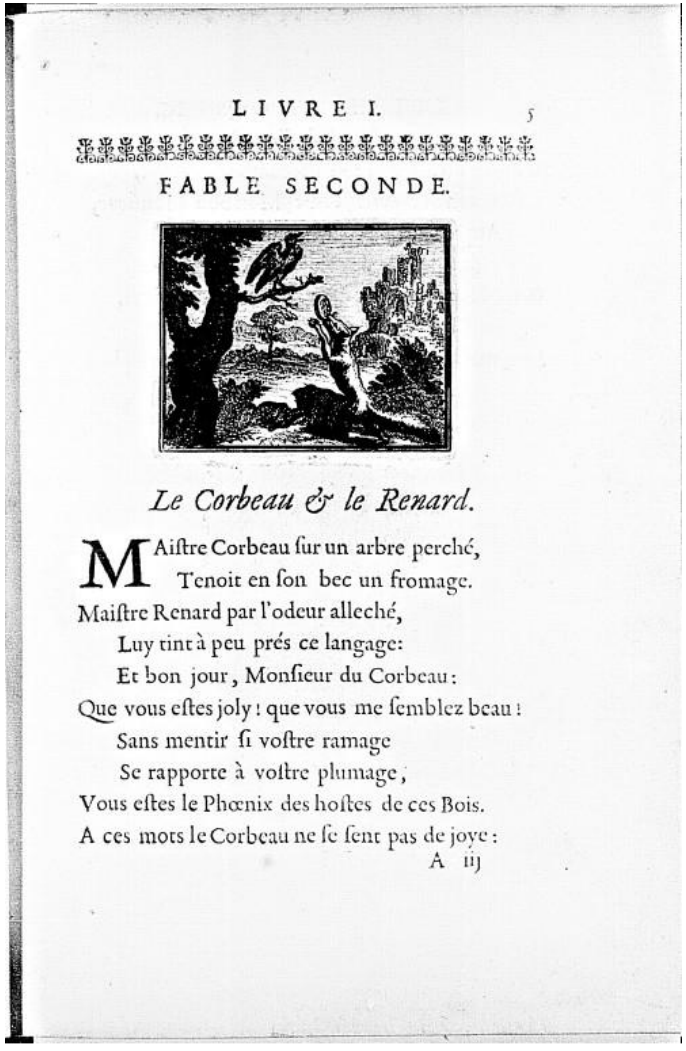
Construction des corpus pour l'analyse lexicale

Lou Burnard Consulting

Objectifs de cette presentation

1. Préciser ce que c'est que l'encodage textuel
2. Un cas d'etude: le *Petit Comtois*
3. Présenter quelques concepts fondamentaux de TEI et d'XML

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



LIVRE I. 5

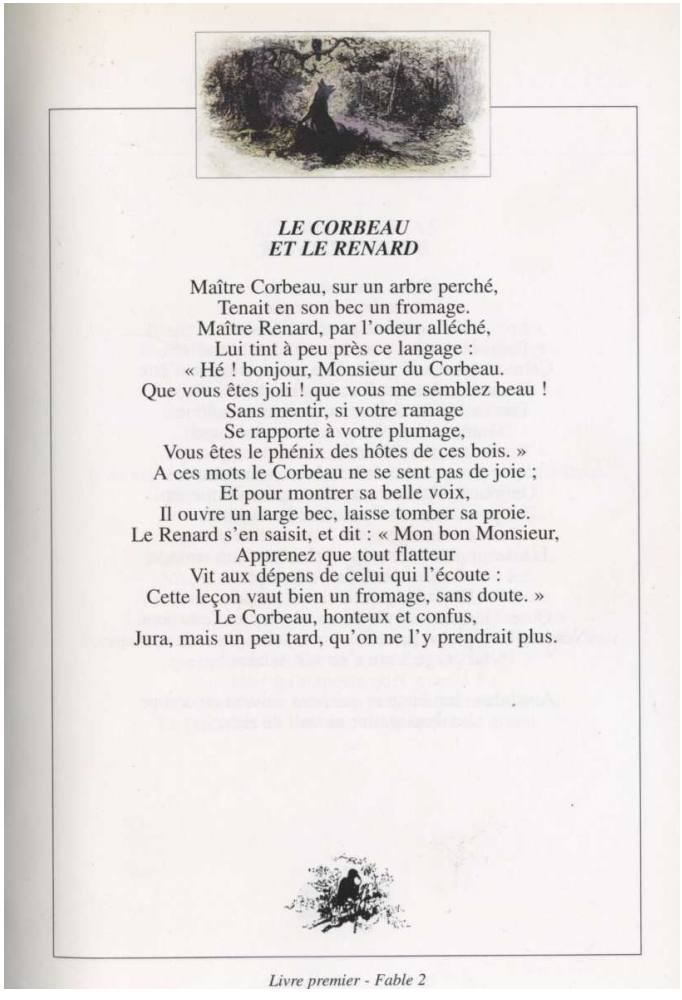
FABLE SECONDE.

Le Corbeau & le Renard.

MAiſtre Corbeau ſur un arbre perché,
Tenoit en ſon bec un fromage.
Maiſtre Renard par l'odeur alleché,
Luy tint à peu prés ce langage:
Et bon jour, Monſieur du Corbeau:
Que vous eſtes joly! que vous me ſemblez beau!
Sans mentir ſi voſtre ramage
Se rapporte à voſtre plumage,
Vous eſtes le Phœnix des hoſtes de ces Bois.
A ces mots le Corbeau ne ſe ſent pas de joye:

A iij

Source gallica.bnf.fr / Bibliothèque nationale de France



LE CORBEAU
ET LE RENARD

Maître Corbeau, sur un arbre perché,
Tenait en son bec un fromage.
Maître Renard, par l'odeur alléché,
Lui tint à peu près ce langage :
« Hé ! bonjour, Monsieur du Corbeau.
Que vous êtes joli ! que vous me semblez beau !
Sans mentir, si votre ramage
Se rapporte à votre plumage,
Vous êtes le phénix des hôtes de ces bois. »
A ces mots le Corbeau ne se sent pas de joie ;
Et pour montrer sa belle voix,
Il ouvre un large bec, laisse tomber sa proie.
Le Renard s'en saisit, et dit : « Mon bon Monsieur,
Apprenez que tout flatteur
Vit aux dépens de celui qui l'écoute :
Cette leçon vaut bien un fromage, sans doute. »
Le Corbeau, honteux et confus,
Jura, mais un peu tard, qu'on ne l'y prendrait plus.

Livre premier - Fable 2

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...

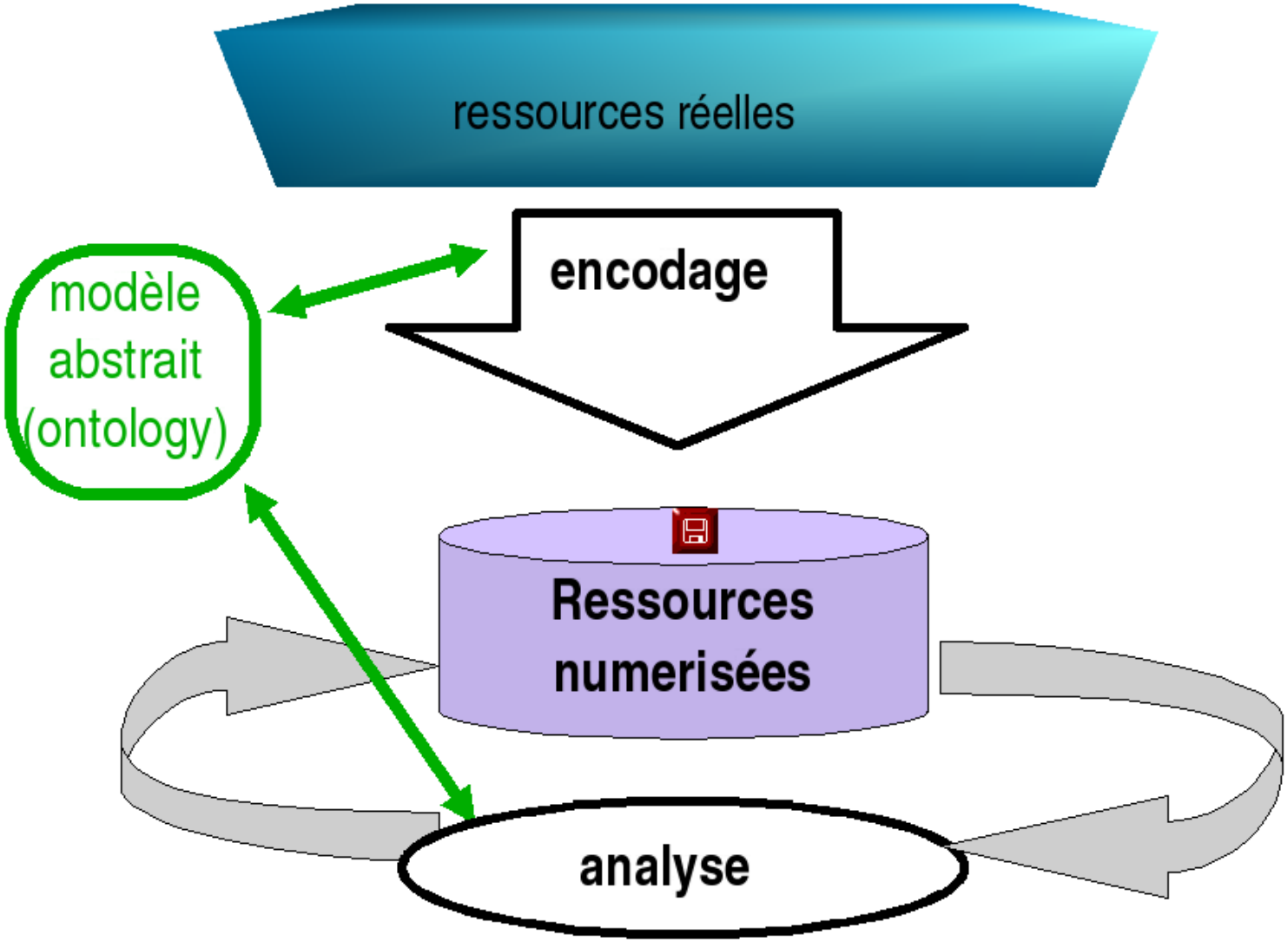
En quoi consiste l'essentiel d'un texte ?

- 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?

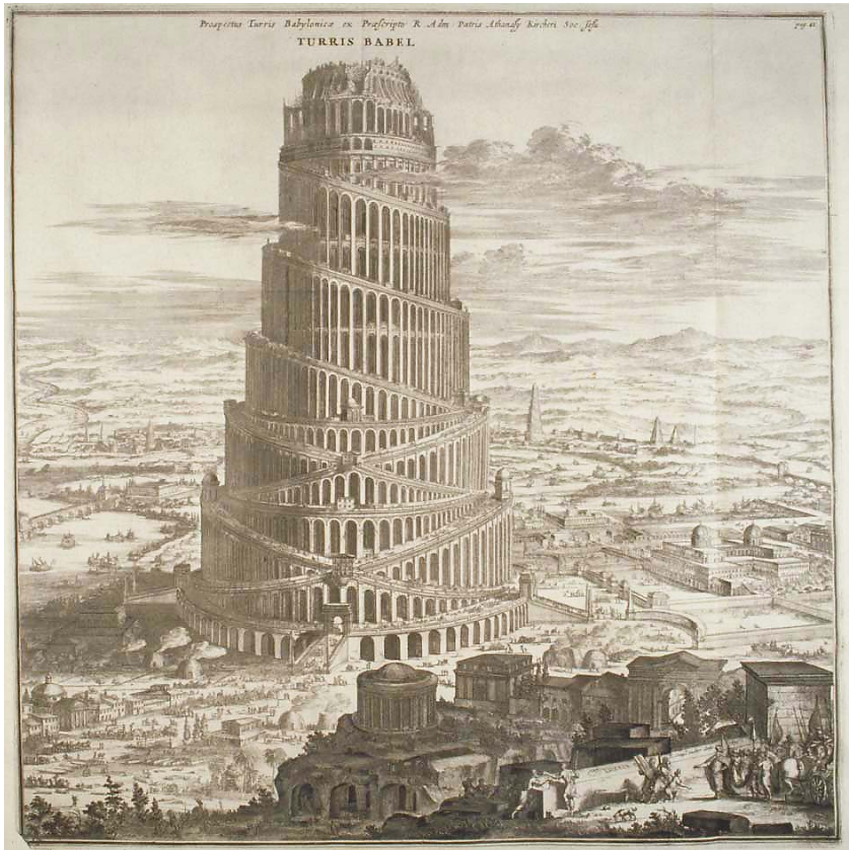


L'encodage

- 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...



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

Par exemple...

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

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
```

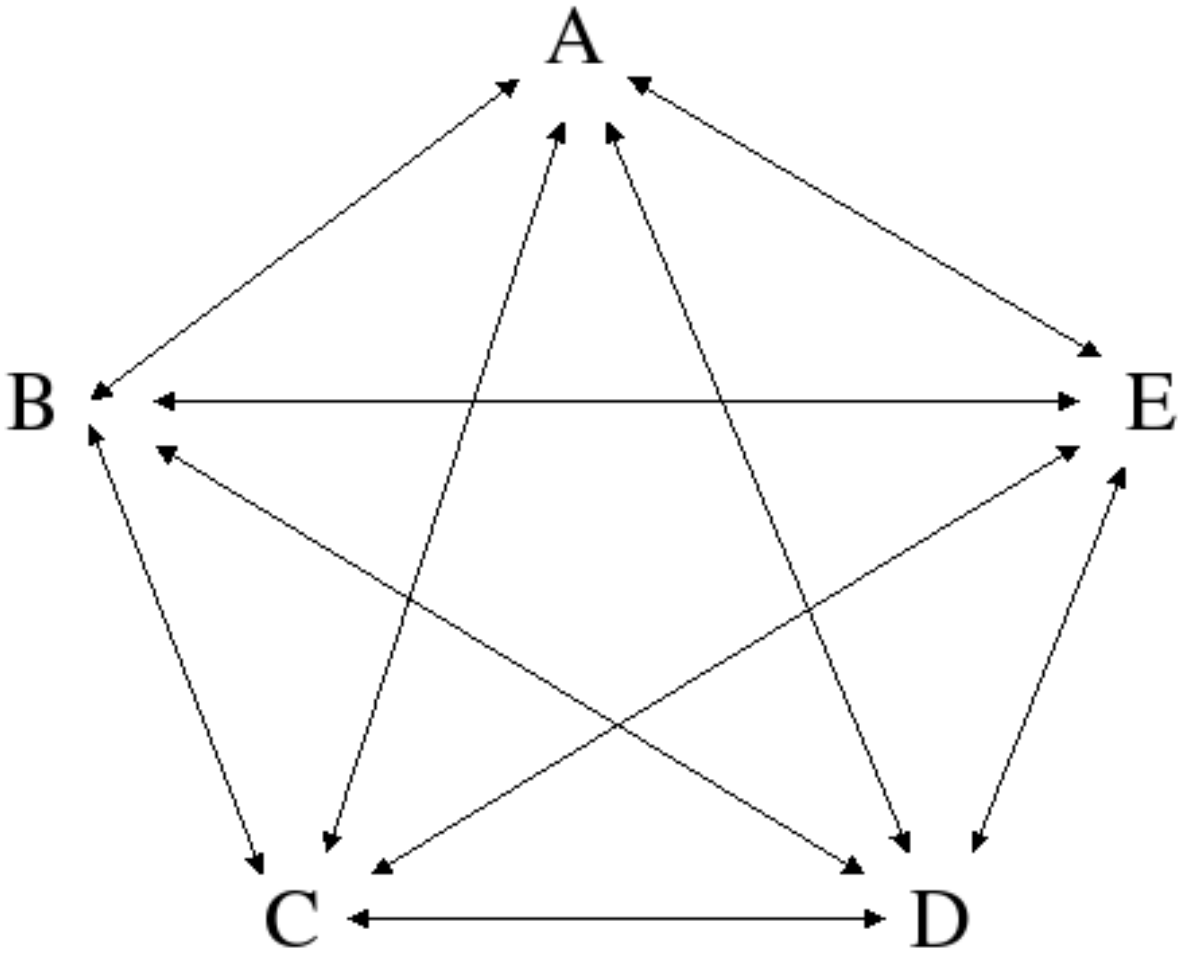
- Bonne nouvelle: il existe des logiciels capables de traduire entre 500 formats divers
- Mauvaise nouvelle: on en a besoin

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
```

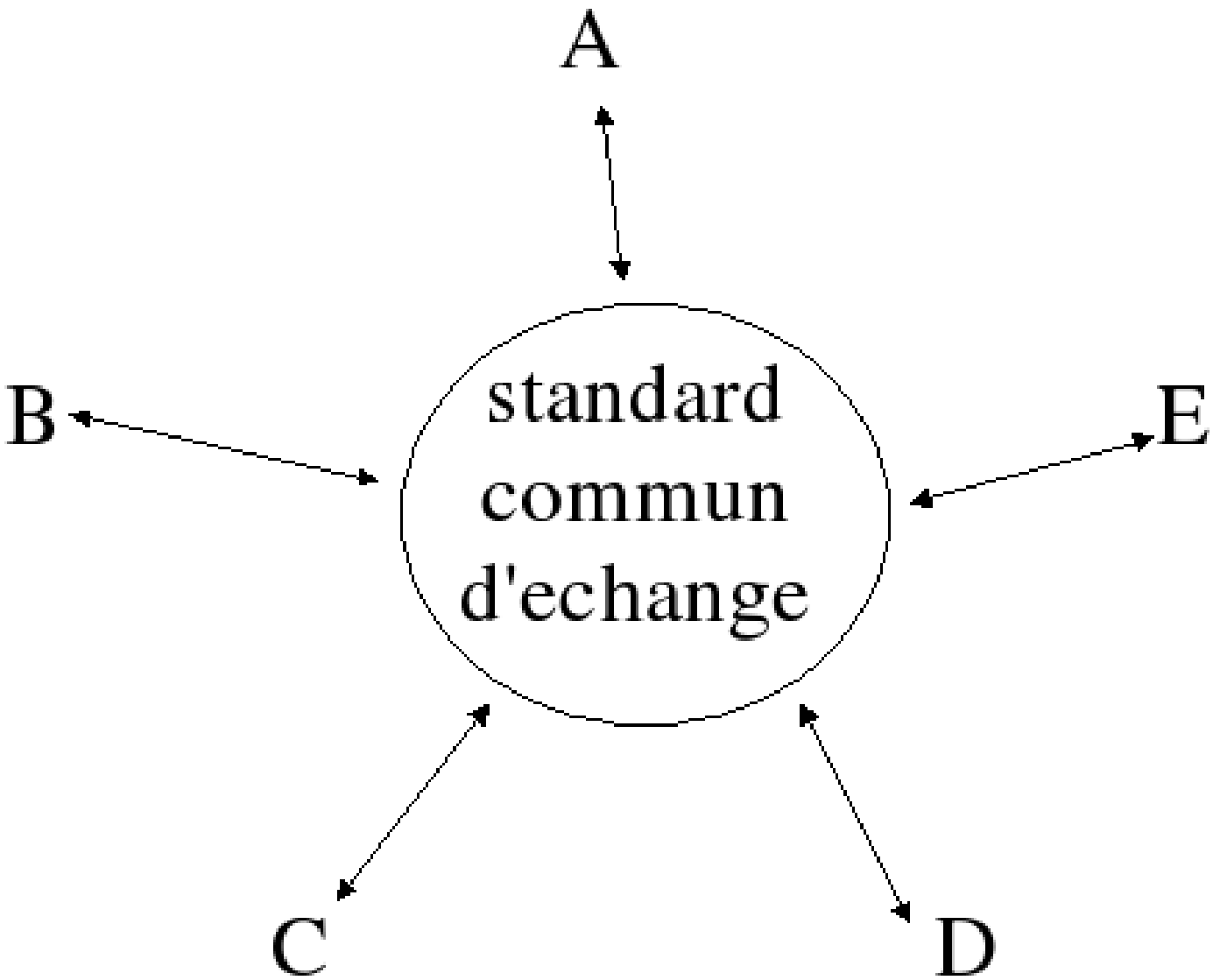
- Bonne nouvelle: il existe des logiciels capables de traduire entre 500 formats divers
- Mauvaise nouvelle: on en a besoin

Echange d'informations (1)



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

Echange d'informations (2)



avec format pivot: 10 passerelles requises (2n)

Objectif d'un format pivot

- permettre la representation de toute information requise
- d'une maniere independente
  - de logiciel specifique
  - de plateforme specifique
  - d'usage specifique
- sans hausse improbable d'effort

Un objectif assez classique, qui s'exprime avec un balisage du texte

Objectif du balisage

Le balisage constitue un langage d'encodage, qui devrait...

- 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.

Le balisage: un technique bien-compris

- 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 va-t-on baliser?

Qu'est ce qu'on balisera?

Comparer:

```
<lb/>LIVRE I.
<lb/>FABLE SECONDE.
<lb/>Le Corbeau & le Renard
<lb/>MAistre Corbeau sur un arbre perché,
<lb/>Tenoit en son bec un fromage.
<lb/>Maistre Renard par l'odeur alleché,
<lb/>Luy tint à peu prés ce langage: ...
<lb/>A ces mots, le Corbeau ne se sent pas de joye:
<lb/>A iii
```

avec

```
<div type="fable" n="I2">
  <head>Le Corbeau & le Renard</head>
  <l>
    <hi rend="lettrine">M</hi>aistre Corbeau sur un arbre perché,</l>
  <l>
    <reg>Tenait</reg> en son bec un fromage.</l>
  <l>Maistre Renard par l'odeur alleché,</l>
  <l>
    <reg>Lui</reg> tint à peu prés ce langage:</l>
<!-- ... -->
  <l>A ces mots, le Corbeau ne se sent pas de joye:</l>
  <fw type="sig">A iii</fw> ...
</div>
```

Ou bien ....

```
<lg>
  <milestone unit="voix" n="narr"/>
  <l>Maître corbeau, sur un arbre perché</l>
  <l>Tenait en son bec un fromage.</l>
  <l>Maître renard par l'odeur alléché</l>
  <l>Lui tint à peu près ce langage:</l>
  <l>
    <milestone unit="voix" n="renard"/>Hé! bonjour Monsieur du Corbeau</l>
  <l>Que vous êtes joli! que vous me semblez beau!</l> ...
</lg>
```

ou bien

```
<lg>
  <sp who="#narr">
    <ab>
      <lb/>Maître corbeau, sur un arbre perché
    <lb/>Tenait en son bec un fromage.
    <lb/>Maître renard par l'odeur alléché
    <lb/>Lui tint à peu près ce langage:</ab>
  </sp>
  <sp who="#renard">
    <ab>
      <lb/>Hé! bonjour Monsieur du Corbeau
    <lb/>Que vous êtes joli! que vous me semblez beau! ... </ab>
  </sp>
</lg>
```

Un autre exemple...

N° 6 — Première année — CINQ CENTIMES — Lundi 6 août 1883

LE PETIT COMTOIS

JOURNAL RÉPUBLICAIN DÉMOCRATIQUE QUOTIDIEN

RÉDACTEUR EN CHEF, JULES GROS

ABONNEMENTS — 3 mois / 6 mois / 1 an
France et colonies . . . 6 f. / 12 f. / 24 f.
Pays étrangers . . . 8 f. / 15 f. / 30 f.
Edition hebdomadaire . . . » / » / 5 f.

RÉDACTION ET ADMINISTRATION
BESANÇON. — 7, Square Saint-Amour, 7 — BESANÇON
Les abonnements partent des 1er et 16 de chaque mois

INSERTIONS
Annonces . . . 20 c. la ligne.
Réclames . . . 50 c.
Faits divers . . . 1 fr.
Les annonces sont reçues à l'avance au bureau du journal, et à Paris, chez nos agents.

CHEMIN DE FER. — Service d'été. — L'heure de la gare est en retard de 10 minutes sur celle de la ville.

Départ de Besançon : Pour Dijon : Matin, 5 h. 45, 8 h. 24 ; Soir, 2 h. 19, 5 h. 25, 8 h. 34. — Pour Belfort : Matin, 5 h. 32, 9 h. 05 ; Soir, 2 h. 05, 6 h. 26. — Pour Auxonne : Matin, 7 h. 45 ; Soir, 2 h. 46, 4 h. 40 pour Montbard et Paris, 8 h. 34, pour Lons-le-Saunier. — Pour Pesmes : Matin, 7 h. 25 ; 2 h. 40, 6 h. 45 ; Soir, 2 h. 34, 7 h. 30. — Pour Gray (par Pesmes) : Matin, 7 h. 25 ; Soir, 2 h. 40, 6 h. 45.

Arrivée à Besançon : De Dijon : Mat., 9 h. 45, 7 h. 21, 9 h. 30 ; Soir, 12 h. 38, 5 h. 58, — De Dole et Chalon (mixte), 8 h., 9 h. 50. — De Belfort : Matin, 7 h. 50, 11 h. 17 ; Soir, 1 h. 47, 4 h. 37, 8 h. 06. — De Auxonne : Matin, 9 h. 12, Soir, 4 h. 55, 9 h. 05, 10 h. 35. — De Pesmes : Matin, 7 h. 34, Soir, 4 h. 45, 8 h. 21, 7 h. 45 ; 10 h. 27. — De Gray : Matin, 10 h. 22 ; Soir, 1 h. 09, 7 h.

LA POLITIQUE COLONIALE

Lorsque, suivant le mot devenu historique, de son ministre Gortschakoff, « *la Russie se recueillait* » après ses désastres de Crimée, et que, renfermée en quelque sorte chez elle dans l'attente des jours meilleurs, elle gardait en Europe une attitude expectante, annonçait, au contraire, libre essor à son activité dans sa grande colonie asiatique et s'abandonnait tout entière à ce magnifique mouvement d'expansion coloniale qui fait d'elle aujourd'hui la principale puissance européenne de l'Asie, alors il ne manquait pas, chez elle d'hommes qui — comme aujourd'hui chez nous — éliminant cette politique, prêchaient l'abstention et dénonçaient comme acte d'imprévoyance tout agrandissement colonial.

Vaines terreurs ! L'avenir, au nom duquel ils prétendaient parler, s'est prononcé contre eux, et l'histoire récente de la Russie qui, après s'être retrempée dans l'œuvre immense de la conquête et de la colonisation de l'Asie centrale, s'est relevée depuis de ses désastres européens, nous démontrent suffisamment que ces hommes timorés, qui déconseillent à tout propos « *la politique d'aventures* », sont eux-mêmes les pires des hésitations et des défaillances du Parlement, qui ont failli compromettre irrémédiablement nos deux expéditions de Tunisie et du Tonkin, comme elles nous ont perdu l'Egypte. nous sommes entrés résolument dans une politique énergique d'expansion coloniale, et nous suivons cette voie, il faut bien le dire, à pas de géant.

Installés aujourd'hui en maîtres à Tunis, où nous avons devancé l'Italie, ne sommes-nous pas en train d'élaborer ce vaste empire colonial africain dont nous avons déjà ainsi jeté les bases par le développement de la colonisation algérienne, par l'occupation successive de la Tunisie, de deux bien-nous devenus immenses, le Niger et le Congo, et par la conquête de Madagascar, cette « *France équinoxiale* » d'où nous dominerons un jour, alors que nous y serons implantés comme race, tout l'océan Indien ?

N'est-ce pas un nouvel empire des Indes, dont nous avons tout récemment conçu la pensée et déjà presque réalisé la conquête sur le continent asiatique, où nous sommes à la veille de décupler notre influence et notre puissance coloniale par l'adjonction de deux vastes et riches contrées, l'Annam et le Tonkin, à nos belles provinces de Cochinchine et du Cambodge ?

Exclusivement guidés dans cette œuvre de vulgarisation par la préoccupation de faire connaître, avec l'impartialité dont notre indépendance est la meilleure garantie, ce que nous croyons être la vérité sur notre politique coloniale, nous estimons superflu de répondre aux arguments sans portée formés de conjectures et d'hypothèses et de chercher la solution de problèmes dont l'avenir seul dira le secret.

Qu'importe, en effet, de savoir si, poursuivant cette politique, nous encourrons la jalousie de l'Angleterre et les témoignages de sa réprobation, qu'elle ne ménage jamais à tout voisin qui s'enrichit.

Il s'agit de savoir si, revenant en arrière, abandonnant cette œuvre grandiose de la conquête du globe à des aventuriers anglo-saxons, nous rappellerons en France ces hommes de cœur et de courage : Brazza, Duveyrier, Pierre, Borgnis-Desbordes, tous ces hardis pionniers qui, portant avec eux notre langue et notre civilisation, sont aujourd'hui sur tous les points du globe l'honneur du pays.

Il s'agit de savoir si, arrivés au terme de notre développement et de notre expansion comme race, nous renoncerons [illegible] dès à présent toute ambition mondiale, et si, conscients de notre vitalité comme race, nous nous sentons encore au cœur l'ardeur de longs espoirs et de vastes pensées.

JEAN DES CHAMPS.

DÉPÊCHES DE NUIT

Service de notre correspondant spécial.

INFORMATIONS GÉNÉRALES.

Les magistrats des commissions mixtes.

La nouvelle loi sur la magistrature aura pour effet d'éliminer les magistrats ayant fait partie des commissions mixtes et étant encore actuellement en activité de service.

M. Camescasse, conseiller à la cour de cassation et père du préfet de police actuel, tombe par le fait sous l'application de la loi. M. Camescasse a demandé sa mise à la retraite et il l'a obtenue, mais la disposition qui fait droit à sa demande ne doit courir que l'hiver prochain.

Incident à Philippeville

Un sieur Hanotte, professeur au lycée de Philippeville, ayant été gravement insulté dans une feuille locale des articles injurieux contre les colons de l'Algérie, a été la cause d'une manifestation tumultueuse

Un petit exercice intellectuel...

Imaginez vous: on a des milliers des pages à baliser....

- Quels traits existent dans ces materiaux?
- Quels traits faut-il baliser? lequels seraient (in)utiles?
- Comment justifier l'étiquettage choisi?
- Comment garantir et la consistence et l'étendue du balisage?

Maintenant, on réduit le budget de 50%. Répétez l'exercice!

Par exemple...

On commence par noter la structuration:

- la page contient des entêtes, des colonnes, des lignes, etc.
- le journal contient des titres, des regroupements de notices, une episode de feuilleton, des notices de publicité, etc.
- dans le texte du journal, il y a partout des noms d'autre journaux, de personnages, des lieux, etc.
- le texte est (principalement) en français, d'un style d'interêt historique
- on note également des références aux evênements historiques
- et on a bien-sûr des infos supplementaires regardant la production, la dissemination, la bibliographie etc. de cette source...

Focalisons...

demandé la révocation de ce professeur imprudent.

L'organisation de l'artillerie de forteresse.

Le général Tricoche, directeur de l'artillerie au ministère de la guerre, sera nommé général de division et chargé de l'organisation des batteries d'artillerie de forteresse, avec le titre d'inspecteur général.

Il sera remplacé dans ses fonctions au ministère par le général Lavocat.

Tremblement de terre en Grèce.

Aujourd'hui, à deux heures du matin, un fort tremblement de terre a été ressenti au Pirée.

---

cour a M. de Sauvelat, et lui demande la

On peut se passer des balises ?

```
Un journal algérien annonce que M. Constantin, recteur de l'académie
d'Alger, a demandé la révocation de ce professeur imprudent.  Le
général Tricoche, directeur de l'artillerie au ministère de la
guerre, sera nommé général de division et chargé des l'organisation
des batteries d'artillerie de forteresse, avec le titre d'inspecteur
général.  Il sera remplacé dans ses fonctions au ministère par le
général Lavocat.  Aujourd'hui, à deux heures du matin, un fort
tremblement de terre a été ressenti au Pirée.  On télégraphie
d'Athènes qu'aucun accident de personnes n'est à déplorer.
```

- on ne mets en evidence que les mots des notices
- il n'y a pas de frontiere entre les deux notices parce que les parties structurantes sont supprimées
- la ponctuation est retenue telle quelle, y compris les liasons
- tous les mots semblent pareilles en terme de leur fonctionnement dans le texte

Balisage ASTARTEX

```
<ASTX_ATTR SOURCE=PETIT_COMTOIS DATE=1883-08-06 NUMERO=0006
PAGE=1 LARGEUR=1 EMPLACEMENT=4M
RUBRIQUE=DEPECHES_DE_NUIT_INFORMATIONS_GENERALES
RUB_GR=DEPECHES_DE_NUIT
TITRE=L'organisation_de_l'artillerie_de_forteresse
SIGNATURE=Service_de_notre_correspondant_spécial. > Le général
Tricoche, directeur de l'artillerie au ministère de la guerre, sera
nommé général de division et chargé des l'organisation des batteries
d'artillerie de forteresse, avec le titre d'inspecteur général.  Il
sera remplacé dans ses fonctions au ministère par le général Lavocat.
<ASTX_ATTR SOURCE=PETIT_COMTOIS DATE=1883-08-06 NUMERO=0006 PAGE=1_2
LARGEUR=1 EMPLACEMENT=4M_1H
RUBRIQUE=DEPECHES_DE_NUIT_INFORMATIONS_GENERALES
RUB_GR=DEPECHES_DE_NUIT
TITRE=Tremblement_de_terre_en_Grèce.
SIGNATURE=Service_de_notre_correspondant_spécial. >
Aujourd'hui, à deux heures du matin, un fort tremblement de terre a
été ressenti au Pirée.  0n télégraphie d'Athènes qu'aucun accident de
personnes n'est à déplorer.
```

- on ne mets en evidence que les mots des notices
- les parties structurantes sont transformées en annotation
- les annotations s'expriment comme propretés des mots

Un balisage TEI

```
<div  type="groupe"  decls="DDN"  n="4">
  <head  type="RUBRIQUE">DEPECHES DE NUIT</head>
  <head  type="SIGNATURE">Service de notre correspondant spécial.</head>
  <head  type="RUBRIQUE">INFORMATIONS GENERALES</head>
  <div  type="NOTICE"  rend="EMP.4M LG.1"
    n="4">
    <head  type="TITRE">L'organisation de l'artillerie de forteresse</head>
    <p>Le général Tricoche, directeur de l'artillerie au ministère de la guerre, sera
        nommé général de division et chargé des l'organisation des batteries d'artillerie de
        forteresse, avec le titre d'inspecteur général. </p>
    <p>Il sera remplacé dans ses fonctions au ministère par le général Lavocat.</p>
  </div>
  <div  type="NOTICE"  rend="EMP.4M LG.1"
    n="5">
    <head>Tremblement de terre en Grèce.</head>
    <p>Aujourd'hui, à deux heures du matin, un fort tremblement de terre a été ressenti au
        Pirée.</p>
    <p>0n télégraphie d'Athènes qu'aucun accident de personnes n'est à déplorer.</p>
  </div>
</div>
```

A noter...

- on met en evidence les mots des notices mais egalement les objets qu'ils constituent (paragraphes, titres...)
- (une selection des) objets textuels sont balisés explicitement avec leur fonction
- les annotations s'expriment comme propretés des objets auxquels elles sont attachées
- on se sert de XML, un syntaxe standardisé

XML: ce que c'est et pourquoi on devrait le connaitre

- XML est une manière de représenter les données structurées en forme de chaîne de caractères
- XML fournit une syntaxe tout à fait générique, à appliquer dans toute domaine de ressources numerisées
- 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

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)

Donc... on a des contenants emboîtés, avec des propriétés.

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

! pas de chevauchement

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

Continuons notre balisage TEI ...

```
<p>
 <s n="42">
  <w ana="#NOM" lemma="XXXX">0n</w>
  <w ana="#VER" lemma="télégraphier">télégraphie</w>
  <w ana="#PRP" lemma="de">d'</w>
  <w ana="#NOM" lemma="XXXX">Athènes</w>
  <w ana="#KON" lemma="que">qu'</w>
  <w ana="#PRO" lemma="aucun">aucun</w>
  <w ana="#NOM" lemma="accident">accident</w>
  <w ana="#PRP" lemma="de">de</w>
  <w ana="#NOM" lemma="personne">personnes</w>
  <w ana="#ADV" lemma="ne">n'</w>
  <w ana="#VER" lemma="être">est</w>
  <w ana="#PRP" lemma="à">à</w>
  <w ana="#VER" lemma="déplorer">déplorer</w>
  <c ana="#SENT">.</c>
 </s>
</p>
```

Encore de balisage TEI

```
<p>
 <persName key="#TRIC01">Le général Tricoche</persName>,
directeur de l'artillerie au
ministère de la guerre, sera nommé général de division et
chargé des l'organisation des
batteries d'artillerie de forteresse, avec le titre
d'inspecteur général.
</p>
<p>Il sera remplacé dans ses fonctions au ministère par
<persName ref="#LAVOC32">le
   général Lavocat</persName>.</p>
<!-- ... -->
<person xml:id="LAVOC32">
 <persName>
  <forename>Jean-Louis</forename>
  <surname>Lavocat</surname>
 </persName>
 <birth when="1823-02-09">ne le 9 fevrier 1823 a
Besancon</birth>
<!-- ...-->
</person>
<relation type="remplacement"
 active="#LAVOC32" passive="#TRIC01"/>
```

Et encore de balisage TEI...

```
<teiHeader>
  <fileDesc>
    <titleStmt>
      <title>Le Petit Comtois: TEI-XML Edition</title>
    </titleStmt>
    <publicationStmt>
      <p>Prepared for MISAT</p>
    </publicationStmt>
    <sourceDesc>
      <bibl>
        <title>Le Petit Comtois: Journal Républicain démocratique quotidien
</title>
        <ref>http://laseldi.univ-fcomte.fr/petit_comtois/archives.php</ref>
      </bibl>
    </sourceDesc>
    <encodingDesc>
      <classDecl>
        <taxonomy xml:id="notice-class">
          <category xml:id="DDN">
            <catDesc>Dépêches de nuit</catDesc>
          </category>
          <category xml:id="CR">
            <catDesc>Chroniques régionales</catDesc>
          </category>
        </taxonomy>
      </classDecl>
    </encodingDesc>
    <revisionDesc>
      <change when="2009-06-15">Taxinomies révisées</change>
    </revisionDesc>
  </fileDesc>
</teiHeader>
```

Le paysage actuel de la 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 (y compris l'oral)
- Documentation de balisage
- Et cetera : voir `http://www.tei-c.org/P5/Guidelines/`

… Bref : une sorte d'encyclopédie du balisage !

Composants de TEI ALL: les modules TEI P5

nom	chapitre P5
analysis	Simple Analytic Mechanisms
certainty	Certainty and Responsibility
core	Elements Available in All TEI Documents
corpus	Language Corpora
dictionaries	Dictionaries
drama	Performance Texts
figures	Tables, Formulae, and Graphics
gaiji	Representation of Non-standard Characters and Glyphs
header	The TEI Header
iso-fs	Feature Structures
linking	Linking, Segmentation, and Alignment
msdescription	Manuscript Description
namesdates	Names, Dates, People, and Places
nets	Graphs, Networks, and Trees
spoken	Transcriptions of Speech
tagdocs	Documentation Éléments
tei	The TEI Infrastructure
textcrit	Critical Apparatus
textstructure	Default Text Structure
transcr	Representation of Primary Sources
verse	Verse

Les TEI Guidelines

Dans ses 1 400 pp imprimées, vous trouverez :

un lexique et une grammaire 22 'modules' regroupant en total 564 éléments qui sont d'ailleurs classifiés en 189 classes

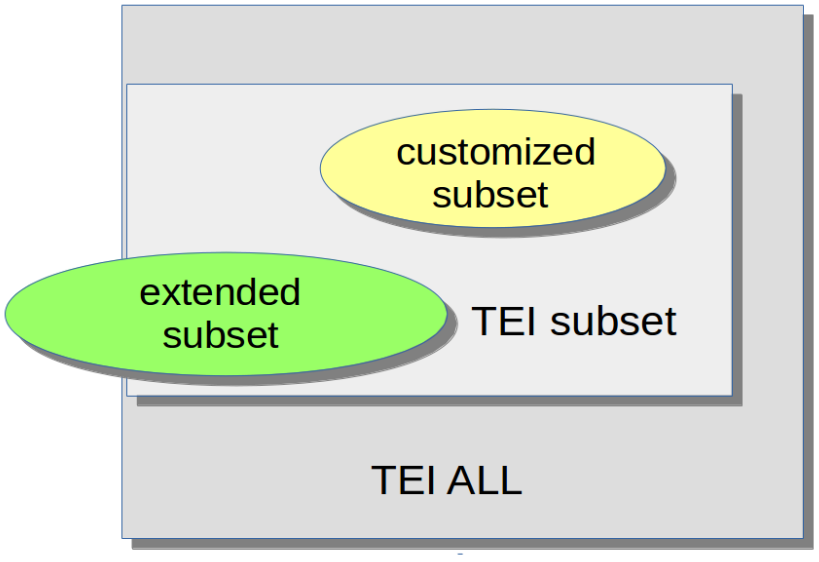
des contraintes additionelles 31 types de données, 63 constrainte/regles Schematron

des conventions d'utilisation *beaucoup* de prose

plusieurs exemples d'utilisation dont au moins un par élément

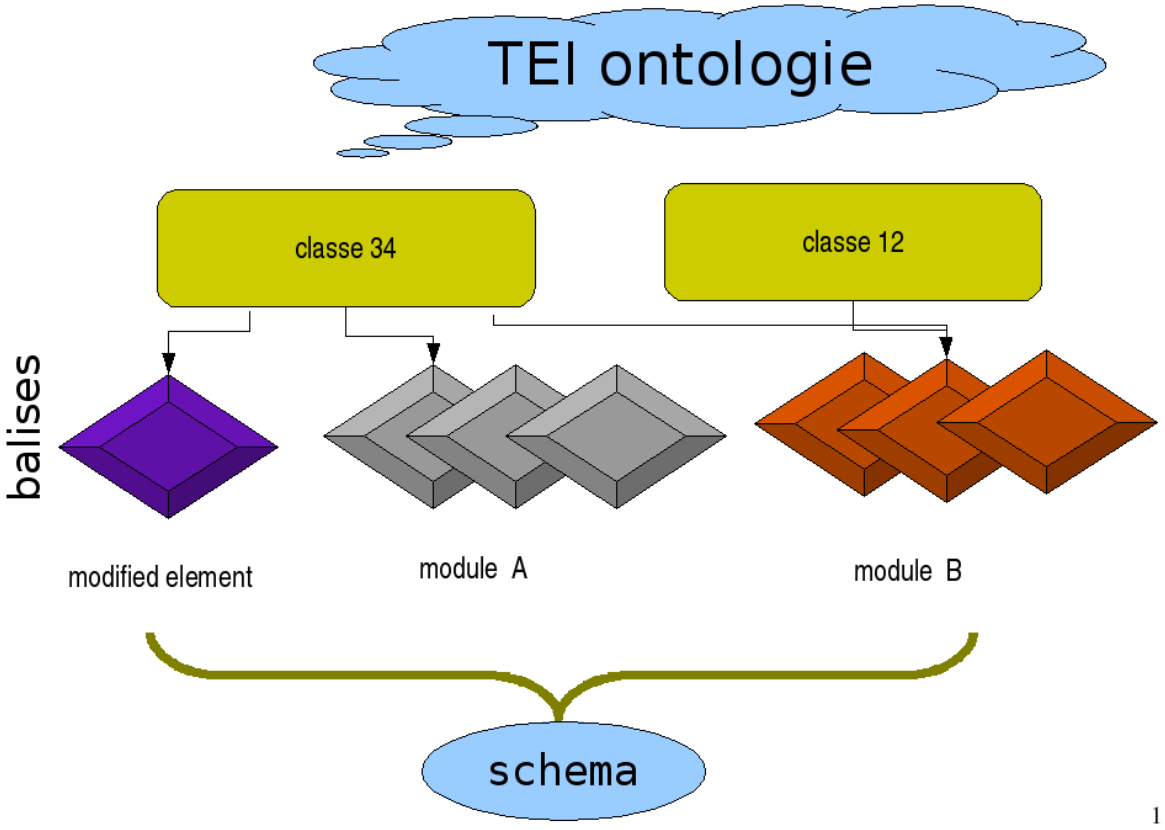
Comment régler ces richesses?

La TEI est conçue pour soutenir une variété d'approches



- on peut simplement utiliser un sous-ensemble de ses propositions (TEI subset)
- on peut y ajouter des contraintes supplémentaires (customized subset)
- on peut y ajouter de nouveaux composants (extended subset)

Architecture de la TEI



1

L'esprit TEI

Qu'est-ce que cela veut dire : utiliser la TEI ?

- une pratique de balisage consensuelle
- un lexique commun
- un respect de l'autonomie

La standardisation ne devrait pas signifier « fais comme moi » mais plutot « explique-moi ce que tu fais. »

Etre conforme à la TEI veut dire quoi?

- **être honnet** : Les éléments XML qui se déclarent comme appartenant au namespace TEI doivent respecter les définitions TEI de ces éléments
- **être explicite** : Pour valider un document TEI, un ODD est fortement conseillé, parce que cela mettra en évidence toutes les modifications effectuées
- Tout document valide TEI est valide par rapport au schéma TEI-ALL

L'objet de ces règles est de faciliter le "blind interchange" des documents.

"One Document Does it all (ODD)"

- Les Guidelines TEI sont produits à partir d'une même ressource XML qui contient:
  - de la prose descriptive (une grande quantité)
  - des exemples d'utilisations (plusieurs)
  - des déclarations formelles pour les constituants du modèle abstrait de TEI
  - les éléments et leurs attributs
  - les modules
  - les classes, et les macros
- On appelle cette ressource un ODD (bien qu'elle consiste en des centaines de petits fichiers)

Et alors?

- Vu ses visées ambitieuses, on ne peut se servir du système TEI qu'en le personnalisant
- Les personnalisations s'expriment également en langue ODD
- Cela permet de communiquer la manière où on a appliqué la TEI, sans perturber les normes

Exemple...

Un fichier TEI ODD est un document TEI comme les autres, avec autant de prose discursif qu'il faut, mais en plus un élément specialisé: un <schemaSpec> qui définit le schéma documenté

```
<text>
 <body>
  <div>
   <head>Notre manuel d'encodage</head>
   <p>Dans ce projet on ....</p>
   <schemaSpec ident="myTei"
    start="TEI">
    <moduleRef key="tei"/>
    <moduleRef key="header"/>
    <moduleRef key="core"
     except="sp stage speaker"/>
    <moduleRef key="textstructure"/>
    <elementSpec ident="head"
     mode="change" module="core">
     <content>
      <textNode/>
     </content>
    </elementSpec>
   </schemaSpec>
  </div>
 </body>
</text>
```

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 interface web permettant de construire des schemas TEI
oxGarage	services web pour transformation entre TEI/docx/html/epub etc.
TEI publisher	systeme d'édition numérique
Versioning Machine	systeme de gestion des variance
txm	outil textometrique capable de gerer directement TEI

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

oxygen	editeur (etc.) XML générique
saxon	moteur de transformation XML générique
basex; exist	systemes de base de données XML

Des références...

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

Depot github `https://github.com/TEIC`

Listes de discussion tei-l@listserv.brown.edu;
tei-fr@groupes.renater.fr (francophone)

Une excellente introduction francophone
http://books.openedition.org/oep/1237