

La Text Encoding Initiative: un standard qui se renouvelle



Qu'est-ce que la Text Encoding Initiative (TEI) ?



- Une organisation, une institution ?
- Un 'club', une mode, une religion ?
- Une spécification technique ?
- Un gabarit pour la construction des spécifications techniques ?



Version française: <http://books.openedition.org/oep/1237>

Vous souvenez-vous de l'an 1987 ?

La Text Encoding Initiative est née dans un monde très différent...

- C'était l'été de *Joe le taxi*, premier tube de Vanessa Paradis...
- le world wide web n'existe pas
- le tunnel sous la manche était en construction
- un état nommé l'Union Soviétique venait de lancer une station spatiale appelée "Mir" .. et de subir un désastre à Tchernobyl
- l'informatique sérieuse s'effectuait uniquement sur les grosses machines dites 'mainframes'



Vous souvenez-vous de l'an 1987 ?

La Text Encoding Initiative est née dans un monde très différent...

- C'était l'été de *Joe le taxi*, premier tube de Vanessa Paradis...
- le world wide web n'existe pas
- le tunnel sous la manche était en construction
- un état nommé l'Union Soviétique venait de lancer une station spatiale appelée "Mir" .. et de subir un désastre à Tchernobyl
- l'informatique sérieuse s'effectuait uniquement sur les grosses machines dites 'mainframes'

Vous souvenez-vous de l'an 1987 ?

La Text Encoding Initiative est née dans un monde très différent...

- C'était l'été de *Joe le taxi*, premier tube de Vanessa Paradis...
- le world wide web n'existe pas
- le tunnel sous la manche était en construction
- un état nommé l'Union Soviétique venait de lancer une station spatiale appelée "Mir" .. et de subir un désastre à Tchernobyl
- l'informatique sérieuse s'effectuait uniquement sur les grosses machines dites 'mainframes'



Vous souvenez-vous de l'an 1987 ?

La Text Encoding Initiative est née dans un monde très différent...

- C'était l'été de *Joe le taxi*, premier tube de Vanessa Paradis...
- le world wide web n'existe pas
- le tunnel sous la manche était en construction
- un état nommé l'Union Soviétique venait de lancer une station spatiale appelée "Mir" .. et de subir un désastre à Tchernobyl
- l'informatique sérieuse s'effectuait uniquement sur les grosses machines dites 'mainframes'



Vous souvenez-vous de l'an 1987 ?

La Text Encoding Initiative est née dans un monde très différent...

- C'était l'été de *Joe le taxi*, premier tube de Vanessa Paradis...
- le world wide web n'existe pas
- le tunnel sous la manche était en construction
- un état nommé l'Union Soviétique venait de lancer une station spatiale appelée "Mir" .. et de subir un désastre à Tchernobyl
- l'informatique sérieuse s'effectuait uniquement sur les grosses machines dites 'mainframes'



Vous souvenez-vous de l'an 1987 ?

La Text Encoding Initiative est née dans un monde très différent...

- C'était l'été de *Joe le taxi*, premier tube de Vanessa Paradis...
- le world wide web n'existe pas
- le tunnel sous la manche était en construction
- un état nommé l'Union Soviétique venait de lancer une station spatiale appelée "Mir" .. et de subir un désastre à Tchernobyl
- l'informatique sérieuse s'effectuait uniquement sur les grosses machines dites 'mainframes'



...mais aussi dans un monde un peu familier...

- Les disciplines "linguistique de corpus" et "intelligence artificielle" avaient établi la nécessité de travailler avec des ressources numérisées et à grande échelle
- Des avancées en traitement de texte commençaient à avoir un effet sur la lexicographie et les systèmes de gestion documentaire (TeX, Scribe, tRoff..)
- L'Internet existait, et les théories sur comment en profiter d'une manière 'hypertextuelle' abondaient
- On confrontait déjà les problèmes de pérennisation des données et d'incompatibilités technologiques (ex. les CD).

Naissance de la Text Encoding Initiative

- printemps 1987 : En Europe, des réunions sur la possibilité de standardisation des données et sources historiques (J.P. Genet, M. Thaller)
- automne 1987 : Aux États-Unis, la NEH finance une réunion internationale sur la possibilité de définir des "text encoding guidelines"



La question qui s'impose :

- Donc, la TEI est *très ancienne* !
- 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 ?



Les enjeux de la TEI

Reconnaissant les possibilités démotiques du numérique...
l'initiative « **Text Encoding for Interchange** » s'est donnée comme mission :

- de faciliter la **création, l'échange, et l'intégration** des données textuelles informatisées
 - pour toute sorte de texte
 - dans toutes les langues
 - de toute origine temporelle ou culturelle
- La TEI s'adresse également ...
 - aux débutants, cherchant des solutions bien connues et consensuelles
 - aux experts, cherchant à créer de nouvelles solutions

Pourquoi cet effort ?



- Parce qu'on s'est aperçu qu'on risquait une nouvelle confusion de langues avec l'arrivée de l'informatique dans la représentation des données textuelles !
- Mais aussi peut-être un désir de mettre à jour les traditions philologiques de la gestion des textes?

Phases de la TEI

- 1988 - 1994 Développement en projet de recherche internationale; versions P1 (1990), P2 (1992), P3 (1994)
- 1995 - 1999 Promotion et prise en main (pas financée !)
- 2000 **Établissement du Consortium TEI**
- 2001 - 2003 Conversion de P3 en XML (TEI P4), lancement d'une révision complète qui apparaîtra comme TEI P5
- 2003 - ? TEI P5 : révisions régulières 2 fois par an; 36 releases sur github depuis 2005; la version 3.0.0 va apparaître



1988: Un temps de transition, et d'évolution

- Les 'Humanities Computing' étaient en train d'apparaître, comme 'interdiscipline'
- les informaticiens et les gens des SHS se regardaient (avec un peu de méfiance)
- dans quelques centres informatiques universitaires on s'est aperçu qu'il fallait faire de la recherche pour maintenir les services au niveau souhaité
- dans quelques centres de recherche on s'est aperçu des possibilités impressionnantes de l'informatique...

The Poughkeepsie Principles

Closing Statement of Vassar Conference The Preparation of Text Encoding Guidelines

Poughkeepsie, New York
13 November 1987

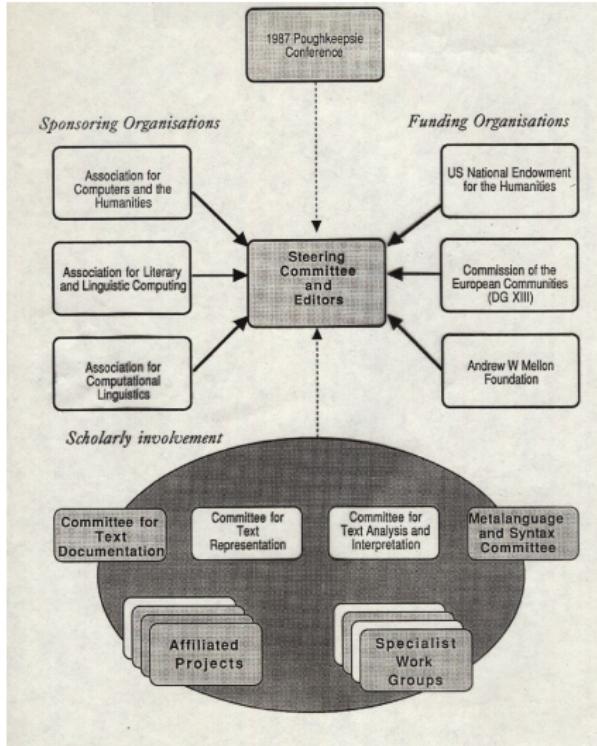
1. The guidelines are intended to provide a standard format for data interchange in humanities research.
2. The guidelines are also intended to suggest principles for the encoding of texts in the same format.
3. The guidelines should
 1. define a recommended syntax for the format,
 2. define a metalanguage for the description of text-encoding schemes,
 3. describe the new format and representative existing schemes both in that metalanguage and in prose.
4. The guidelines should propose sets of coding conventions suited for various applications.
5. The guidelines should include a minimal set of conventions for encoding new texts in the format.
6. The guidelines are to be drafted by committees on
 1. text documentation
 2. text representation
 3. text interpretation and analysis
 4. metalanguage definition and description of existing and proposed schemes,
coordinated by a steering committee of representatives of the principal sponsoring organizations.
7. Compatibility with existing standards will be maintained as far as possible.
8. A number of large text archives have agreed in principle to support the guidelines in their function as an interchange format. We encourage funding agencies to support development of tools to facilitate this interchange.
9. Conversion of existing machine-readable texts to the new format involves the translation of their conventions into the syntax of the new format. No requirements will be made for the addition of information not already coded in the texts.

The principles agreed upon at the Poughkeepsie Planning Conference are expounded in more detail and supplemented with other material in the sections which follow.

<http://www.tei-c.org/Vault/ED/edp01.htm>



Organisation de la TEI (1991)



Les travaux de la TEI ont été pris en main par les deux 'editors' et par quatre 'working committees'

- Documentation : bibliothécaires/archivistes
- Métalanguage : informaticiens
- Text Analysis and Interprétation : linguistes théoriques
- Text Representation : ... le reste

Opposition
analyse/représentation

Travaux de mutualisation

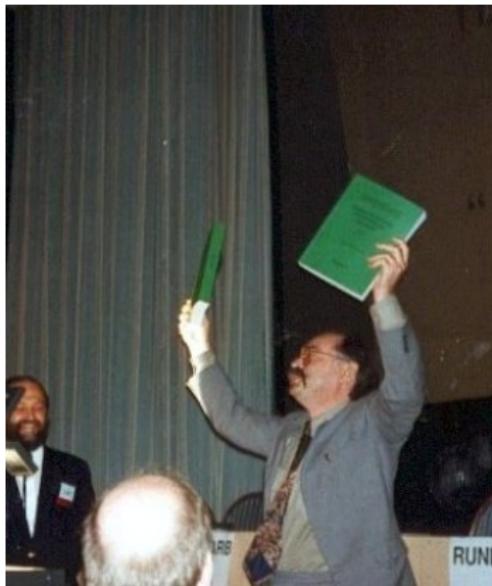
On a très vite compris qu'il y avait beaucoup de chevauchements parmi ces travaux. Les deux TEI Editors essayaient de participer aux débats de chaque comité, et d'appliquer, aussi rigoureusement que possible, le célèbre **rasoir d'Ockham**.

Néanmoins, la TEI propose plusieurs systèmes de représentation pour :

- la segmentation linguistique
- les annotations interprétatives (à plusieurs niveaux) avec des codes
- la documentation des codes interprétatifs
- des balisages effectués en ligne, et également en 'standoff'
- ...

(Encore une raison d'éviter l'usage de TEI All)

1994 : P3



- avril : TEI P3 est annoncé au colloque ALLC-ACH à Paris
- mai : Les 'green books' apparaissent enfin au colloque international SGML à Montreux
- déc : Le premier 'TEI Metaworkshop' a lieu à Chicago

1994-1999

L'adoption de la TEI, et l'influence de ses idées est difficile à tracer, parce qu'elle est devenue une partie de l'écosystème informatique qui était en état très rapide d'évolution à cette époque.

- En 1996, Michael Sperberg McQueen, l'éditeur principal de la TEI, fut nommé co-éditeur du standard W3C XML
- En 1997, on célébrait le dixième anniversaire de la TEI par un colloque à Brown University
- En 1998, une réunion organisée par le DLF à Washington parlait déjà de migrer la TEI de SGML en XML.
- En 1999 apparaissait une version de P3 légèrement révisée avec des corrections, et un ajout (la balise `<ab>`, à savoir)

Who owns the output of an international collaborative research project? Who has the right and duty to maintain it?

2000 : Naissance du TEI Consortium

Suite à des travaux sérieux de la part de plusieurs utilisateurs de la TEI (notamment à Londres, Virginia, Brown, Oxford, Bergen...) le TEI Consortium a été établi comme association à but non lucratif en 2000.

Enjeux du consortium (à part les détails bureaucratiques) :

- de garantir l'entretien du système TEI
- de mettre en place des mises à jour urgentes :
 - version XML
 - élargissement des sujets traités
- définir un modèle économique et scientifique qui permette une pérennité aux efforts de la communauté TEI

<http://www.ariadne.ac.uk/issue24/tei/>

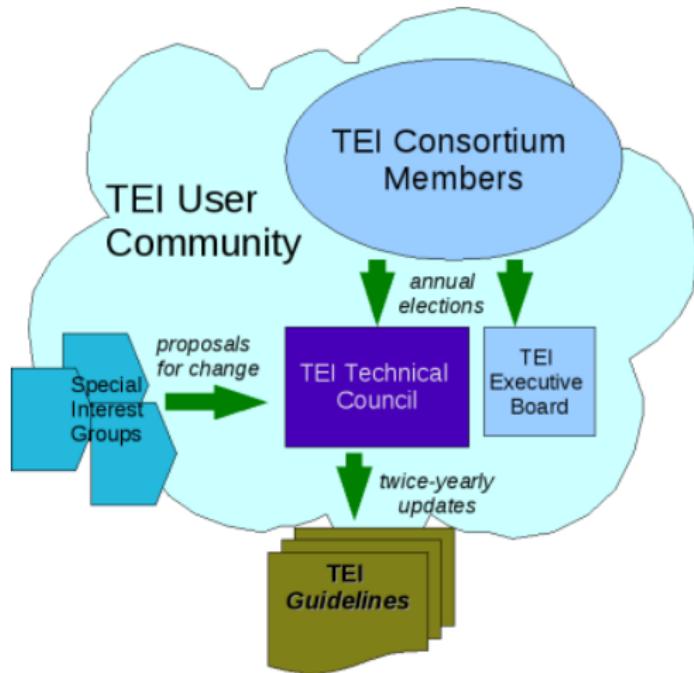


2001 : Première réunion annuelle des membres du TEI Consortium (à Pise)

- voir <http://www.tei-c.org/Membership/Meetings/2001/index.xml>
- Intervention de Michael Sperberg McQueen : The TEI is Dead : Long Live the TEI
- Intervention technique de Syd Bauman et Lou Burnard sur la feuille de route pour la P5



TEI organigramme (aujourd'hui)



TEI n'est plus un projet de recherche

- Un projet basé sur la communauté
- Évolution, gestion, et maintenance par un Conseil Scientifique de 12 personnes
- Les conseillers sont élus pour une période fixe par les membres payants du consortium
- Possibilité d'adhérer au Consortium à titre personnel ou institutionnel

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 seulement sur : les métadonnées, les textes, les analyses textuelles et linguistiques

Nous avons changé tout cela

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 !



Le cornucopia de balises TEI



Les TEI Guidelines

Dans ses 1 400 pp imprimées, vous trouverez :

un lexique et une grammaire 22 'modules' regroupant en total 521 élément qui sont d'ailleurs classifiés en 146 classes

des règles d'usage 7 185 lignes de règles exprimées en grammaire RELAXNG

des contraintes additionnelles 21 types de données, plusieurs règles formalisées en Schematron

des règles / conventions d'utilisation beaucoup de prose

plusieurs exemples d'utilisation dont au moins un par élément

Comment régler ces richesses?

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. »

... d'où les variations TEI

Par exemple : éléments pour description bibliographique : On a le choix entre

- **<bibl>** qui contient n'importe quel mélange de composants bibliographiques ... ou aucun
- **<biblStruct>** qui contient une sélection prédéfinie d'éléments, strictement structurés

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.

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

Niveaux de validation

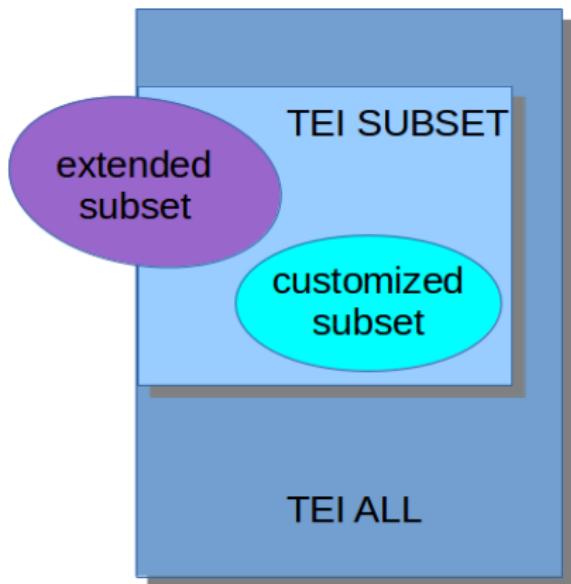
Un document TEI-XML doit:

- ① respecter les règles syntaxiques d'XML;
- ② être valide par rapport à un schéma quelconque;
- ③ respecter la sémantique définie de chaque élément TEI utilisé

Un schéma exprime d'une manière formelle et validable par logiciel une partie importante de ces règles

Un TEI ODD nous permet de construire un schéma adéquat à nos besoins

Mais la TEI est quand-même 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)

Besoins d'un projet d'informatisation

Nous aurons besoins de plusieurs choses :

- un schéma formel (en langue informatique tel que DTD, RELAXNG, W3C Schema, Schematron) qui peut contrôler :
 - quelles balises sont disponibles ?
 - dans quels contextes ?
 - avec quels attributs ?
 - avec quelles valeurs ?
 - en respectant quelles contraintes ?
- une documentation pour expliquer nos principes éditoriaux, nos principes de choix de balises, etc. aux utilisateurs/developpeurs :
 - en plusieurs langues naturelles
 - en plusieurs formats bureautiques (PDF, Word, HTML, epub...)
- des outils informatiques pour transformer et gérer nos données XML

Propositions de la TEI

Ayant elle-même ces mêmes besoins, la TEI vous propose :

ODD Un vocabulaire XML pour définir les vocabulaires XML
Stylesheets Un ensemble de feuilles de style XSLT très générique pour la conversion des documents XML TEI

Roma

Un outil web pour créer et traiter des documents ODD, pour en faire ressortir

- des schémas RELAXNG, DTD, etc.
- des manuels "mode d'emploi" en HTML, PDF etc.

OxGarage

Un outil web pour effectuer plusieurs transformations documentaires, par ex

- de DOCX en TEI, et l'inverse
- de TEI en HTML, et l'inverse
- de TEI en JSON



L'idée essentielle de ODD

One Document Does it all

Un vocabulaire spécialisé pour la définition

- des schémas
- des types d'élément XML, indépendant des schémas
- des types de donnée ou 'datatype's
- des patrons (MLE macros)
- des classes (et sous-classes)

Pour la définition des références également, ainsi permettant de réunir dans un schéma

- des objets identifiables (dans la liste ci-dessus)
- des objets prédéfinis notamment par la TEI
- des objets appartenant à d'autres espaces de nommage

et qui serait intégrable à un système de balisage documentaire assique

Par exemple ...

Dans notre projet de transcription collaborative nous souhaitons utiliser

- une gamme très réduite des balises TEI
- une liste très contrainte de possibilités de valeur pour quelques attributs
- un élément pas encore prévu par la TEI

L'usage d'un schéma XML pour renforcer ces contraintes simplifie énormément la création d'une interface ergonomique, bien adaptée aux utilisateurs prévus

Exemple fictif (1)

```
<body>
  <head>Une personnalisation TEI pour la
transcription collaborative</head>
  <p>Cette personnalisation propose un
schéma minimal pour la transcription
collaborative des documents archivés.
</p>
  <schemaSpec ident="transMin"
    start="TEI text div" docLang="fr">
    <moduleRef key="tei"/>
    <moduleRef key="header"
      include="teiHeader fileDesc
titleStmt publicationStmt sourceDesc"/>
    <moduleRef key="textstructure"
      include="TEI text body div"/>
    <elementRef key="ab"/>
    <elementRef key="pb"/>
    <elementRef key="unclear"/>
    <elementRef key="hi"/>
    <elementRef key="name"/>
    <elementRef key="title"/>
    <classRef key="att.global.rendition"
      except="rendition style"/>
    <classSpec type="atts"
      ident="att.declaring" mode="delete"/>
    <classSpec type="atts"
      ident="att.edition" mode="delete"/>
    <classSpec type="atts"
      ident="att.editLike" mode="delete"/>
  <schemaSpec>
    dy>
```

- un peu de documentation
- un `<schemaSpec>` identifiable, précisant une langue de documentation et des éléments racine
- une sélection d'éléments
- suppression de plusieurs attributs

Exemple fictif (2)

```
<body>
  <head>Une personalisation TEI pour la
transcription collaborative</head>
  <p>Cette personalisation propose un
schéma minimal pour la transcription
collaborative des documents archivés.
</p>
  <schemaSpec ident="transMin"
    start="TEI text div" docLang="fr">
<!-- ... -->
  <elementSpec ident="hi"
    mode="change">
    <attList>
      <attDef ident="rend"
        mode="replace">
        <valList type="closed">
          <valItem ident="underline"/>
          <valItem ident="superscript"/>
        </valList>
      </attDef>
    </attList>
  </elementSpec>
<!-- ... -->
  </schemaSpec>
</body>
```

- la spécification existante pour `<hi>` est modifiée
- la spécification de son attribut `@rend` est remplacée
- la liste des valeurs possibles pour cet attribut est fermée

Exemple fictif (3)

```
<body>
  <head>Une personnalisation TEI pour la
transcription collaborative</head>
  <p>Cette personnalisation propose un
schéma minimal pour la transcription
collaborative des documents archivés.
</p>
  <schemaSpec ident="transMin"
    start="TEI text div" docLang="fr">
<!-- ... -->
  <elementSpec ident="botName"
    ns="http://monexcellentprojet.com">
    <desc>nom botanique</desc>
    <classes>
      <memberOf key="model.phrase"/>
      <memberOf key="att.global"/>
    </classes>
    <content>
      <macroRef key="macro.paraContent"/>
    </content>
  </elementSpec>
<!-- ... -->
  </schemaSpec>
</body>
```

- nous ajoutons une spécification pour un élément non-TEI, appartenant à une autre espace de nommage
- cette spécification comporte
 - une description
 - une indication des classes TEI auxquelles l'élément appartiendrait
 - une indication de son contenu possible

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

Comment faire mûrir une théorie ?

Dans son TEI ODD, on peut :

- limiter les valeurs possibles d'un attribut plus ou moins strictement
- proposer des règles Schematron sur le contenu (p.e. co-dependency)
- enlever quelques éléments facultatifs
- ajouter de nouveaux éléments, labellisés dans votre propre espace de nommage

Donc on peut évoluer et tester une théorie précoce, 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 metadonnées
 - ...
- 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)

Mais, au fond, le modèle textuel proposé par la TEI reste proche à un modèle très répandu: très intuitif

L'évolution darwinienne, ça marche...

- faites vos extentions dans votre propre espace de nommage
- 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...
- Il y a une version nouvelle de TEI P5 deux fois par an...

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

