Introduction à l'XML-TEI

Modéliser ses données textuelles

Maison des Sciences de l'Homme en Bretagne

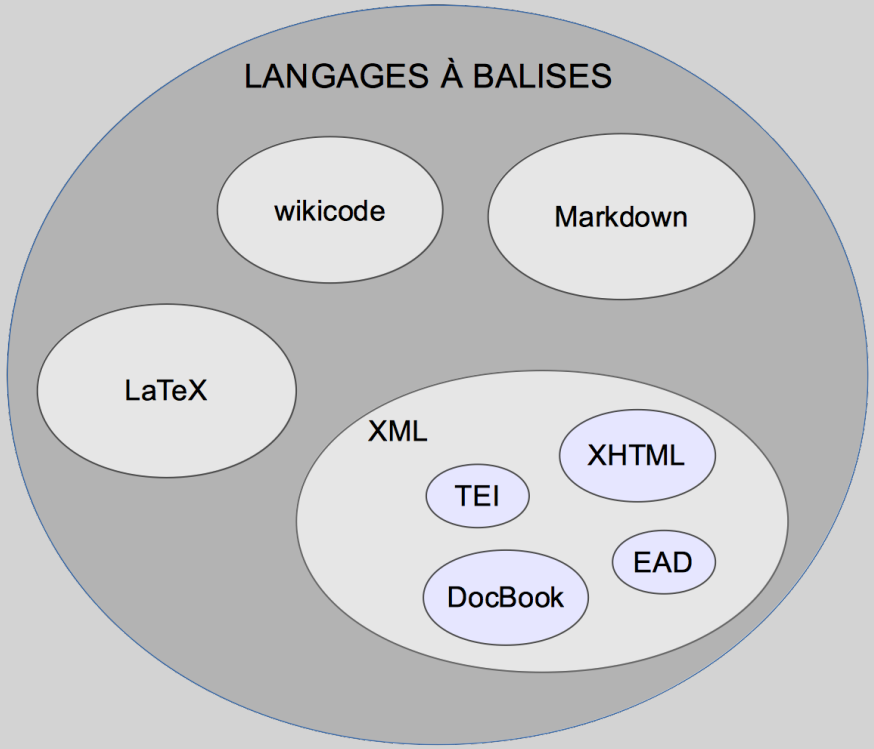
Séance 1 Édition électronique

Master mention Humanité numériques Rennes 2, 30/11/2021

Avant de modéliser, se poser la question : Qu'est-ce qu'un texte ?

- Une notion
  - Un texte se caractérise par différents aspect : une fois identifiés, comment les décrire ? Le format XML (eXtended Markup langage) est particulièrement adapté.

- Une matérialité
  - Un texte brut composé de caractères issus de plusieurs répertoires (ASCII, American Standard Code for Information Interchange - ISO 8859-1, Latin 1 - UTF-8, Universal Character Set Transformation Format pour les principaux). Les formats HTML et LaTeX servent notamment à « styler » le texte.



Avant de modéliser, se poser la question : Qu'est-ce qu'un texte ?

- Distinguer balisage typographique et balisage sémantique

- Intérêt d'un balisage sémantique
  - Exemple d'édition scientifique : The Shelley-Godwin Archive ; Testaments de Poilus
  - Pour aller plus loin : *DH in Practice - Digital Scholarly Editions* par E. Pierazzo ; Why do we encode par E. Pierazzo

Langage	Balise typographique	Balise sémantique
LaTeX	\emph{ad hoc}	\selectlanguage{latin}{ad hoc}
HTML 5	ad hoc	ad hoc
XML-TEI	<hi rend="i">ad hoc</hi>	<foreign xml:lang="la">ad hoc</foreign>

Introduction au format XML

- **Définition** : Format de données pur, langage à balises extensible, conçu pour la description des données textuelles. Son intérêt réside notamment dans la séparation du contenu et de la présentation, permettant d'afficher un même document sur des applications ou des périphériques différents sans pour autant nécessiter de créer autant de versions du document.
- **Un standard international**
  - langage libre et documenté depuis 1998 (spécifications *XML 1.0* ont été reconnues comme recommandations par le W3C)
  - respect des recommandations du **W3C** (World Wide Web Consortium)
  - faciliter la lisibilité par les machines ou par l'œil humain ; l'échange de données ; la migration vers d'autres plates-formes, d'autres logiciels, d'autres formats.

- **Structure générale**

Cf. Fichiers en XML-TEI de l'édition numérique des Testaments de Poilus

Données structurées sous formes de chaînes de caractères délimitées par un balisage les décrivant.

L'unité comprenant les données et le balisage est appelée « élément ».

Exemple : <nomElement>chaîne de caractères</nomElement>

Les éléments XML suivent un principe d'arborescence par imbrication.

Exemple :

```
<elementParent>
    <elementEnfant>chaîne de caractères</elementEnfant>
</elementParent>
```

Les éléments enfants héritent donc des propriétés des éléments parents.

- **Contexte de naissance**

SGML (1970), *Standard Generalized Markup Language*, langage de description à balises qui a donné lieu à … :

- XML, version contrainte de la syntaxe de SGML, afin d'éviter les ambiguïtés dans la structure des données textuelles ;
- HTML (*HyperText Markup Language*), langage de description pour afficher les données sur le web.

- **Éléments structurant un document XML :**

  - Les éléments, ce qui délimite un ensemble cohérent dans le texte selon un tag donnée

    <element>texte</element> ou <elementVide/>

  - Les attributs, ce qui qualifie l'élément

    <MiseEnValeur rendu="rouge italique" position="centrePage">texte </MiseEnValeur>

  - Les commentaires

    <!-- texteCommentaire -->

- Les entités

  &entité;

Les entités sont des appels pour insérer dans le XML des caractères interdits ou bien des séquences de code définies au préalable dans une DTD.

Cf. convertisseur de caractères (pour obtenir le code hexadécimal).

Règles importantes pour que l'encodage soit dit **bien formé** :

- à chaque balise de début doit correspondre une fin de balise (sauf pour les balises auto-fermantes) ;
- les éléments peuvent être imbriqués, mais ils ne doivent pas se recouvrir ;

  Contre-exemple : `<paragraphe> <MiseEnValeur rendu="rouge italique" position="centrePage">texte</paragraphe></MiseEnValeur>`

  Bon exemple : `<paragraphe><MiseEnValeur rendu="rouge italique" position="centrePage">texte</MiseEnValeur></paragraphe>`
- il ne doit y avoir qu'un seul élément racine ;
- un élément ne doit pas avoir deux attributs avec le même nom, mais un attribut peut avoir plusieurs valeurs séparées par des espaces.

- **Instruction de traitement et déclaration XML :**

Les instructions de traitement sont un autre moyen de fournir des informations aux applications auxquelles est destiné le document. Une instruction de traitement commence par "< ?" et se termine par "?>".

Ces dernières sont des balises et pas des éléments. Elles doivent donc être en dehors d'une balise.

Les instructions de traitement les plus courantes sont l'appel d'une feuille de style, d'un schéma et l'appel d'une version de XML. Ces appels doivent être placés avant l'élément racine.

Exemple : <?xml version="1.0" encoding="UTF-8"?>

Présentation de la TEI

- **Pourquoi faire de la TEI ?**
  - Une structuration des données textuelles autour du sens plus que de l'apparence ;
  - Un standard interopérable, indépendant de tout environnement logiciel particulier ;
  - Un langage conçu pour et par la communauté scientifique, qui est aussi en charge de son développement continu.

Cf. Lou Burnard, « Rêve ou cauchemar : comment maîtriser le tigre TEI », http://lb42.github.io/Talks/2021-05-rennes.html.

- **Comment est née la TEI ?**

« La TEI a été d'abord développée, il y a plus de trente ans, comme un projet de recherche dans le champ alors émergent du " *Humanities computing* ". L'idée originelle était de proposer un ensemble de recommandations sur la façon dont les chercheurs devraient créer des ressources textuelles " lisibles par ordinateur ", qui soient adaptées aux besoins de la recherche – dans la mesure où un consensus existait sur le sujet –, mais qui soient également extensibles, puisque ces besoins changent et évoluent. », BURNARD, Lou. *Qu'est-ce que la Text Encoding Initiative ?* Nouvelle édition [en ligne]. Marseille : OpenEdition Press, 2015 (généré le 28 avril 2021). Disponible sur Internet : <http://books.openedition.org/oep/1237>. ISBN : 9782821855816. DOI : https://doi.org/10.4000/books.oep.1237.

- **Quelques dates**
  - 1987 : établissement de la *Text Encoding Initiative*;
  - 1990 : TEI P1 (proposal 1), dir. Michael Sperberg-McQueen et Lou Burnard ;
  - 1992-1993 : TEI P2, expansion ;
  - 1994 : TEI P3, première version complète ;
  - 2000 : naissance du TEI Consortium ;
  - 2001-2004 : TEI P4, introduction du XML ;
  - 2007-... : TEI P5, abandon de SGML.

➔ Un standard en constante évolution.

- **La communauté TEI**

La communauté TEI est animée par le *TEI consortium*, fondation interdisciplinaire à but non lucratif.

Il se compose des unités suivantes:

- TEI Board of Directors ;
- TEI Technical Council ;
- Membres institutionnels et individuels ;
- TEI Workgroups, par exemple :
  - TEI Manuscripts Special Interest Group ;
  - Correspondence SIG ;
- Special Interest Groups.

La communauté peut échanger et se rencontrer grâce à :

- Une liste de diffusion : TEI-L mailing list;
- Une liste francophone : TEI-FR et un wiki;
- Des *members meetings* (congrès annuels) : TEI Conference;
- Une revue : *Journal of the Text Encoding Initiative*;
- Des **Guidelines** ("recommandations") qui documentent notamment chaque élément.

- **TEI Technical Council**

À l'écoute de la communauté scientifique, le *council* est en charge du maintien de la TEI et de son adaptation en fonction des besoins des utilisateurs.

Moyens de communication :

- La mailing liste **TEI-fr**, https://groupes.renater.fr/wiki/tei-fr/index;
- Github TEI où vous pouvez ouvrir des « issues » : https://github.com/TEIC/TEI/issues, soit des discussions autour de problèmes et questions pouvant intéresser la communauté, ou bien trouver de nouveaux outils : https://github.com/TEIC.

Pour aller plus loin : https://docs.google.com/presentation/d/16cVewiMmMI7LcA4tqaWVSC-XLgelwo7TE_CaMiztEz0/edit#slide=id.p

- **Mode d'emploi de la TEI**

TEI est un set de balises prédéfini et documenté dans les TEIguidelines qui permet de procéder à une description « scientifique » et « sémantique » d'un texte.

Pour utiliser un set de balises TEI, il faut déclarer le nom de domaine TEI (*name space*) dans l'élément racine du document XML grâce à ce qu'on appelle une adresse URI qui pointe vers une description du set de données. C'est ce que l'on appelle la déclaration de nom de domaine.

Exemple : <TEI xmlns="http://www.tei-c.org/ns/1.0">

TEI (All) n'est pas un schéma à proprement parler, mais plutôt un framework, utile à la conception de son propre schéma. Il est fortement déconseillé d'utiliser un schéma englobant l'intégralité de la TEI. **La conception d'un modèle adapté à ses données et à son projet est extrêmement importante.**

Structuration générale d’un fichier TEI

Structuration générale d’un fichier TEI

Tout document TEI a au moins deux parties :

- Un en-tête, représenté au moyen d’un élément **<teiHeader>** contenant des métadonnées décrivant le document ;
- le texte lui-même, représenté par un élément **<text>**.

- Le **teiHeader** minimal comporte les trois sections suivantes au sein de l’élément **<fileDesc>** :
  - **<titleStmt>** : informations identifiant le document lui-même ;
  - **<publicationStmt>** : informations sur la façon dont il est distribué ou publié ;
  - **<sourceDesc>** : indications sur ses origines.

Structuration générale d'un fichier TEI

- L'élément **<text>** contient les trois parties suivantes :
  - **<front>** : pour les préfaces, et tous les éléments liminaires du texte ;
  - **<body>** : pour le corps du texte proprement dit ;
  - **<back>** : pour tous les appendices, épilogues, postfaces, etc.

Pour réviser, cf. « Why Do We Standardize? », par Elena Pierazzo.

- **Le <teiHeader> :**

L’en-tête TEI peut posséder quatre composants principaux :

- **<fileDesc>** : description bibliographique du document **(obligatoire)** ;
- **<encodingDesc>** : description de l’encodage ;
- **<profileDesc>** : description détaillée des aspects non bibliographiques ;
- **<revisionDesc>** : résumé de l’historique des révisions pour un fichier.

NB : Il est conseillé de normaliser les valeurs d’attributs en se calant sur des standards internationaux. Par exemple, si je déclare une langue dans le profileDesc, j'utilise la norme ISO 639 pour renseigner la valeur
de @xml:lang: https://fr.wikipedia.org/wiki/ISO_639.

Le **teiHeader**, regroupant les **métadonnées** du texte encodé, est extrêmement complexe, car il est conçu pour s’adapter à des usages scientifiques très divers.

Le header doit garantir la **qualité scientifique** du document, permettre l’échange des données et leur conservation. On peut donc, la plupart du temps, chercher à reproduire des modèles qui ont déjà fait leurs preuves comme le *Dublin Core*.

Certaines informations du ***Dublin Core*** ne sont pas comprises en TEI dans le <fileDesc>, comme la langue de l’œuvre ou sa date de création. Ces informations primordiales sont à ajouter dans le **<profileDesc>**.

Les Teiguidelines

Guide des pratiques d'encodage à consulter au moindre doute !

- **Les "principes de Poughkeepsie" (1987)**

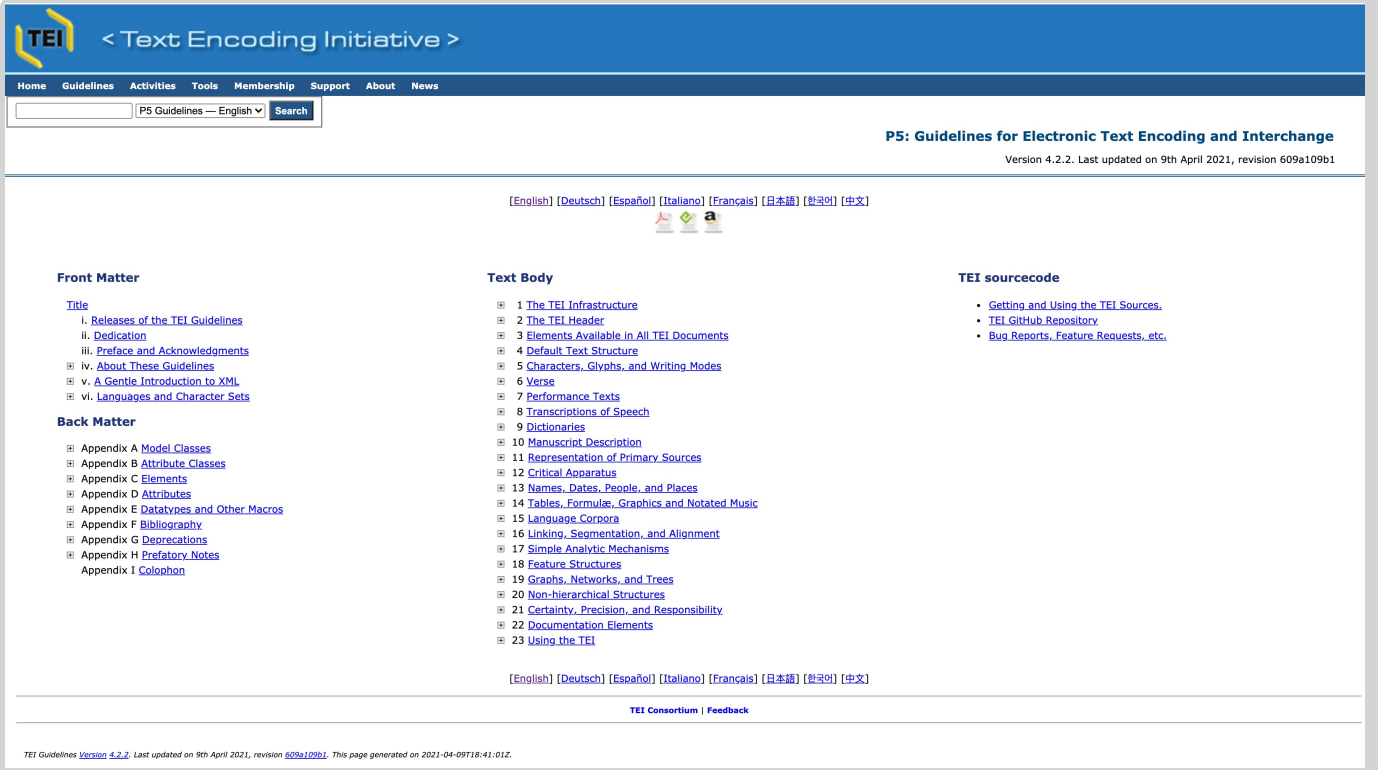
Les recommandations visent à :

- Fournir un format standard ;
- Favoriser l'échange de textes dans les humanités ;
- Suggérer des principes abstraits pour l'encodage des textes ;
- Inclure un ensemble minimal de conventions pour l'encodage de nouveaux textes ;
- Proposer des ensembles de conventions d'encodage adaptés à plusieurs applications différentes.

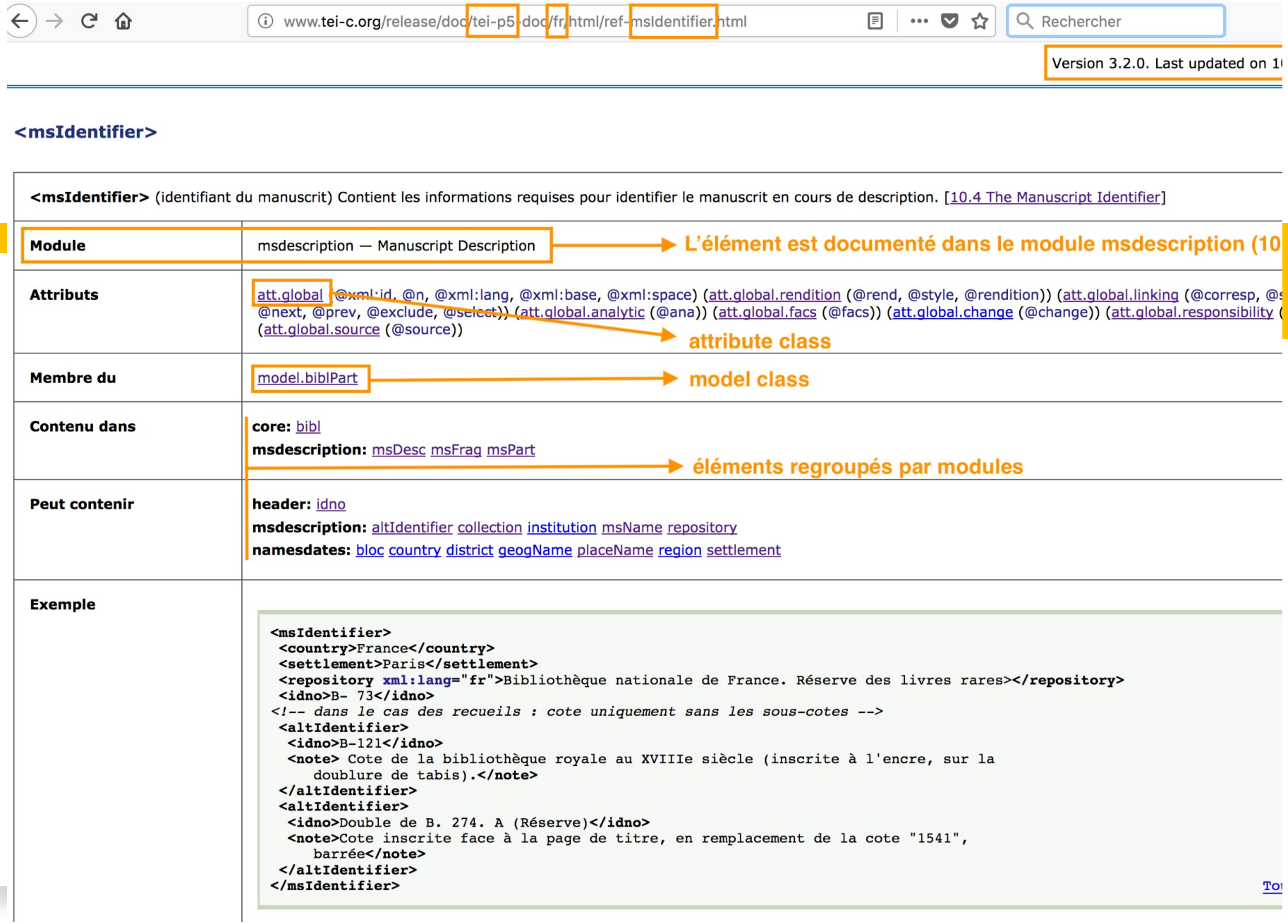
Les Teiguidelines

- Comment lire les guidelines ?

Un sommaire avec les différents types de texte que l'on pourrait encoder.



Les Teiguidelines



- **Les modules**

Chaque module est documenté par un chapitre des Guidelines.

4 modules sont obligatoires (communs à tous les documents TEI) :`

- tei : 1 The TEI Infrastructure (définition des modules, des classes et des macros, soit les modèles de contenu et les types de données) ;
- header : 2 The TEI Header (métadonnées communes) ;
- core : 3 Elements Available in All TEI Documents (paragraphe, ponctuation, citations, …) ;
- textstructure : 4 Default Text Structure (éléments de base pour structurer un texte de type livre).

Les Teiguidelines

Les modules sont relatifs à un type d'objet, une approche, une discipline, par ex. :

- analysis : analyse linguistique
- drama : textes d'art dramatique
- gaiji : caractères non standard et glyphes
- linking : liens, segmentation, alignements
- msdescription : description des manuscrits
- namesdates : noms, dates, lieux
- textcrit : apparat critique
- transcr : transcription des sources primaires
- …

Les Teiguidelines

- **Les classes d'attributs**

La classe ***att.global*** fournit un jeu d'attributs communs à tous les éléments dans le système de codage TEI.

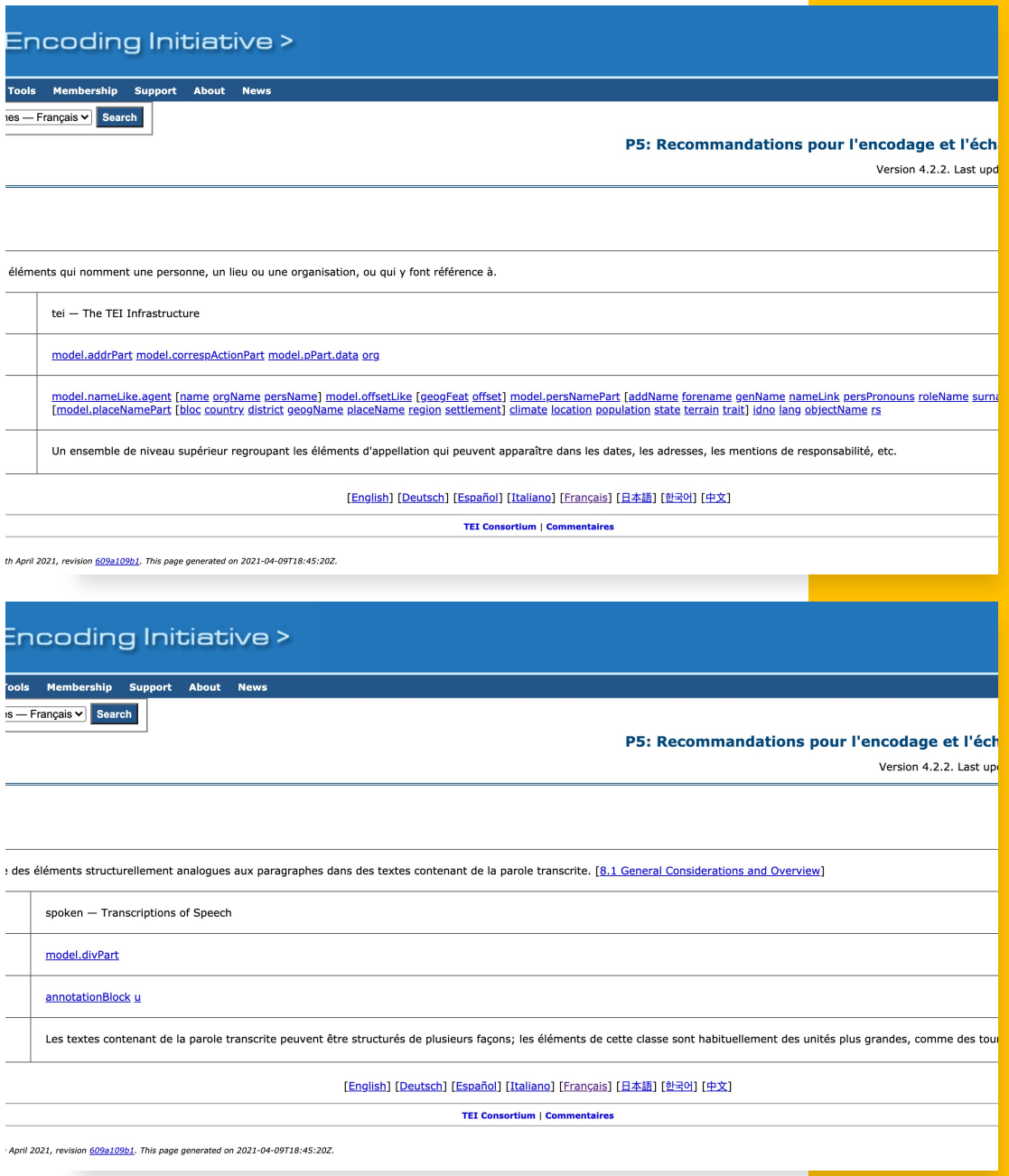
- **xml:id** (identifiant) fournit un identifiant unique pour l'élément qu'il porte.
- **n** (nombre) donne un nombre (ou une autre étiquette) pour un élément, qui n'est pas nécessairement unique dans le document TEI.
- **xml:lang** (langue) indique la langue du contenu de l'élément en utilisant les codes du RFC 3066 **rend** [att.global.rendition]
- **rendition** [att.global.rendition] pointe vers une description du rendu ou de la présentation utilisée pour cet élément dans le texte source.
- **xml:space** signale que les applications doivent préserver l'espace blanc.
- **source** [att.global.source] spécifie la source du document.
- **cert** [att.global.responsibility](certitude) donne le degré de certitude associée à l'intervention ou à l'interprétation.
- **resp** [att.global.responsibility] donne l'identité de la personne à l'origine de l'élément encodé.

Les Teiguidelines

- **Les modèles de classe**

  - Pour les sections d'un modèle : model.*divPart.??*

  - Pour regrouper les éléments d'un même domaine : *model.??Like*



- **Les macros**
  - **macro.limitedContent** (contenu du paragraphe) définit le contenu des éléments textuels qui ne sont pas utilisés pour la transcription des contenus existants.
  - **macro.phraseSeq** (suite de syntagmes) définit un ordre de données et d'éléments syntagmatiques.
  - **macro.specialPara** (contenu "spécial" de paragraphe) définit le modèle de contenu des éléments tels que des notes ou des items de liste.
  - **macro.xtext** (texte étendu) définit une suite de caractères et d'éléments gaiji.