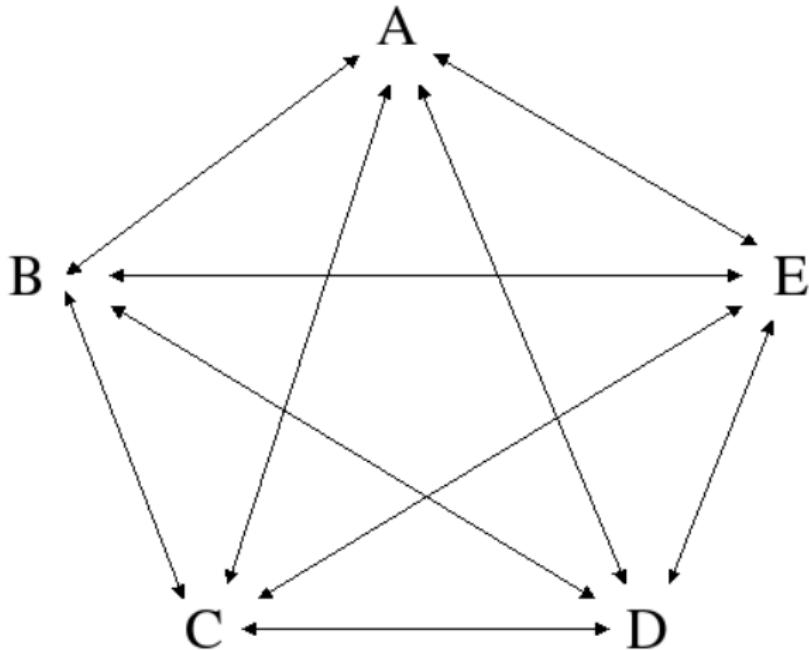
- Bonne nouvelle : il existe des logiciels capables de traduire entre 500 formats divers
- Mauvaise nouvelle : on en a besoin !

## 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
```

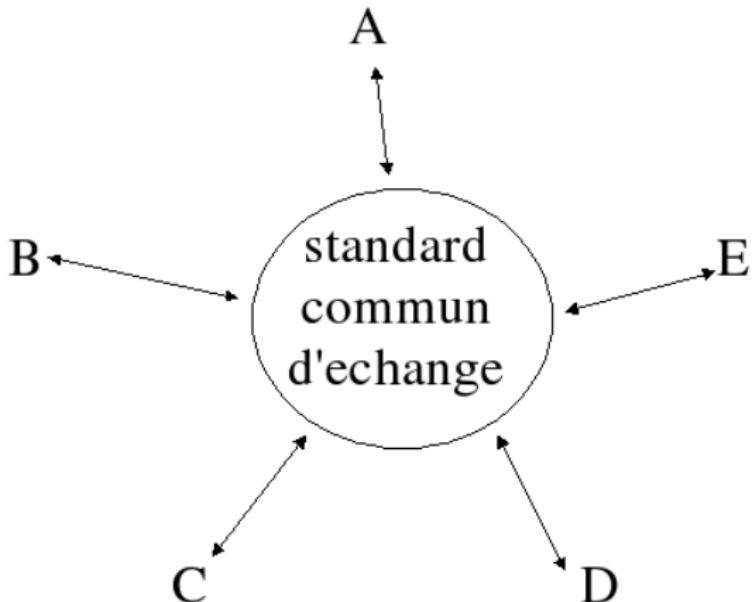
- Bonne nouvelle : il existe des logiciels capables de traduire entre 500 formats divers
- Mauvaise nouvelle : on en a besoin !

## Échange d'informations (1)



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

## Échange 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 réutilisation
  - et augmente la flexibilité
- Un balisage présentationnel, par contre, s'intéresse plus à l'affichage qu'à sa sémantique

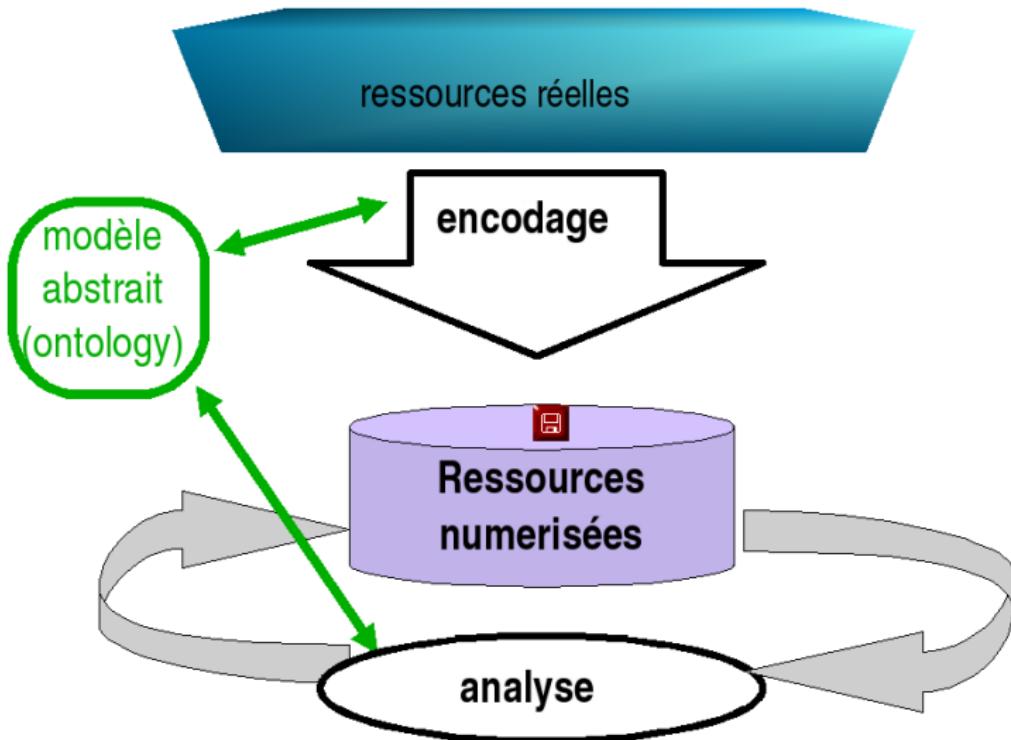
## Séparation de forme et contenu

- Un balisage descriptif s'intéresse plus au contenu qu'à sa mise en forme
  - cette séparation facilite la réutilisation
  - et augmente la flexibilité
- Un balisage présentationnel, 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 réutilisation
  - et augmente la flexibilité
- Un balisage présentationnel, par contre, s'intéresse plus à l'affichage qu'à sa sémantique

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



Ceci n'est pas un arbre



*Un modèle textuel réduit les complexités des textes, en utilisant un syntaxe simple pour les exprimer*

## XML, par exemple

```
texte texte <tag attribut="valeur">element</tag> texte
```

On introduit des balises dans le flux de texte

- pour identifier un empan de ce flux
- pour lui associer un nom ou type ...
- ... et peut être des attributs

Par exemple

# A MONSEI.

GNEVR LE R E V E .

rendissime Cardinal

du Bellay.

S.



E V le Personnaige,  
que tu ioues au Specta-  
cle de toute l'Europe,  
uoyre de tout le Mon-  
de en ce grand Thea-  
tre Romain, ueu tant  
d'affaires, & telz, que  
seul quasi tu soutiens à  
l'Honneur du sacré Cola-  
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

Qu'est ce qu'on balisera?

# Balisage du mise en evidence

A MONSEI-  
GN E V R LE R E V E -  
rendissime Cardinal  
du Bellay.  
S.



EV le Personnage,  
que tu ioues au Specta-  
cle de toute l'Europe,  
uovre de tout le Mon-

```
<pb n="4"/>A MONSEI-  
<lb/>GN E V R LE R E V E -  
<lb/>rendissime Cardinal  
<lb/>du Bellay.  
<lb/>S  
  
<lb/>  
<c rend="lettrine">V</c>EV le  
Personnage,  
<lb/>que tu ioues au Specta-  
<lb/>cle de toute l'Europe...
```

# Balisage de structure

A MONSEI.  
GN E V R L E R E V E .  
rendissime Cardinal  
du Bellay.  
S.



EV le Personnaige,  
que tu joues au Specta-  
cle de toute l'Europe,  
nouvre de tout le Mon-

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

## Mais on peut aller plus loin...

```
<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>
```

## à ne rien dire de ...

```
<head> A MONSEIGNEUR LE REVERENDISSIME  
<persName ref="#dubellay03">CARDINAL DU  
BELLAY</persName>  
</head>  
<!-- .... -->  
<person xml:id="dubellay03">  
  <persName>Jean du Bellay</persName>  
  <birth>  
    <date>1492</date>  
    <placeName>Souday</placeName>  
  </birth>  
  <death>  
    <date when="15600216">16 February 1560</date>  
    <placeName>Roma</placeName>  
  </death>  
<!-- .... -->  
</person>
```

## Conclusions préliminaires

- Avant de commencer un exercice de balisage, il faut bien préciser son choix des balises
- Ce choix sera déterminé par les distinctions et métainformations qu'on considère d'importance
- L'XML nous aide en définissant un syntaxe formel pour notre balisage
- La TEI nous aide en fournissant un lexique très complet des balises disponibles

Revenons d'abord sur le syntaxe XML

## Notre système d'encodage devrait être capable de...

- spécifier les caractères d'un texte
- expliciter la/les structures aperçue/s dans un texte
- linéariser le texte
- spécifier les méta-information, renseignements contextuels etc.
- prendre en compte la sémantique de la texte

Jusqu'à présent, XML semble une bonne solution...