

# La TEI pour les corpus linguistiques: un standard qui se renouvelle



# Combien de formats standardisés faut-il dans le monde ?



|<

< PREV

RANDOM

NEXT >

>|

PERMANENT LINK TO THIS COMIC: [HTTP://XKCD.COM/927/](http://xkcd.com/927/)

IMAGE URL (FOR HOTLINKING/EMBEDDING): [HTTP://IMGS.XKCD.COM/COMICS/STANDARDS.PNG](http://imgs.xkcd.com/comics/standards.png)

# Standards : un paysage complexe

Agences officielles de standardisation nationales : AFNOR, ANSI, BSI, DIN ; internationales : ISO, IEC, W3C, OASIS, ...

## Regroupements des Personnes/Industries Interessées

Plusieurs... par exemple

- LISA (Localisation Industry Standards Association)
- MPEG (Moving Pictures Expert Group)

Projets ayant des enjeux pré-normatifs En France par ex Cahier, Ortolang ....

Infrastructures de recherche Internationales : DARIAH, CLARIN ; Française : Huma-Num



# Comment caractériser les standards existants ?

WKWBFY un seul : solution centralisée

NWEUMP aucun : solution anarchiste

FTH autant qu'il en arrive : solution laissez-faire

## Les normes ne s'imposent pas dans la vie intellectuelle

- soit elles émergent d'un besoin de la communauté
- soit leur usage dérive de la nécessité d'utiliser une technologie particulière
- mais on ne renonce pas volontairement à son indépendance !

## Standards : on peut s'en passer?

Pour les scientifiques, les standards pourraient constituer un inconvénient :

- ils figent un état de la connaissance
- leur production est chronophage et nécessite des compétences interdisciplinaires
- leur utilisation peut être également chronophage

Quand même, il y a des avantages qu'il faut souligner

## Quelques besoins scientifiques

- ① Comment sur le web identifier et retrouver des ressources numériques ayant un intérêt linguistique ?
- ② Comment valider les résultats scientifiques obtenus par d'autres personnes ?
- ③ Comment enrichir ou intégrer les ressources existantes avec ses propres idées ?
- ④ Comment séparer les ressources des outils qui les gèrent/analysent ?

Pour tout cela, les standards restent essentiels

## Quelques besoins techniques

- ① possibilité de recombiner ou de réutiliser les systèmes existants
- ② évolution modulaire des logiciels
- ③ réduction des coûts de formation
- ④ existence de 'frequently answered questions' — des solutions qui s'appliquent dans plusieurs domaines

Les standards offrent ces possibilités !

## A big claim

Dès sa conception initiale, la TEI s'est proposée comme standard pour la construction, la description, la structuration, et l'annotation des corpus linguistiques et littéraires.

De tous types, de toutes périodes, dans toutes les langues



## La question qui s'impose :

- La TEI est *très ancienne* (née en 1987) !
- Elle précède le Web, le DVD, le téléphone portable, la télévision cablée, Microsoft Word..
- Les technologies informatiques qui survivent plus de 5 ans sont assez rares...
- Pourquoi et comment la TEI a-t-elle survécu plus de 30 ans ?

## Communautés scientifiques

Du point de vue historique, la TEI est un produit d'une rarissime conjonction d'interêts parmi ...

- littéraires, stylométriciens, critiques de l'école 'close reading'
- historiens, archivistes, éditeurs
- bibliographes, bibliothécaires
- linguistes, notamment, mais pas exclusivement, de corpus
- informaticiens ...

Tous concernés par le passage au numérique des textes écrits ou oraux à grande échelle



# Pourquoi tout cet effort ?

Parce qu'on s'est aperçu qu'on risquait une nouvelle confusion des langues avec l'arrivée de l'informatique dans la représentation des données textuelles !



# Oppositions

En même temps, la TEI prétend fournir une réponse pratique aux oppositions typiques des 'humanités numériques' :

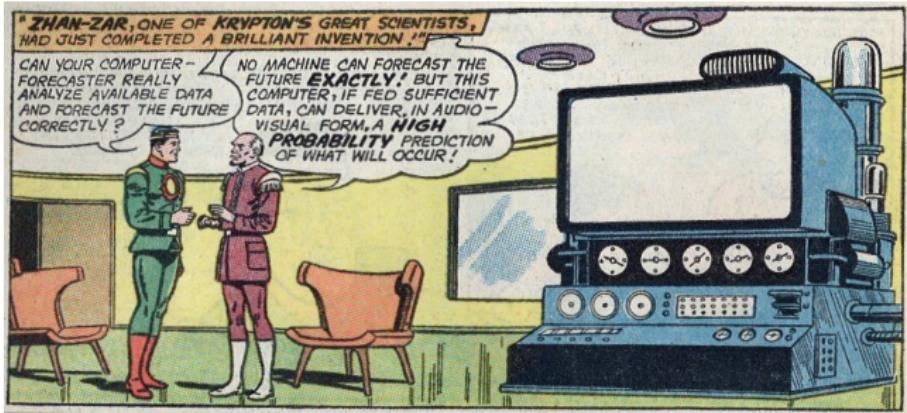
- ① entre les besoins des débutants et ceux des experts
- ② entre les besoins et les intérêts des scientifiques et ceux des ingénieurs

# Première opposition



- D'un côté, la TEI doit présenter des solutions préconnues, consensuelles, bien établies: les topoï sur lesquels ‘tout le monde s'est mis d'accord’ – et que les débutants doivent donc arriver à comprendre
- De l'autre, elle doit soutenir la recherche, et donc la découverte des solutions aux questions pas encore posées, ou posables – ainsi permettant aux experts de partager leur expertise.

## Deuxième opposition



- D'un côté, la TEI propose aux ingénieurs un système bien adapté à l'implantation avec des outils informatiques courants
- De l'autre, elle s'exprime dans une langue compréhensible aux non-techniciens et se base sur un modèle conceptuel adapté aux disciplines SHS

## Impérialisme

En dépit de son nom, la TEI ne s'adresse pas uniquement au texte proprement dit.

Même dans P1 (la version initiale) on trouve déjà des propositions assez complètes pour

- les métadonnées bibliographiques
- l'encodage des transcriptions orales
- les analyses linguistiques abstraites en termes de structures des traits

en complément des propositions pour l'encodage des structures traditionnelles du livre et leurs composants typiques

## Les non-enjeux de la TEI

À l'origine, la TEI ne s'intéressait pas à...

- le web (ça n'existe pas)
- la mise en page (tex, scribe...)
- l'intégration des pages-images/facsimilés numérisés
- la représentations des faits ou des objets (les bases de données)
- la production des logiciels

Elle se focalisait quand même sur : les métadonnées, les textes, les analyses textuelles et linguistiques

Nous avons changé tout cela



# La TEI actuelle facilite un balisage ‘intelligent’

Elle s'applique à l'encodage des...

- composants structuraux et fonctionnels d'un texte
- transcriptions diplomatiques des sources historiques, des images, des annotations
- liens, correspondances, alignements
- données et entités : par exemple de temps, personnes, lieux ou événements
- annotations péritextuelles et métatextuelles (correction, suppression, ajouts)
- analyses linguistiques
- métadonnées de plusieurs types
- ... et définitions formelles de schéma XML !

mais il faut sélectionner



# Interdisciplinarité

La TEI peut ainsi être considérée comme un des plus grands efforts d'interdisciplinarité de son époque: d'où deux principes de son architecture –

- l'application du rasoir d'Ockham
- les mécanismes de personnalisation

# Application du rasoir

- Il arrive souvent que les mêmes objets portent de noms divers et (moins souvent) que des objets divers ne sont pas distingués par le nom
- En souhaitant éviter une multiplication ingérable de concepts, la TEI fournit typiquement
  - une proposition générique (par ex `<div>`, `<q>`)
  - moins souvent, des propositions spécifiques (par ex `<said>`, `<quote>`, `<mentioned>`... )



## La personnalisation



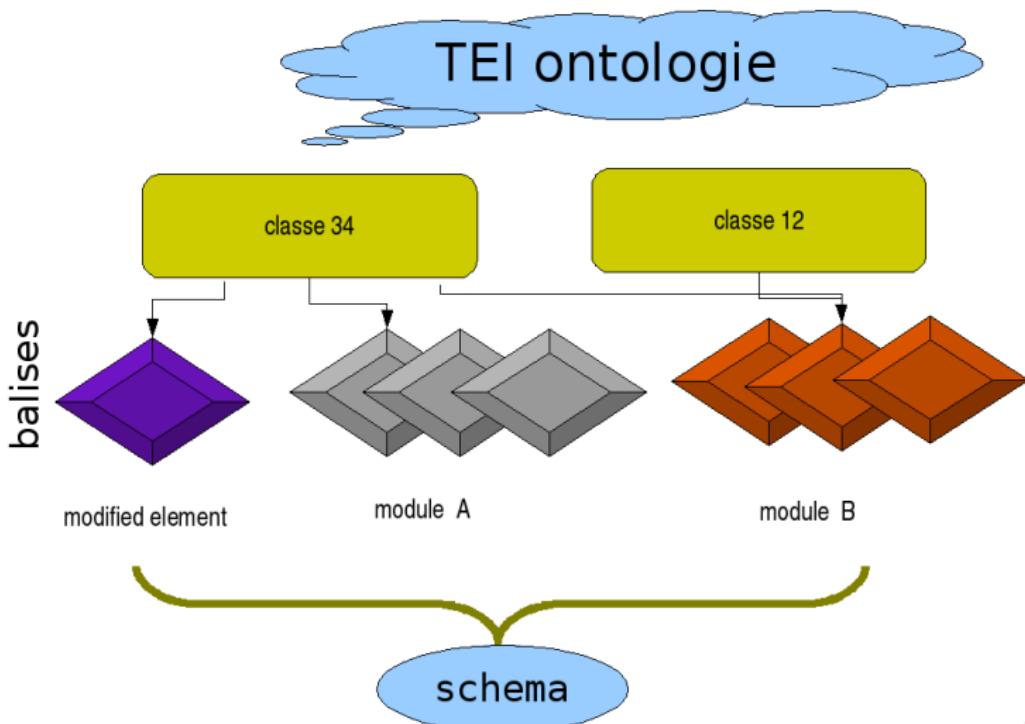
Le système TEI fournit un gabarit ou un kit lego pour la construction d'un système d'encodage bien adapté aux besoins de l'utilisateur, qui reste en même temps compréhensible par d'autres personnes ou systèmes, et dans lequel les modifications éventuelles se déclarent clairement.

*Extrait du ICAME Journal, 1992*

## Il n'y a pas de "TEI dtd"

- TEI est un système *modulaire*. On s'en sert pour créer un système d'encodage selon ses propres besoins, en sélectionnant des *modules* spécifiques
- Chaque module définit une brique contenant un groupe d'éléments (et leurs attributs)
- on peut sélectionner les éléments souhaités, et même en changer des propriétés
- on peut y mélanger des éléments nouveaux, ou bien natifs ou bien d'autres standards

# Organisation logique de la TEI



# Liste des modules TEI (1)

module	contents
analysis	Mécanismes simples d'analyse interprétative
certainty	Indications de (in)certeritude et de pertinence
core	Eléments communs à presque tout document
corpus	Eléments spécialisés pour la description des corpus
declarefs	Déclarations de système de traits
dictionaries	Eléments spécialisés pour l'encodage des dictionnaires
drama	Eléments spécialisés pour l'encodage des pièces de théâtre
figures	Encodage des tableaux, formules et figures
gaiji	Documentation des caractères non Unicode et des glyphes
header	Définition des métadonnées
iso-fs	Représentation des structures de traits
linking	Eléments spécialisés pour l'encodage des liens, de la segmentation et de l'alignement

## Liste des modules TEI (2)

module	contents
msdescriptions	Eléments spécialisés pour la description des sources manuscrites
namesdates	Encodage des entités nommées
nets	Encodage des graphies, réseaux, et arborescences
spoken	Eléments spécialisés pour la transcription orale
tagdocs	Eléments spécialisés pour la definition et documentation des systèmes d'encodage
tei	Definition des datatypes, des classes, et des macros utilisés par tout TEI module
textcrit	Eléments spécialisés pour l'apparat critique traditionnel
textstructure	Eléments structurants par défaut applicables à tout texte
transcr	Eléments spécialisés pour la transcription diplomatique ou génétique des sources primaires
verse	Eléments spécialisés pour le textes poétiques

# Les TEI Guidelines



# Les TEI Guidelines

Dans ses 1 400 pages imprimées, vous trouverez :

- **un lexique** de 564 éléments
- **des règles d'usage et des contraintes**
  - exprimées en langue naturelle
  - exprimées en langages formels
- **beaucoup de discussions** notamment de plusieurs exemples d'utilisation

L'accès online est maintenant plus pratique...

Par exemple : une spécification typique online

<http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-particDesc.html>



## TEI P5: derniers nouveautés

Il y a une version nouvelle de TEI P5 au moins 2 fois par an; le plus souvent contenant des fautes corrigées, de nouveaux exemples, etc.

La plus récente contient deux modifications importantes des possibilités fournies par le langage ODD de personnalisation. On peut désormais utiliser de nouveaux éléments TEI

- pour la définition des modèles de contenu
- pour la documentation des modèles de traitement

Prochainement, le dossier d'exemplaires contiendra TEI Simple un nouveau schéma réduit 'adapté aux besoins de 90% des utilisateurs TEI', qui est censé remplacer le vénérable TEI Lite



## Pourquoi continuer de s'intéresser à la TEI ?

Deux raisons pour lesquelles les standards échouent le plus souvent :

- ils sont basés sur une théorie pas encore mûre
- 'not invented here': la communauté envisagée est trop diverse ou fragmentée

Le secret de la réussite de la TEI et de sa longévité serait-il dans ses capacités d'évolution ?



# Comment faire mûrir une théorie ?

Une personnalisation TEI peut :

- faire une sélection explicite des éléments et des attributs considérés comme utiles
- limiter les valeurs possibles d'un attribut plus ou moins strictement
- proposer des règles Schematron sur le contenu d'un élément (p.e. co-dependency)
- ajouter de nouveaux éléments (ou classes d'élément), labellisés dans votre propre espace de noms, mais appartenant aux classes sémantiques prédéfinies par la TEI

Donc on peut évoluer et tester sa théorie, en restant toujours TEI-conforme.



## Not Invented Here?

- TEI P5 a des possibilités très extensives pour l'I18N...
- TEI héberge volontairement d'autres espaces de noms
- Donc on peut se servir des autres schémas existants :
  - SVG pour les graphiques
  - MathML pour les maths
  - DCMI pour les métadonnées
  - ...
- La définition d'un élément TEI peut inclure (s'il y en a) son mapping avec d'autres ontologies, formalisé par un élément `<equiv>` (équivalent)



# Un standard existe pour qu'on s'y conforme, non ?

## The TEI Commandments

- I. Thou shalt have no other encoding scheme but this one
- II. Honour the consensus that thy days may be long in this land
- III. Thou shalt not take the GIs of this scheme in vain
- IV. Thou shalt not commit polysemy

⟨Text Encoding Initiative

650

November 1991⟩



## L'esprit TEI

Qu'est-ce que cela veut dire : « être conforme » à la TEI ?

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

La standardisation ne doit pas signifier « fais comme moi » ; elle veut dire « explique-moi ce que tu fais. »



## L'évolution darwinienne, ça marche...

- faites vos modifications dans votre espace de nom
- documentez-les dans un ODD
- faites discuter vos propositions sur la liste TEI-L, ou dans un SIG !
- proposez les modifications efficaces au Conseil Scientifique de la TEI, en faisant une "feature request" sur sourceforge
- Il y a une version nouvelle de TEI P5 deux fois par an...

... et n'oubliez pas de vous abonner au Consortium !

