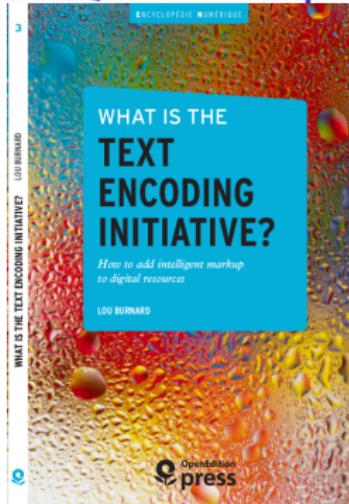


# Une brève histoire de la Text Encoding Initiative

# 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 ?

# Vous souvenez-vous de l'an 1987 ?

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

- le world wide web n'existait 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 "mainframes"



# Vous souvenez-vous de l'an 1987 ?

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

- le world wide web n'existait 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 "mainframes"



# Vous souvenez-vous de l'an 1987 ?

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

- le world wide web n'existait 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 "mainframes"

# Vous souvenez-vous de l'an 1987 ?

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

- le world wide web n'existait 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 "mainframes"

# Vous souvenez-vous de l'an 1987 ?

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

- le world wide web n'existait 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 "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 tout 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 !

## Phases de la TEI

- 1988 - 1990 Développement cycle 1 : Production de TEI P1  
(consultation avec une cinquantaine d'experts mondiaux)
- 1990 - 1992 Développement cycle 2 : production des fascicules TEI P2 (en consultation avec plusieurs groupes de travail... un ensemble de quelques centaines d'experts)
- 1993 - 1994 Intégration des fascicules P2 comme TEI P3 : la version "finale"
- 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 sur sourceforge ; des révisions régulières jusqu'à présent (on est à la version 2.7.0)



## 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 recherches on s'est aperçu des possibilités impressionnantes de l'informatique...

## Les 'experts' dans quelles domaines?

- ceux qui s'intéressaient à la création des corpus numérisés pour :
  - linguistique de corpus
  - fonds littéraires, études stylistiques etc.
  - édition scientifique (cfr. systèmes bureautiques naissants)
- ceux qui s'occupaient de la recherche linguistique :
  - lexicologie et lexicographie
  - systèmes de "compréhension" artificielle
  - systèmes de génération de langue naturelle
- ceux qui voulaient expérimenter l'application du numérique à l'édition et à la diffusion des sources littéraires ou historiques
- ceux qui s'occupaient du catalogage et de la documentation des ressources numérisées...
  - aux bibliothèques universitaires
  - aux archives numérisées des sciences sociales

# 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.

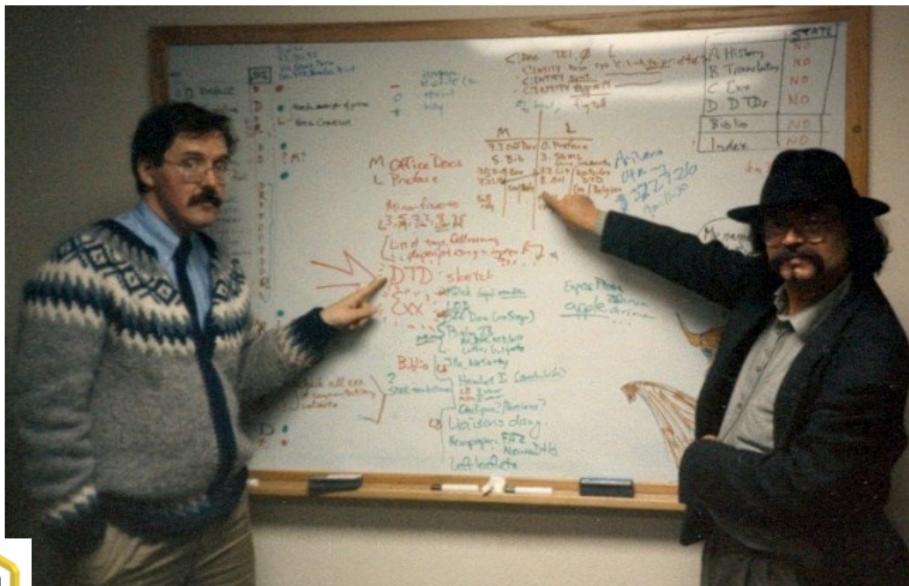
The principles agreed upon at the Poughkeepsie Planning Conference are expounded in more detail and supplemented with other material in the sections which follow.

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

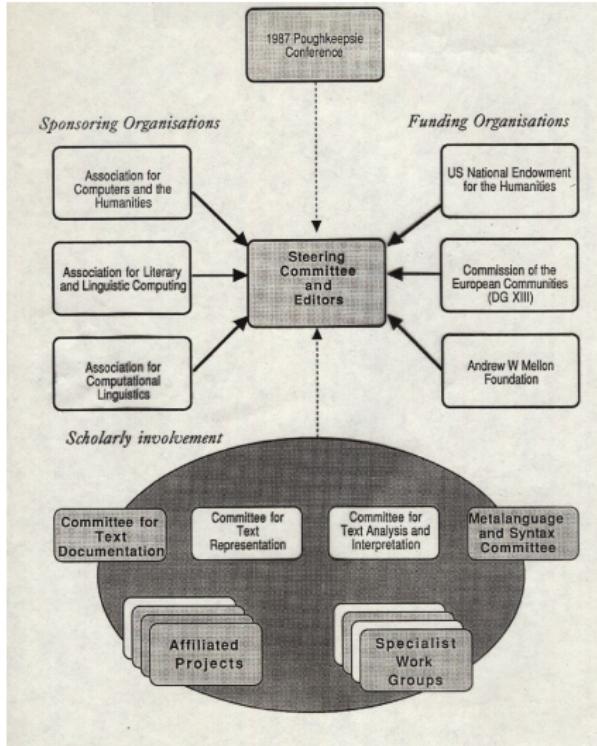


# TEI jadis...

- Un projet de recherche
- Au centre, deux éditeurs et un comité de pilotage, gérant les propositions d'un ensemble des comités d'experts/groupes de travail



# 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

# Text Documentation

: <para> The purpose of the Text Documentation Committee is to define an SGML-based tagset adequate to document electronic texts. This tagset will describe the tags used in an electronic text and also the relationship between it and its non-electronic source or sources. The Committee is able to take advantage of substantial work already carried out by the library and archive community in establishing bibliographic methods appropriate to machine readable datasets, and its work should be completed by the end of the first funding cycle.

Comment décrire une ressource numérisée, afin de la faire figurer correctement et avec utilité dans les catalogues des BU ?

Ils ont inventé : le TEI Header comme 'source d'information unique', analogue numérique du page de titre classique

Opposition métadonnées obligatoires standardisées/besoins de répondre aux besoins des chercheurs (dans un domaine inexploré...)

# Metalanguage

<para> The fourth committee, on Metalanguage and Syntax, has a somewhat different role from the others. Rather than proposing tagsets, it is charged with identifying a subset of SGML features appropriate to the needs of the Guidelines, and with assisting in the task of defining formal DTDs derived from the tagsets where these are thought advisable. It will also provide assistance in translating between other existing encoding schemes and that finally proposed by the Guidelines.

Il s'est décidé très vite d'adopter la norme ISO SGML -- pourvu que ce dernier serve à répondre aux besoins de modélisation exprimés par les chercheurs.

Dès le début, la TEI modélise un système de balisage indépendamment de sa réalisation dans une syntaxe concrète.

Voir <http://www.tei-c.org/Vault/ED/edw05.txt>

# Text Representation

<para> The Text Representation Committee is charged with the larger task of identifying tagsets for all those structural features of running text for which typographic conventions already exist in printed texts. Workgroups have been set up to consider character set representations, the structural elements of particular types of texts (notably technical, historical, philosophical, legal and literary texts) and of corpora or representative collections of text samples. It is expected that a set of common or core tags will emerge from this work, together with suggestions for recommended extensions.

Coordonné par un linguiste de corpus (Stig Johansson), ce comité était constitué de gens que l'on appellerait aujourd'hui des 'digital humanists'

Sa mission : identifier et donner les noms aux 'particularités signifiantes'

## Particularités signifiantes des textes écrits

- structuration et référencement (paragraphs, lists...)
- variations de mise en forme considérées d'importance
- ponctuation, usage des caractères, etc.
- annotations et liens hypertextuels
- variations éditoriales (correction, ajouts, etc.)
- ...

What is text **really**?

# Analysis and Interpretation

<para> By contrast, the Text Analysis and Interpretation Committee addresses structural aspects of text for which no agreed typographic conventions exist. Clearly this is an open-ended task: initially therefore, the committee will focus on the specific area likely to be of most general benefit: that of identifying features of use in linguistic analyses. Three of its workgroups will define tagsets for use in phonological and phonetic analysis, in lexical and morphological analysis, and in syntactic analysis respectively; a fourth, which has already completed a draft proposal, will address the tagging of natural language dictionaries, and the relationship between such tagsets and those appropriate to tagging the structure of electronic lexica.

Coordonné par un linguiste théorique (Terry Langendoen), ce comité était constitué de gens que l'on appelerait aujourd'hui les 'linguistes computationnels'

Sa mission ambitieuse : identifier et représenter toute la gamme des analyses linguistiques et littéraires.

# Analyses linguistiques

Voir [http://www.tei-c.org.uk/Vault/AI/air03.txt...](http://www.tei-c.org.uk/Vault/AI/air03.txt)

- Underspecification, uncertainty, multiple hierarchies...
- Phonology and prosody
- Morphology and word-level tagging
- Higher level syntactic analysis
- Structural ambiguity
- Anaphora and Deixis
- Idioms
- Figures of speech (not for this cycle)
- etc. etc.

On propose des sous-comités...



## L'impératif EAGLES

En Europe il y avait un fort désir de la standardisation des ressources numériques, associé à un financement majeur pour les systèmes d'«ingénierie linguistique» à l'époque.

- l'affaire Eurotra
- une convergence d'intérêts scientifiques/commerciaux
- l'établissement des industries dites langagières

Malheureusement une énumération complète des concepts linguistiques n'est pas tellement évidente...

## Structures de traits

Au lieu de cela, la TEI propose un "meta-modèle" pour décrire n'importe quelle espèce d'annotation linguistique... et de les rendre mutuellement compréhensible, voire unifiable

Ce modèle opère à deux niveaux :

- **représentation** de l'analyse comme un lot de traits, typés, structurés, et linéarisés en XML
- définition d'un **système de traits**, représentant les contraintes sur les valeurs intégrées, et les règles à suivre pour les interpréter, surtout du point de vue d'une grammaire d'unification

## Travaux de mutualisation

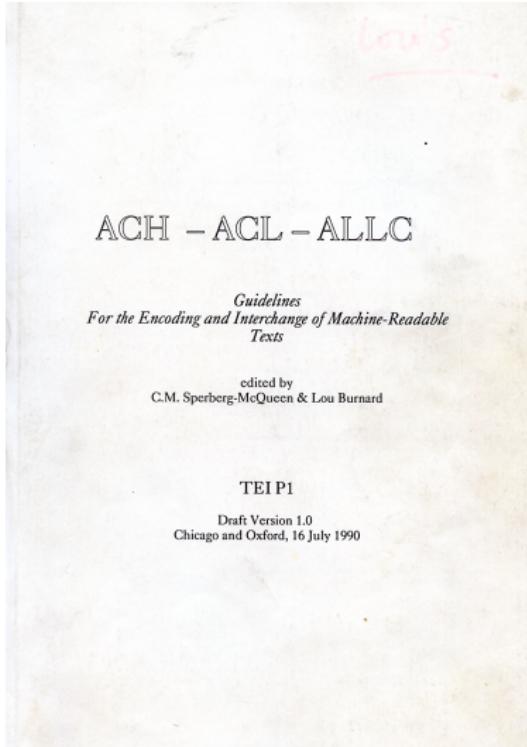
On a très vite compris qu'il y avait beaucoup de chevauchements parmi ces travaux. Les deux TEI Editors essayaient de participer aux débats de chaque comité, et d'appliquer, aussi rigoureusement que possible, le célèbre **rasoir d'Ockham**.

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'
- ...

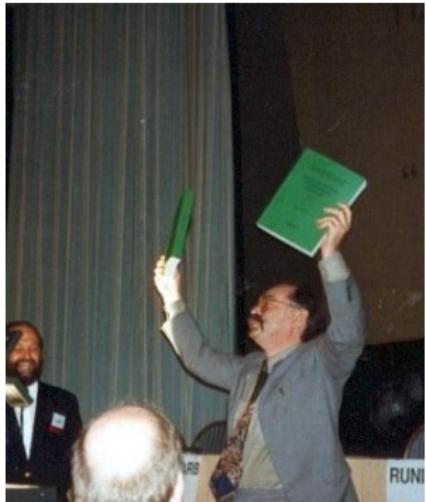
(Encore une raison d'éviter l'usage de TEI All)

# De P1 à P3



- P1 est distribué en c. mille exemplaires imprimés en 1991-1992.
- Avec des financements supplémentaire aux États-Unis, en Europe, et au Canada, une deuxième phase de développement et de validation est entretenue
- Les commentaires et les propositions qui en ressortent sont considérés ensemble à un premier 'TEI Technical Review Meeting' en mai 1993.

## 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



## Révisions majeures prévues pour TEI P5

- élaboration et intégration du système ODD pour la documentation des systèmes de balisage
- description détaillée de manuscrits
- nouvel élément `<choice>` (suite à la 'Guerre des Attributs')
- nouveaux éléments pour la prosopographie
- nouveaux éléments pour représenter et documenter les caractères et les glyphes non-Unicode
- révisions majeures du système de classification des éléments

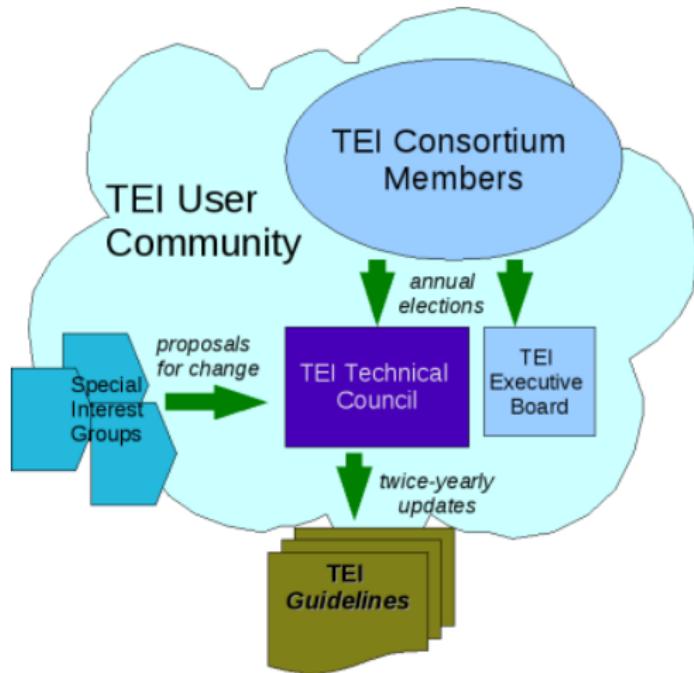
## Et en plus ...

Des grands efforts communautaire sur...

- les entités nommées
- l'intégration des 'facsimilés numériques'
- la représentation physique des documents pour l'édition génétique
- la représentation des analyses linguistiques déportées en ('standoff markup')

Évolutions générées et gérées par les communautés scientifiques intéressées

# 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 !



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

# 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

# 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 noms

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)

## 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 !



## Pour en savoir plus

- <http://www.tei-c.org>
- <http://tei.sf.net>
- [http://listserv.brown.edu/archives/cgi-bin/  
wa?SUBED1=tei-l&A=1](http://listserv.brown.edu/archives/cgi-bin/wa?SUBED1=tei-l&A=1)
- [tei-fr@cru.fr](mailto:tei-fr@cru.fr)

