

Principes et enjeux du balisage

Lou Burnard

Texte et texte numérique

Un texte peut être considéré selon trois axes :

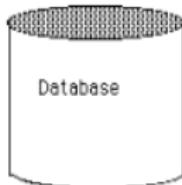
- Un texte a une existence physique, ayant des **traits visuels** qu'on peut (plus ou moins) transférer automatiquement d'une instance à une autre
- Un texte possède des **propriétés linguistiques et structurelles**, qu'on ne peut transcrire, traduire, ou transmettre qu'avec une compréhension humaine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

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

Traitements numériques du texte



- * word processing
- * indexing
- * database



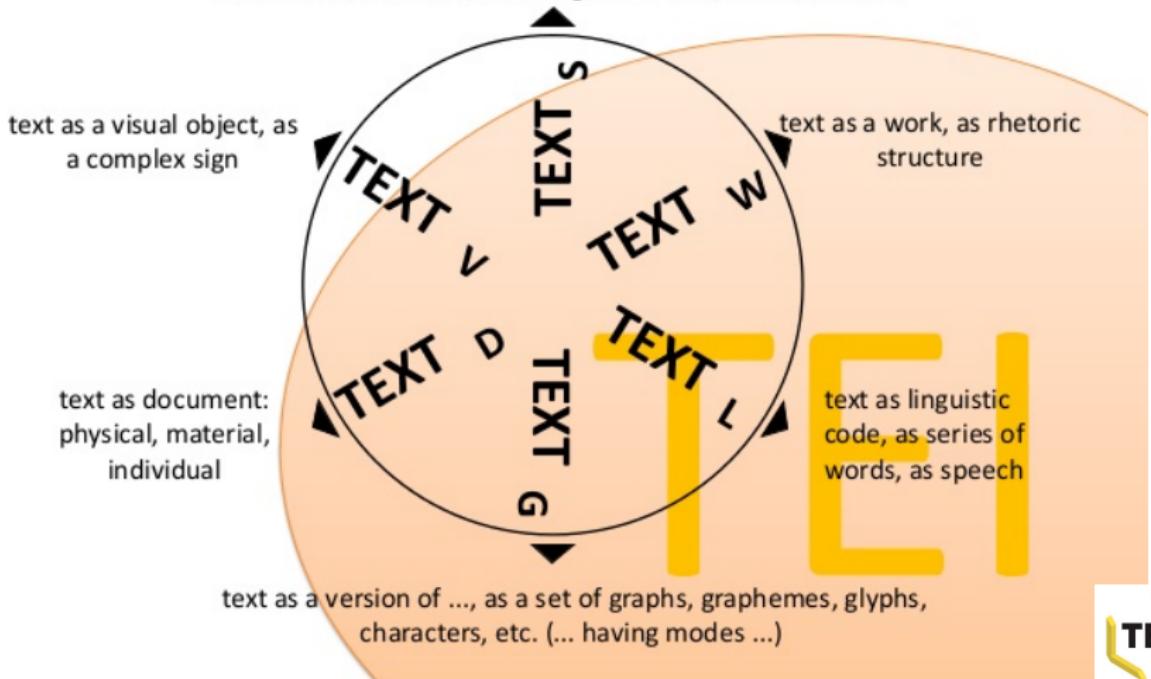
aardvark
abacus
abracadabra
acidulation
adumbrate
aetherial
affection
agony
aha
ai
ajacanthus

(C'est peut-être plus qu'un triptyque)

The Textwheel

(Patrick Sahle: Digitale Editionsformen, 2013)

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



Les familles de logiciels

Les logiciels existants ont tendance à se spécialiser...

- outils de gestion et de production de documents
- de gestion et de production d'images
- outils d'analyse et de gestion linguistique
- systèmes de gestion de bases de données

La convergence

Cependant la convergence fait désormais partie du quotidien numérique de chacun d'entre nous. Lorsqu'on fait un "mashup" qui articule

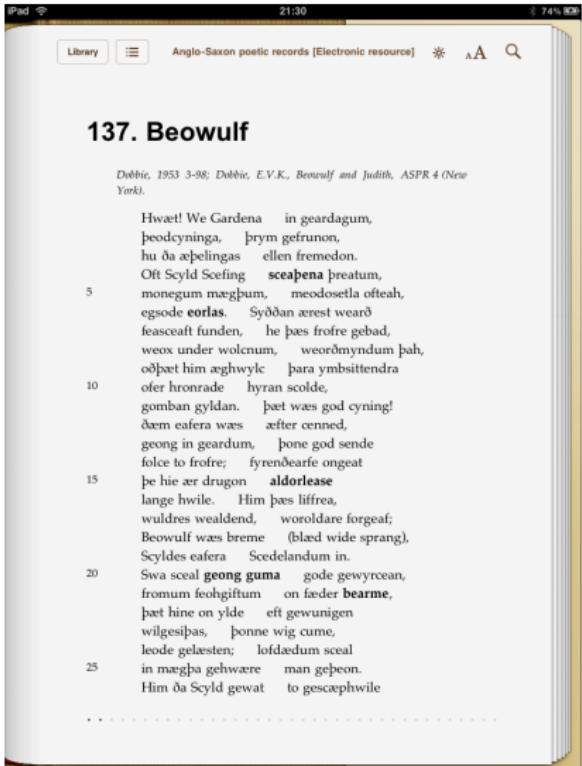
- une base de données de type système d'information géographique sur les lieux de la mer Egée
- un index des noms de lieux situés dans la même aire géographique
- un corpus de textes qui mentionnent ces lieux

on a besoin de combiner la puissance d'une base de données avec des outils d'analyse linguistique et de restitution spatiale de l'information.

Quelques exemples :

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

Le problème



137. Beowulf

Dobře, 1953 3-98; Dobře, E.V.K., *Beowulf and Judith*, ASPR 4 (New York).

Hwæt! We Gardena in geardagum,
þeodcyniga, þrym gefrunon,
hu ða æfelingas ellen fremedon.
5 Oft Scyld Sefering **sceapena** þreatum,
monegum mægþum, meodosetla ofteah,
egode **eorlas**. Syððan ærest weard
feasceaft funden, he þas frofre gebad,
weox under wolcnum, weordmyndum þah,
oðþæt him aghwylc þara ymsittendra
10 ofer hronrade hyran scoldé,
gomban gyldan. þæt wæs god cyning!
ðær esfara wæs æfter cenned,
geong in geardum, þone god sende
folce to frofre; fyrendearle ongeat
15 þe hie ær drugen **aldorlease**
lange hwile. Him þas liffrea,
wuldfres wealend, woroldare forgefæ;
Beowulf was breme (blad wide sprang),
Scyldes esfara Scedelandum in.
20 Swa sceal **geong gumna** gode gewyrcean,
fromum feohgiftum on feder **bearme**,
þæt him on ylde eft gewunigen
wilgesiðas, þonne wig cumē,
leode gelesten; lorfædum sceal
25 in mægþa gehwære man geþeon.
Him ða Scyld gewat to gescæphwile.

- Les applications actuelles des bibliothèques numériques se concentrent encore sur la fourniture de pages virtuelles pour le lecteur : la métaphore du livre est tellement omniprésente que nous pouvons tout juste la voir
- Il est évident que la numérisation permet d'offrir des substituts moins chers et plus accessibles des pages imprimées ou manuscrites.
- Mais ce n'est pas assez... textes numériques devront aller 'au-delà de la page'

A quoi sert un texte numérique ?

Les applications numériques permettent de faire plus avec un texte, en particulier avec une collection de textes !

- plus que simplement lire le texte du début à la fin
- plus que lui associer des annotations pour que les autres les lisent
- plus qu'exécuter une fouille de texte brute

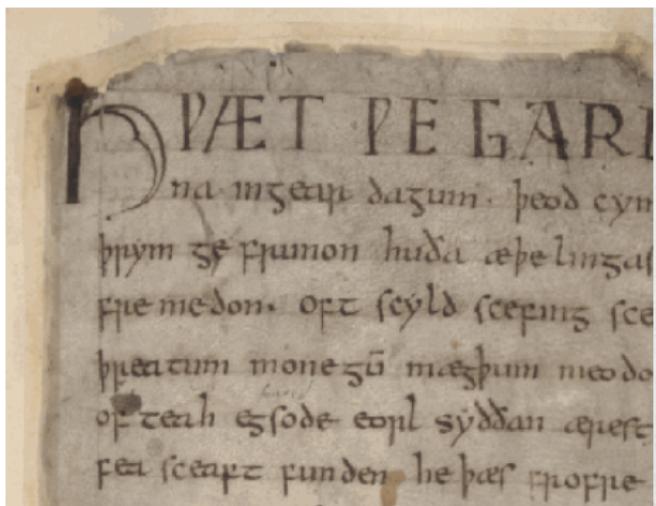
Le contenu de la bibliothèque numérique doit donc être enrichi, même si cela requiert l'utilisation de techniques qui ne sont pas automatisables aujourd'hui.

Que se passe-t-il dans la bibliothèque numérique ?



- Une édition numérique devrait capturer les intentions et le sens d'un texte, non seulement son apparence
- Sans quoi, il ne peut y avoir d'analyse au-delà du niveau documentaire, pas de 'conversation entre les livres'

Enrichissement ou représentation ?



Quand nous allons de ceci...

... à cela, que se passe-t-il ?

Hwæt wē Gār-Dena
þēod-cyninga brym ^{glory}
hū ðā æfelinges ellen
Oft Scyld Scēfing
5 monegum mægþum
egsode Eorl[e], - syðða
fēasceaft funden; hē
wēox under wolcnum,
oðþæt him ēghwylc
10 ofer hron-rāde hȳran

L'édition

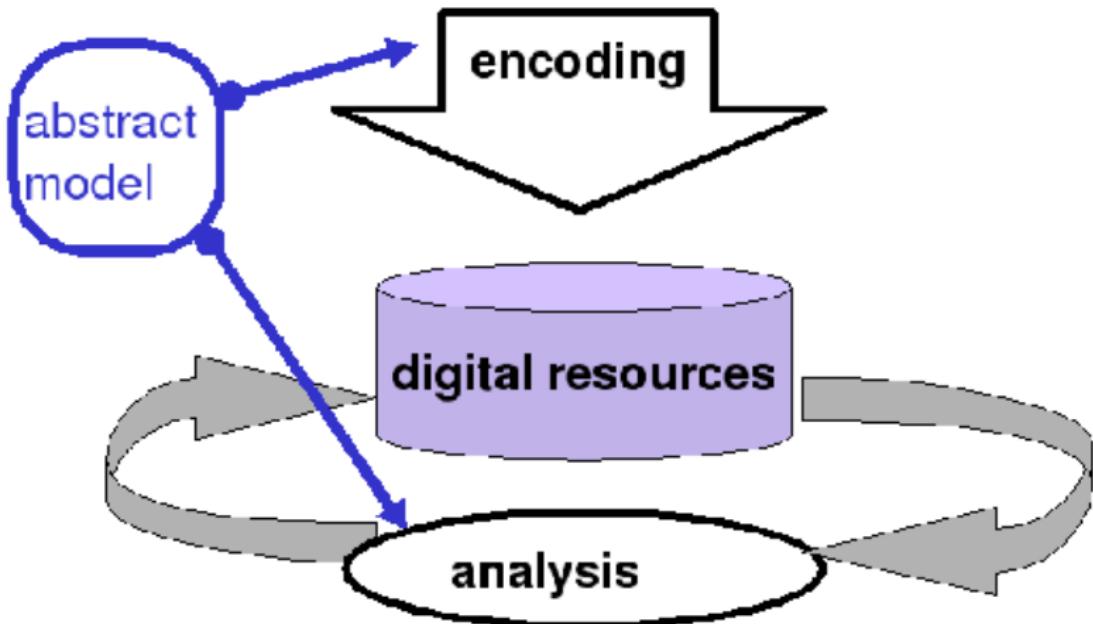
Il est courant de distinguer au moins les catégories ou niveaux suivants d'interprétation d'un texte :

- le niveau paléographique : identification des caractères et autres graphèmes
- le niveau documentaire ou diplomatique : détermination de ce qui a été écrit à l'origine
- le niveau éditorial ou sémantique : détermination des modalités de lecture

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

Le cercle herméneutique de l'enrichissement numérique

Resources



Enrichissement

Ajouter du balisage à un document détermine comment il peut être traité automatiquement. Le balisage peut concerner de nombreux aspects différents :

- la présentation du document – son usage des styles d'écriture ou des polices, son apparence et sa mise en page
- l'organisation rhétorique du document – ses sections et sous-sections, ses paragraphes et listes, ses intertitres et notes de bas de page ;
- les aspects métatextuels du document – corrections, additions, suppressions, erreurs, lacunes
- les propriétés linguistiques du document – sa syntaxe, sa morphologie et sa sémantique
- le document comme un objet – les informations sur sa création, l'histoire de sa conservation, sa transmission, sa réception, sa fonction sociale, sa catégorie...
et beaucoup d'autres aspects.

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 spectacles.

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

« 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			

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**¹). I went to **Fábrica**[RG] **Margarita** to order my grey **Americano** ², & to **Posada** to open windows (shutting shutters) turn out lights, take **into**[RG] away perishables.

Then worked at **Gordon's** Life³, 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** ⁴. More work on **Gordon**⁵. 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

¹ [Old Soldier Sahib](#), eds.
² Spanish [slang?] for "jacket" KG; KG also replaces the final "o" with "a". eds.
³ See [A Mistake Somewhere](#), eds
⁴ bus. KG
⁵ 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 de 30 ans)
- adapté aux besoins pratiques
 - la conversion de données existantes
 - la création de données nouvelles
 - l'intégration des données déjà existantes mais répandues dans plusieurs sources
- adapté aux besoins scientifiques
 - dérive d'un consensus et de 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 françoys (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'-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 ? 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
```

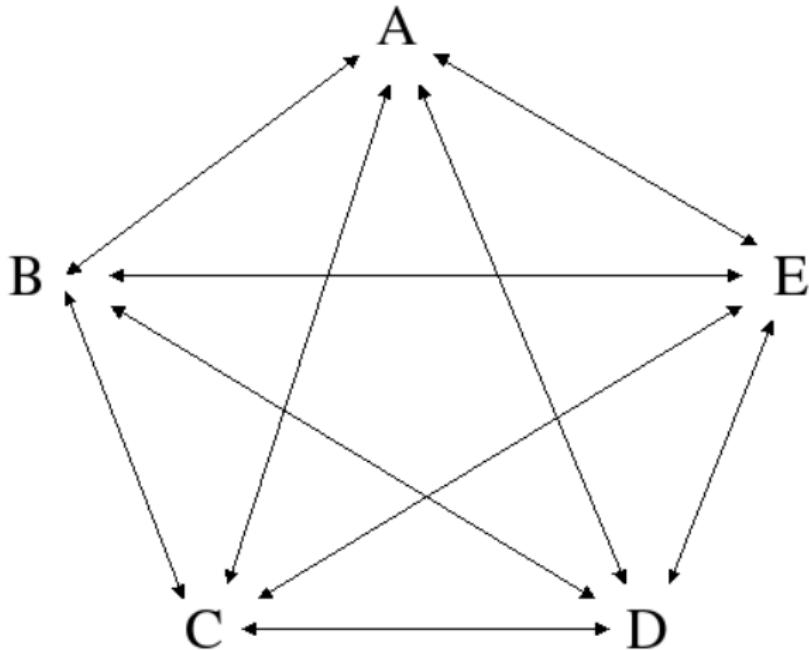
- Bonne nouvelle : il existe des logiciels capables de traduire entre 500 formats divers
- Mauvaise nouvelle : on en a besoin !

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
```

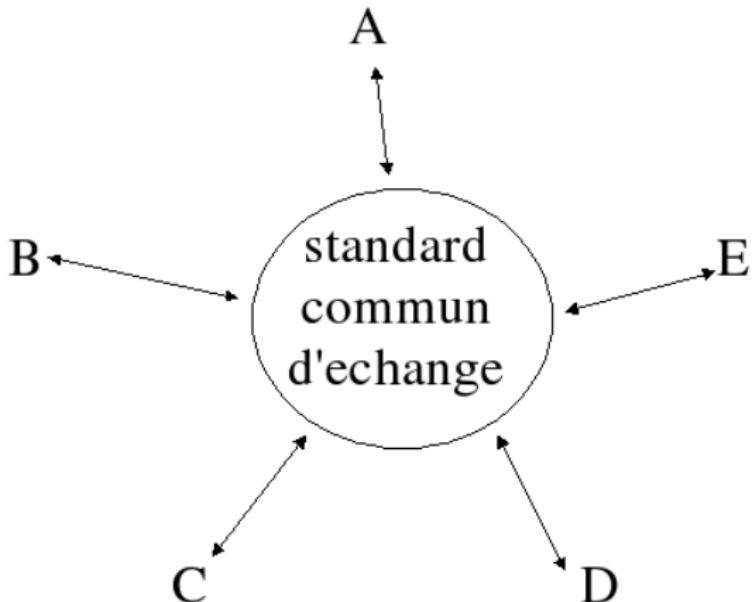
- Bonne nouvelle : il existe des logiciels capables de traduire entre 500 formats divers
- Mauvaise nouvelle : on en a besoin !

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 la forme et du 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

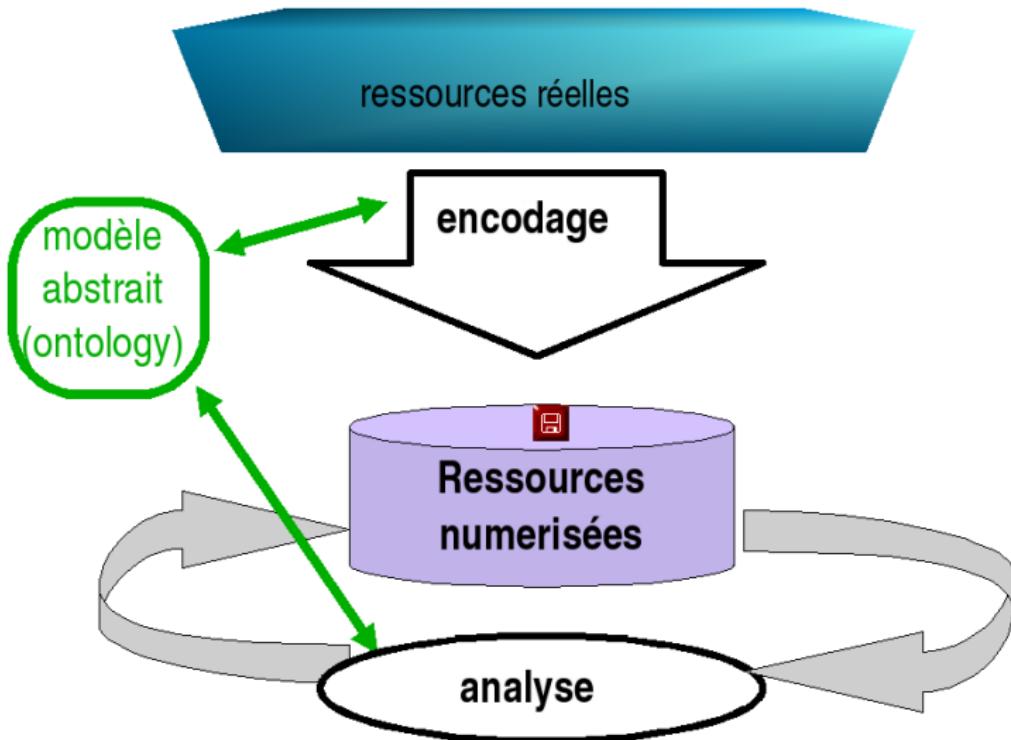
Séparation de la forme et du contenu

- 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 la forme et du 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 une 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 segment de ce flux
- pour lui associer un nom ou type ...
- ... et peut-être des attributs

Par exemple

A MONSEI.

GNEVR LE REVE-

rendissime Cardinal

du Bellay.

S.



EV 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 de la mise en page

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.
A N 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>
```

sans parler 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
- XML nous aide en définissant une syntaxe formelle pour notre balisage
- La TEI nous aide en fournissant un lexique très complet des balises disponibles

Revenons d'abord sur la 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 du texte.

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