

Les possibilités de la TEI P5 pour les sources historiques : l'exemple d'un recueil de chartes

Introduction

En guise d'introduction et pour poser les bases de mon intervention, je voudrais rappeler quelques généralités sur le format XML et le travail d'encodage. Ce format est en passe de s'imposer comme le format standard d'encodage des données dans les applications informatiques.

Mis au point au sein du W3C, sur la base du SGML, il offre de nombreux avantages :

- Format normalisé, ouvert et libre ;
- Indépendance par rapport aux logiciels et plates-formes ;
- Il n'impose pas l'utilisation d'un logiciel particulier pour le lire ;
- Séparation des données relatives à la mise en forme ou l'exploitation de l'information et les données en elles-même ;
- Souplesse d'utilisation ;
- Support par de plus en plus d'applications ;
- Perspective d'interopérabilité.

Le XML a pour but de décrire et structurer l'information selon son organisation logique, hiérarchique et intellectuelle au moyen d'une syntaxe particulière. Il n'indique donc pas le sens d'un texte, mais sa structure logique dans le contexte précis du document. Chaque type d'informations, comme par exemple un paragraphe, un titre de partie, une emphase, un lien... est décrit au moyen d'une balise qui encadre la portion d'information concernée.

La norme définissant le XML laisse libre tout à chacun de décider des noms des balises et de leurs règles d'emboîtement. Pour autant, pour offrir des possibilités d'interopérabilité, il existe un système de grammaire ou de schéma qui décrit le nom des éléments, leurs règles d'utilisation et d'agencement. Ces grammaires peuvent être elle-même exprimées selon plusieurs syntaxes : DTD, XML schema ou Relax NG. Ainsi, il existe une multitude de schémas qui répondent à des buts différents :

- EAD, encoding Archival Description, pour structurer des inventaires d'archives
- XHTML, eXtensible Hypertext Markup language pour structurer les pages Web
- MODS, pour décrire des notices bibliographiques
-

Malgré tout, l'utilisation de ces grammaires ne rend pas automatique le travail d'encodage, car elles ne répondent bien souvent pas à des besoins très précis, mais plutôt au besoin d'encodage d'un type plus ou moins vaste de documents. Le choix des éléments utilisés, la façon de les utiliser et leurs applications constituent donc des actes réfléchis et répondent à une perspective et une problématique précise, scientifique dans le cas des chercheurs. Cette constatation est d'autant plus vraie dans le cas

de l'encodage de sources historiques dont on connaît l'hétérogénéité de présentation et de structure ; on pourrait d'ailleurs dire qu'à un certain niveau de détail, chaque source présente une structure unique. Et si vous ajoutez à ce constat le fait que chaque chercheur veut étudier une source avec un point de vue particulier, la transcrire et/ou l'éditer avec des règles plus ou moins précises et standardisées et l'analyser selon sa propre problématique, l'encodage des sources historiques revient finalement à un travail personnel presque impossible à normaliser, car chaque encodeur développera sa propre stratégie d'encodage par rapport à sa propre perspective pour chaque type de sources voire chaque source.

Malgré ce tableau qui pourrait effrayer, nous sommes tous parfaitement conscients de l'intérêt d'échanger les corpus de sources mis au point dans nos différentes institutions. C'est pourquoi, depuis 1987, au sein de la Text encoding Initiative, la TEI, un groupe composé de chercheurs en sciences humaines, en linguistique, en informatique, de bibliothécaires et plus généralement de professionnels de l'information et de la documentation travaille à l'élaboration d'un guide de balisage qui comprend à peu près 450 éléments et qui vise à l'établissement d'un format pivot d'échange. Malgré les efforts de ce groupe, il est impossible d'être exhaustif. Comme je l'ai déjà dit, à chaque communauté voire à chaque individu au sein même d'une communauté correspond des besoins précis dont l'introduction gonflerait abusivement le guide. C'est pourquoi a été mis au point un format : ODD qui ouvre de nouvelles perspectives en matière d'élaboration de règles d'encodages pour des besoins précis et en particulier la possibilité de mettre en place des modèles conceptuels, servant de référentiels.

Avant de continuer ma communication, je précise que je me place dans la perspective de l'édition critique et scientifique de sources. Mon propos ne recouvre donc pas la seule description des sources, à des fins archivistiques, en particulier, prise en charge par l'EAD.

I- Présentation de la TEI et de ODD

La TEI, organisée depuis 2000 sous la forme d'un consortium, travaille, en priorité, à l'élaboration d'un guide comprenant la description de l'ensemble des éléments utiles pour l'encodage d'un texte en vue de son échange, d'où le nom du guide : « *Guidelines for Electronic Text Encoding and Interchange* ». Les éléments sont classés en différents modules selon leurs buts. 22 modules sont ainsi disponibles rassemblant à peu près 450 éléments. Parmi ces modules, outre ceux contenant les éléments communs à tous les documents encodés en TEI (*core*, *header* et *tei*) et les modules de base répondant aux grands types de documents encodés (par exemple, *prose* pour les documents en prose, *verse* pour les documents en vers, *drama* pour les pièces de théâtre et les scénarios), des modules additionnels répondent à des besoins particuliers comme *textcrit* pour les éléments spécifiques à l'édition critique de textes ou *transcr* pour les éléments spécifiques à la transcription de sources historiques. En choisissant ces différents modules en fonction de ces besoins, un utilisateur crée sa propre grammaire qui suit les prescriptions du guide. Il peut alors générer le fichier qui va lui permettre de valider ses fichiers XML par rapport à cette grammaire dans la syntaxe qui lui convient le mieux : DTD, XML schema ou Relax NG.

Pourtant, comme tout travail de ce type, les éléments proposés par le guide de la TEI est le fruit d'un compromis entre les différentes communautés et les différents participants. C'est pourquoi il peut paraître insuffisant voire aux yeux de certains « dictatorial », ne laissant que peu de marges aux spécificités d'un type de sources ou de la problématique d'un chercheur. Il existe bien des éléments génériques qu'il est possible de préciser avec l'attribut « type », comme <div>, pour une division d'information, <rs> pour une chaîne de caractères référencée, <seg>, pour une portion à l'intérieur d'une phrase ou encore <name> pour un nom propre ou <persName> pour un nom de personne. Pour autant, dans une perspective d'interopérabilité, il sera essentiel de normaliser les valeurs possibles de l'attribut « type » et ces solutions ne sont pas forcément satisfaisantes aux yeux de tous les utilisateurs. De plus, les possibilités ouvertes par la TEI peuvent avoir pour conséquence que pour une même source encodée dans une perspective à peu près équivalente, l'encodage pourra s'avérer différent. Enfin, l'appropriation de la TEI s'avère fastidieuse et avant de pouvoir réellement manipuler la TEI et en utiliser tout le potentiel, un important travail est nécessaire en amont.

Or, le but du consortium n'est absolument pas de mettre au point une norme qui serait suivie par tous, mais plutôt d'offrir des outils facilitant l'encodage des textes en vue de leur échange. Conscient de ces problèmes, le consortium, et plus précisément le TEI council en charge des questions techniques en son sein, a mis en place dans la version actuellement en développement, dite P5 pour Proposal 5, un nouveau format : ODD, One Document Does it all et a ajouté un deuxième système pour organiser les éléments et les attributs : les classes qui correspondent aux comportements de l'élément dans l'arbre XML, c'est à dire aux règles d'emboîtement que suit l'élément.

ODD permet de documenter précisément sous la forme d'un fichier XML la grammaire d'un utilisateur en prenant appui sur le guide de la TEI. Au niveau le plus simple de l'utilisation de ce format, il permet d'indiquer les modules choisis par l'utilisateur, ce qui correspond à l'essence même de la démarche de la TEI décrite précédemment. Mais, il offre d'autres perspectives qui dépassent ce cadre strict :

- changer le nom d'un élément ;
- changer le nom d'un attribut ;
- changer les attributs d'un élément ;
- restreindre le schéma aux éléments et aux attributs utilisés et pas simplement aux modules ;
- changer le comportement d'un élément et d'un attribut, c'est à dire le changer de classes ou le classer dans une ou plusieurs autres classes ;
- ajouter un élément, en lui définissant un comportement, une ou plusieurs classe(s) d'élément à laquelle il appartient, un type de données, une classe d'attribut voire en le rattachant à un élément précis ;
- ajouter un attribut, en lui définissant un comportement, un type de données et en le rattachant à une ou plusieurs classes :
- restreindre la valeur d'un attribut ou d'un élément à une liste.

ODD permet donc de décrire et documenter tous les changements par rapport au guide de la TEI et assure ainsi l'interopérabilité avec les autres projets. Concrètement, le TEI consortium a mis en place une interface en ligne, Roma, qui permet de générer très facilement un fichier ODD, la DTD, le XML schema ou le Relax NG qui correspond et la documentation des éléments choisis, et éventuellement spécifiques aux projets, puisqu'il est possible de documenter chacun des éléments.

DESCRIPTION DE ROMA --- cf le power point et les copies d'écran effectuées.

Puisque ODD est un fichier XML, il est aussi possible de l'écrire directement avec un éditeur XML ce qui offre d'ailleurs un peu plus de possibilités que l'interface qui n'implémente pas encore toutes les fonctionnalités. Il suffit ensuite de charger dans l'interface en ligne le fichier XML pour générer le schéma qu'il vous convient.

Grâce à ODD, il est ainsi possible de pallier aux manques et aux défauts de souplesse qui peuvent apparaître au premier abord. L'interopérabilité est assurée à plusieurs niveaux :

- Au niveau d'une communauté précise, il est possible de s'accorder sur un fichier ODD et ainsi d'utiliser un schéma très précis et donc plus facile à utiliser et à partager. Cette solution est d'ailleurs celle choisie par les épigraphistes, au sein du projet Epidoc. Cela pourrait aussi être le cas pour l'encodage des chartes dans le cadre de leur édition critique, par exemple.
- Grâce au fichier ODD, il est très facile de repasser à une TEI « canonique » puisque tous les changements ont été documentés, ODD prévoit d'ailleurs de lier le fichier avec une feuille de style XSL qui décrirait les transformations pour retrouver un fichier XML suivant

scrupuleusement le guide.

- Même, sans passer par le fichier ODD, il s'avère que les indications de macro-structure sont identiques. Dans ce cas, l'interopérabilité n'est pas assuré à 100%, mais pour autant certaines passerelles sont rendus possibles par l'utilisation de la TEI.

Pour autant, la mise en place d'un fichier ODD demande deux préalables :

- une bonne connaissance de la TEI, son fonctionnement, les différents éléments et son organisation ;
- une réflexion attentive de tous les éléments qui vont être utiles pour encoder le document, en l'occurrence la source.

En partant d'un travail qui pouvait sembler technique à la base, force est de constater que la démarche scientifique reste essentielle. Or, on s'aperçoit rapidement que cette réflexion amène bien souvent à faire évoluer notre propre perception du document à encoder, comme le montrera demain Paul Bertrand à propos de la description des Cartulaires. Il faut garder aussi en tête que l'encodage de l'information par ces moyens ne peut accepter les exceptions et les approximations et que certaines pratiques qui n'étaient pas strictement normalisées traditionnellement, comme la rédaction des notes d'apparat critique, doivent absolument l'être. Dans ce cas, une réflexion sur les pratiques est aussi indispensable.

Une fois ces réflexions abouties, la mise en place du fichier ODD pourra servir de référentiels pour une communauté qui cherche à encoder un type précis de documents. En guise d'exemples et pour revenir à nos préoccupations, je vous propose donc de vous faire part du début d'une réflexion modeste sur un modèle conceptuel pour les chartes médiévales dans le cadre de l'édition critique d'un chartier, qui, je l'espère, pourrait aider les travaux pour la mise en place du schéma de la CEI.

II- Un modèle conceptuel pour les chartes médiévales

Cette réflexion s'appuie sur le travail engagé au sein de l'École nationale des chartes, depuis 4 ans, et depuis peu, au sein du centre de ressources numériques TELMA qui a abouti à la mise en ligne de plusieurs éditions critiques de chartes originales ou numérisées encodées selon le guide de la TEI.

Il est possible d'aborder une charte suivant différentes problématiques de recherche : paléographique, diplomatique, historique ou linguistique et philologique. A chacune d'entre elles correspondent des couches d'annotations associées à des descripteurs spécifiques. Néanmoins, ces couches d'annotations se recoupent en partie selon deux modalités différentes :

- D'une part, les différents niveaux de profondeur d'appréhension du document qui correspondent à trois réalités :
 - **L'événement décrit dans la charte ;**
 - **Le texte** en tant que médium de l'événement décrit dans la charte, mis au point à partir de l'original s'il est toujours conservé ou de la copie qui semble la plus pertinente au chercheur dans le cas contraire ;
 - **Les différents témoins du texte**, c'est à dire les caractéristiques matérielles et graphiques du ou des supports du texte.
- D'autre part, les règles de l'édition critique scientifique qui entremêle bien souvent ces différentes réalités et ces différentes problématiques de recherche, sans que les chercheurs en aient réellement conscience, pris dans leurs habitudes de travail.

Il est donc assez complexe de classer tous les descripteurs pouvant être utiles à l'encodage des chartes et d'en définir la nature. Il serait possible d'avoir un ensemble commun relevant de l'exercice de l'édition critique proprement dite et de ses habitudes, et différents modules rassemblant des descripteurs très spécialisés. Cela reviendrait à classer les descripteurs en fonction de leurs contenus et de leurs usages avec toutes les difficultés que cela représente.

Une autre possibilité plus simple consiste à s'appuyer sur la structuration de l'encodage pour classer les descripteurs en deux groupes principaux :

- **Les informations placées à l'extérieur du texte de la source** permettent d'identifier et de contextualiser le document par rapport à une période historique, un corpus, une problématique de recherche et les recherches antérieures. On pourrait, en partie, rassembler ces informations sous le vocable de métadonnées.
- **Les informations placées à l'intérieur du texte de la source** permettent de structurer, décrire ou annoter le texte ou le support d'un témoin du texte.

C'est cette approche que je me propose d'étudier.

Informations à l'extérieur du texte de la charte

Ces informations sont placées à l'extérieur de la charte et permettent l'identification et la contextualisation de l'événement, du texte et des témoins du texte.

Contexte documentaire

Le contexte documentaire est de deux types :

- **La place de la charte dans l'ensemble documentaire édité** que ce soit une compilation médiévale dans le cas d'un cartulaire ou contemporain dans le cas d'un chartrier. Ce contexte sous-entend le codage de la charte par rapport à l'ensemble documentaire, mais aussi les moyens d'identifier la charte dans l'ensemble documentaire : identifiant informatique et titre de la charte.
- **Le tableau de la tradition** listant et décrivant l'ensemble des témoins du texte : original, copie manuscrite ou imprimée. Il contient les informations suivantes :
 - copie ou original
 - Identification géographique du témoin - référence bibliographique ou archivistique
 - référence de l'institution conservant le manuscrit
 - référence cote du manuscrit
 - citation de la rubrique
 - citation du texte
 - référence au folio qui contient le document
 - description physique du témoin
 - identifiant du témoin. Dans le cas des actes, il faut suivre la typologie ; dans le cas des manuscrits littéraires, il faut utiliser l'identifiant canonique (le plus souvent une lettre attribuée au manuscrit)
 - mesure du témoin
 - support physique du témoin
 - particularités physique du témoin (dessins, *signa*, sceau, chirographe)
 - présence d'une rubrique
 - la référence au témoin dont dépend celui décrit
 - deux types de références : « d'après »/ « avec référence à »

Contexte historique

Le contexte historique rassemble toutes les informations relatives à l'événement ou à l'élaboration d'un des témoins du texte :

- Les différents acteurs de la charte sous une forme normalisée : auteur de l'action juridique,

bénéficiaire de l'action juridique, auteur de la charte s'il est différent de celui de l'action juridique ;

- Date de l'événement, au besoin restituée (accompagné d'une date normalisée selon la forme YYYY-MM-DD) et, éventuellement si différente, date de rédaction du témoin édité, si elle est connue ;
- Résumé détaillé de l'événement appelé aussi regeste dans l'édition scientifique traditionnelle de charte ;
- Une note critique constitue un texte de un ou plusieurs paragraphes abordant différents sujets :
 - discussion sur la datation de l'événement
 - Analyse d'une (ou plusieurs) particularité d'un témoin (cas du chirographe par exemple) ;
 - Analyse de la véracité de la charte (l'analyse du faux, même si elle n'est pas aussi importante qu'avant, reste une spécialité de la diplomatique)
 - Analyse du contenu relaté dans la charte.

Parmi les informations qui ne sont pas spécifiques à l'édition critique, on pourrait aussi placer dans cette partie une indexation matière de la charte ou un classement de la charte selon des typologies relevant de la ou des problématiques des chercheurs.

Contexte bibliographique

Le contexte bibliographique permet de replacer l'édition de la charte par rapport aux recherches ou mentions de la chartes dans des études ou inventaires antérieurs. Cette information correspond à la rubrique traditionnellement dénommée « Indiqué » dans l'édition critique de chartes et placée juste après le tableau de la tradition.

Contexte linguistique

La langue utilisée dans le document doit pouvoir être indiqué de façon normalisée.

Informations à l'intérieur du texte de la charte

Outre les descripteurs permettant de refléter la structure logique du texte (paragraphe ou vers), les informations à l'intérieur du texte peuvent être classés en trois groupes :

- Les informations relatives aux caractères externes, c'est à dire les informations sur le ou les support de la charte ;
- Les descripteurs portés sur le texte concernant les informations relatives aux caractères internes, c'est à dire le texte en lui-même ;
- Les notes de l'éditeur/encodeur/chercheur portées sur le texte.

Les caractères externes

Les informations relatives à l'analyse des caractères externes s'apparente en partie à la critique génétique et qu'on pourrait aussi appeler l'édition imitative. Au passage, Il est intéressant de noter que l'encodage permet de mettre fin au débat entre les tenants de l'édition critique et de l'édition imitative puisqu'ils ne s'excluent plus, mais se complètent. Ces informations recourent des problématiques linguistiques, comme l'étude de la ponctuation originale, des problématiques paléographiques comme l'étude des abréviations ou diplomatiques comme l'étude des *signa* et autres *chrismon*. Les descripteurs propres à l'étude des caractères externes sont les suivants :

- Une figure (Lettrines, enluminures, signum, croix, chrismon, ruche, note tironnienne...);
- Abréviation pour lequel il est utile d'encoder le mot abrégé, les lettres restituées et éventuellement le type d'abréviation.
- Saut de ligne
- Saut de colonne
- Saut de page
- Ponctuation originale
- Ratures/biffures
- Changement de mains
- Changement d'encre
- Changement d'ordre typographique des lettres (lettres allongées, lettres décorées)
- Trou dans le parchemin (avec éventuellement la restauration. cf au-dessus)
- Ajout postérieur
- Note portant sur un détail de mise en page d'un des témoins du texte (présentation en colonnes par exemple)

Les caractères internes

Les descriptifs relatifs aux caractères internes relève de différentes couches d'annotations. Elles sont

donc de différentes natures :

- Descripteurs sur le texte en tant que médium de l'événement ;
- Ajout de l'éditeur scientifique pour permettre une lecture fluide du texte ;
- Descripteurs repérant des informations utiles à l'analyse historique.

Descripteurs sur le texte :

- **Les différentes lectures/variantes d'un passage** selon les témoins du document ou une correction introduite par l'éditeur selon la méthode dite « parallel segmentation method » dans le guide de la TEI, c'est à dire le codage du passage choisi par l'éditeur et les différentes lectures des autres témoins semblant pertinents à l'éditeur scientifique voire relevé de manière exhaustive comme le préconise Peter Robinson.
- Bourde identifiée avec sa correction

Ajout de l'éditeur scientifique

- Lacune comblée par d'autres sources
- Mot ou passage restaurés par conjectures
- Passage ne pouvant être transcrits par l'éditeur scientifique.
- Informations d'ordre linguistique

Analyse historique sur le texte

- Identifications des personnes
- Identifications des lieux
- Encodage d'informations relevant de l'analyse historique. Index rerum. Termes à indexer selon une typologie
- Les parties du discours diplomatique

Les notes portées sur le texte

Il est possible d'associer un point du texte à une annotation qui peut être de deux types :

- Référence bibliographique ou note d'ordre historique
- Identifications des citations