
Le codage informatique des appareils critiques : évaluation des recommandations de la *Text Encoding Initiative*

François ROLE

DISTNB

Ministère de l'Éducation nationale, de l'Enseignement supérieur et de la Recherche

1, rue Descartes, 75005 Paris

email: role@distb.mesr.fr

Résumé. Après une rapide définition de la notion d'apparat critique et une présentation détaillée des recommandations de la TEI en la matière, nous examinons les extensions qui pourraient être proposées à la TEI pour mieux prendre en compte les caractéristiques particulières de certains types de textes (manuscrits modernes, textes présentant un intérêt du point de vue de leur aspect physique).

1. Introduction

La TEI *Text Encoding Initiative* [7] est un projet de recherche international qui a débuté en 1988 et a été soutenu par de nombreuses institutions scientifiques (ACL, ACH, ALLC, etc.).

L'objectif de ce projet est de définir un format d'échange, basé sur la norme SGML [6, 9], pour faciliter la circulation des textes électroniques au sein de la communauté scientifique et en particulier dans celle des sciences humaines.

Au delà du processus d'échange *stricto sensu* les éléments de données définis par la TEI peuvent également se prêter au codage d'analyses et de traitements effectués sur les textes. Les experts de la TEI se sont notamment penchés sur les problèmes posés par l'édition critique.

Avant d'aborder ce point particulier, nous commençons par rappeler la structure des recommandations publiées dans le cadre du projet TEI. Ce document, intitulé *Guidelines for Electronic Text Encoding and Interchange* [8] est en effet un texte complexe dont il est utile de comprendre l'organisation d'ensemble.

2. Structure des recommandations de la TEI

Les recommandations de la TEI ont une structure très modulaire. les éléments SGML et les attributs correspondants sont regroupés dans des fragments de DTD appelés *tag sets*. Ces fragments sont invoqués selon les besoins par la DTD TEI proprement dite, la *main* DTD¹.

On distingue trois types de *tag set* :

core tag sets

Ce sont des fragments de DTD inclus obligatoirement dans la DTD principale. Ils sont définis dans la partie II.

base tag sets

Ce sont des ensembles d'éléments et d'attributs correspondant aux principaux types de texte traités par la TEI : prose, poésie, théâtre, dictionnaires, etc. À chaque *base tag set* correspond un chapitre de la partie III.

additional tag sets

regroupent les éléments nécessaires au codage d'analyses et de traitements effectués sur les textes. Les *additional tag sets* sont définis dans la partie IV. L'un d'entre eux, nommé *TEI.textcrit* est spécifiquement dédié à l'édition critique.

```
<!DOCTYPE TEI.2 system 'tei2.dtd' [  
<-- Spécification d'un base tag set à utiliser -->  
  
<!ENTITY % TEI.prose 'INCLUDE'>  
<-- Spécification d'un additional tag set à utiliser:  
ici celui de l'édition critique -->  
<!ENTITY % TEI.textcrit 'INCLUDE'>  
>
```

FIGURE 1 – Exemple d'invocation de tag sets

1. Cette DTD est dite *main* car à côté de la DTD principale, la TEI a également défini quelques DTD dites auxiliaires, par exemple la DTD dite *independent header*. Ces DTD auxiliaires sont décrites dans la partie V de la TEI.

3. Analyse des recommandations de la TEI pour le codage des apparats critiques

La section de la TEI dédiée à l'édition critique aborde principalement deux questions : le codage de l'apparat critique et la façon d'associer cet appareil au texte.

3.1. Notion d'apparat critique

Les textes littéraires antérieurs à l'apparition de l'imprimerie nous sont connus au travers de copies manuscrites comportant souvent des divergences.

Lorsqu'il édite un texte de ce genre, le philologue doit tenir compte de ces différences (ces «variantes»). Il doit cependant également présenter au lecteur un texte cohérent. Pour ce faire, il choisit celles des variantes qui lui paraissent le plus plausibles et renvoie les autres en notes avec l'indication des copies manuscrites dans lesquelles on les trouve. Ces copies manuscrites sont en général identifiées par un symbole (un «sigle», qui est souvent une lettre majuscule). L'éditeur ajoute également souvent des notes et des commentaires pour permettre une meilleure compréhension.

L'ensemble constitué par les indications de variantes et les éventuelles notes associées constitue ce que l'on appelle «l'apparat critique» du texte².

Nous donnons, figure 2, un exemple simple d'apparat critique. Cet échantillon met en évidence un certain nombre de variantes signalées par des lettres minuscules en exposant. Au moyen de ces renvois, l'éditeur critique signale en note des variantes apparaissant dans deux manuscrits désignés par les lettres majuscules P et G.

3.2. Codage de l'apparat critique

Les informations relatives aux différentes variantes signalées dans l'édition critique sont regroupées dans l'élément `<app>` pour *apparatus entry*.

3.2.1. Modèle de contenu de l'élément `<app>`

Chaque entrée de l'apparat comporte un ensemble d'au moins un élément `<rdg>` (*reading*). La suite d'éléments `<rdg>` peut être éventuellement précédée

2. Contrairement à la majorité des systèmes de traitement de texte commerciaux, il faut donc pouvoir disposer de deux systèmes de notes de bas de page. Ceci est possible avec certains éditeurs structurés comme Grif/Thot [1] et bien sûr en (L^A)T_EX [11].

64 DUX SUMMERSETI REGIMEN NORMANNIAE ASSEQUITUR

vectus et inde in Angliam revectus fuerat, cum nautis contenderet, nec eis pro voto satisfacere vellet, ab eisdem, commota sedicione, trucidatus est¹. Erat enim multis popularium exosus ex ea causa quod pro treugis cum Francis ineundis prorogandisque multociens laborasset et sepe propterea in Franciam trajecisset.

Aliorum vero^a principum Anglie cedes, que in suis civilibus^b discensionibus facte fuerunt, postea a nobis suis^c ordine atque loco perstringentur. Post dedicionem autem non sponte sed coacte ab Anglicis factam Cenomannorum civitatis, in dies, quamvis adhuc treuge per multos menses desponse^d essent ulro citroque durature, increscebant tamen odia, et particulares dietim cedes in limitaneis finibus tam Francorum quam Anglorum vel^e insidiose vel eciam in patulo agebantur^f.

CAPITULUM XI

QUOMODO DUX SUMMERSETI^g ASSECUTUS EST REGIMEN NORMANNIE, ET DE OPIDI FULGERIIS CAPCIONE ET DIREPCIONE^h.

Venerat autem dux Summerseti³ in Normanniam, missus ex Anglia ad regendam provinciam^h, pro cuius regentia diu multumque inter eum et duces Eboraci⁴ fuerat antea concertatum. Habebat quisque in consilio regis Anglorum quamplures suarum parcium studiosos, et vehementer uterque ipsorumⁱ ad assequendum provincie regimen aspirabat. Unde evenit ut per fautores parcium in Anglicano consilio, uni hodie provincia regenda, alteri

a. autem P. — b. civilibus suis P. — c. suo P. — d. desponsate P. — e. que vel P. — f. patrabantur P. — g. Summirseti G. — h. patriam corrigé en provinciam, dans la marge, de la main de l'auteur, en G. — i. eorum P.

FIGURE 2 – Exemple de texte avec apparat critique (ici, les notes 1, 2, etc. sont rejetées en fin de tome) – extrait de Thomas Basin, Histoire de Charles VII, éd. Belles Lettres, Paris, 1944.

par un élément `<lem>` (*lemma*) correspondant à la leçon la plus communément admise ou considérée comme la meilleure. Cet élément est optionnel et la TEI n'oblige pas le critique à utiliser la notion de texte de référence.

```
<app>
<lem wit='A D'>Eutaces</lem>
<rdg wit='B'>Eustasses</rdg>
<rdg wit='C'>Euthaices</rdg>
</app>
```

FIGURE 3 – Exemple de codage de l'élément `<app>`

Les éléments `<rdg>` peuvent apparaître directement sous l'élément `<app>` (comme dans l'exemple de la figure 3) ou être regroupés au sein d'un élément intermédiaire `<rdgGrp>`. On peut en effet par exemple juger bon de regrouper ensemble deux variantes pour indiquer qu'elles ne divergent du texte de référence que par l'orthographe.

```
<app>
<lem wit='A D'>though</lem>
<rdgGrp type=orthographic>
<rdg wit='B'>thogh</rdg>
<rdg wit='C'>thouh</rdg>
</rdgGrp>
</app>
```

FIGURE 4 – Regroupement de variantes

L'attribut *wit* de l'élément `<rdg>` permet de noter un sigle de manuscrit. L'élément `<rdg>` comporte plusieurs autres attributs (*resp*, *hand*, *vaseq*, etc.) qui permettent respectivement d'attribuer la responsabilité d'une leçon à un érudit donné³, d'indiquer qu'une variante donnée a été ajoutée par un scribe différent du scribe principal, de noter quand c'est possible l'ordre d'enchaînement des variantes, etc.

3. Par exemple, lorsqu'un passage difficile à déchiffrer a donné lieu à des interprétations divergentes.

3.2.2. Attributs de l'élément <app>

La plupart des attributs de l'élément <app> sont utilisés pour associer l'apparat critique au texte. la TEI propose pour ce faire plusieurs méthodes différentes (elles sont présentées en détail à la section 3.3). Selon la méthode retenue, certains des attributs décrits ci-dessous peuvent être ou non utilisés.

Les attributs pouvant être associés à une entrée de l'apparat sont :

<i>type</i>	cet attribut permet d'intégrer la ou les variantes contenues dans l'entrée à une classe de variante définie par l'éditeur ;
<i>from</i>	cet attribut identifie le début du lemme dans le texte de base ⁴ ;
<i>to</i>	cet attribut identifie la fin du lemme dans le texte de base ;
<i>loc</i>	cet attribut indique l'emplacement de la variante ⁵ .

3.3. Méthodes recommandées pour associer l'apparat critique au texte

Deux approches sont possibles. Certaines retiennent la notion de texte de base ou de référence, d'autres pas.

Si l'on estime qu'il existe un texte de référence, un ensemble de variantes peut être lié à une place donnée dans le texte, en spécifiant par exemple un numéro de ligne. Ce type de liaison est employé dans deux méthodes proposées par la TEI :

- la méthode dite *Location-referenced Method* ;
- la méthode dite *Double End-point Attachment Method*.

Si l'on n'utilise pas la notion de texte de référence, la TEI propose une méthode désignée sous le nom de *Parallel Segmentation Method*.

Le choix de telle ou telle de ces méthodes a des conséquences sur l'emplacement physique des balises correspondant à l'apparat critique. Dans le cas des méthodes *location-referenced* et *double-end-point-attached*, le balisage de l'apparat peut être mêlé au texte de référence ou stocké à l'extérieur du texte (à la suite dans le même fichier, dans un autre fichier, etc.)

4. Certaines méthodes recommandées par la TEI supposent l'existence d'un texte de référence.

5. Cet attribut n'est utilisé que lorsque l'on a recours à la méthode dite *Location-referenced Method*, que nous décrivons plus bas.

Dans le cas de la méthode *parallel segmentation*, l'apparat ne peut pas être inséré dans le texte. En effet, avec cette méthode, il n'existe pas de texte de référence : c'est la suite de paradigmes constitués par les empilements de variantes qui constituent le «texte»⁶.

3.3.1. Méthode dite location-referenced Method

Cette méthode est adaptée à la réalisation d'éditions imprimées. L'apparat est lié au texte de base en désignant globalement le bloc de texte contenant une ou des variantes.

La désignation du bloc s'appuie sur l'attribut *loc* de l'élément `<app>` qui prend pour valeur une division canonique du texte et/ou un numéro de ligne.

L'inconvénient de cette méthode est qu'elle ne désigne pas de façon précise les fragments de texte concernés par les variantes. C'est l'œil du lecteur qui doit retrouver ce fragment au sein du bloc.

3.3.2. Méthode dite double-end-point-attached

Pour remédier à l'imprécision qui caractérise la version précédente, la méthode *double-end-point-attached* permet de spécifier le début et la fin de la portion du texte de référence affectée par les variantes⁷.

Les attributs *from* et *to* de l'élément `<app>` sont utilisés pour noter ces limites. Leurs valeurs sont des identifiants SGML attachés aux éléments structurant le texte (les paragraphes, les lignes, etc.) ou, si ces derniers ne suffisent pas, à des éléments `<anchor>` spécialement insérés pour permettre de marquer une limite.

Nous donnons ci dessous (figure 6) un exemple d'utilisation de la méthode *double-end-point-attached*. Ici, on a choisi de coder l'apparat séparément.

6. Les deux aspects qui viennent d'être abordés (méthode pour lier l'apparat au texte et emplacement physique de l'apparat) sont spécifiés dans l'élément `<VariantEncoding>` de l'en-tête TEI. Exemple :

```
<VariantEncoding Method='location-referenced' location=external>
```

7. Cette liaison plus fine entre le texte et l'apparat permet d'envisager des traitements informatiques plus poussés que dans le cas de la méthode précédente.

```

<teiHeader>
<!-- ... -->
<variantEncoding method='location-referenced'
                  location='external'>
<!-- ... -->
</teiHeader>

<text>
<body>
<!--...-->
<div n=x1>
<seg id=1>Interim Romae C. Mamilius Limetanus tribunus plebis
rogationem ad populum promulgat uti quaeretur in eos quorum
consilio Jugurtha senati decreta neglegisset.
</seg>
</div>
</body>
</text>
<app loc='x1 1'>
<rdg wit=A>Interim</rdg>
<rdg wit='Q D F'>Interea</rdg>
</app>
<app>
<rdg wit='A Q'>rogationem</rdg>
<rdg wit='D F'>rogatione</rdg>
</app>
<!--...-->

```

FIGURE 5 – *Utilisation de la location-referenced method*

```

<teiHeader>
<!-- ... -->
<variantEncoding method='double-end-point' location='external'>
<!-- ... -->
</TeiHeader>
<text>
<body>
<!-- .. -->
<div n='i'>
<seg n=6 id='xxxii.6'>molles<anchor id =a1>aetate<anchor id=a2>
et fluxi<anchor id=a3></seg>
<!-- ... -->
</text>
<!--...-->
<app from='xxxii.6' to=a2>
<rdg wit='A'>molles aetate</rdg>
<rdg wit='B'>molles aestate</rdg>
</app>

<app from=a1 to=a3>
<rdg wit='A'>aetate et fluxi</rdg>
<rdg wit='C'>et aestate fluentes</rdg>
</app>

```

FIGURE 6 – Utilisation de la méthode de l'attachement double

Soient les trois versions données par les manuscrits A, B et C :

A : *molles aetate et fluxi*

B : *molles aestate et fluxi*

C : *molles et aestate fluentes*

On souhaite d'une part mettre en rapport les syntagmes *molles aetate* de A et *molles aestate* de B et d'autre part opposer la séquence *aestate et fluxi* de A et la séquence *et aestate fluentes* de C.

Ce type de chevauchement complexe (*overlapping variant*) peut être traité simplement par la méthode de l'attachement double.

3.3.3. Méthode dite Parallel Segmentation Method

Cette méthode n'est utilisable que dans les cas simples où les différents manuscrits peuvent être divisés en parallèle en un même nombre de segments.

Elle ne permet pas de traiter les cas complexes comme par exemple les chevauchements de variantes. Pour traiter ces cas, il est nécessaire de recourir à la méthode de l'attachement double qui a été décrite à la section précédente.

Soient les deux versions A et B

A : *Quamobrem in sententiam non addidisti uti prius
verberibus in eos animadverteretur?*

B : *Quamobrem in sententia non addidisti ut prius
verberibus in eos animadverteretur?*

Le codage de ces deux versions est donnée à la figure 7.

```
<div n=1>
<seg n=50 id=1.50>
Quamobrem in <app><lem wit=A>sententiam</lem>
<rdg wit=B>sententia</rdg></app>non addidisti <app>
<lem wit=A>ut</lem><rdg wit=B>uti</rdg></app>verberibus
in eos animadverteretur?
</seg>
</div>
```

FIGURE 7 – Utilisation de la méthode dite de «segmentation parallèle»

Dans l'exemple de la Figure 7, l'apparat est mêlé directement au texte. C'est d'ailleurs la seule possibilité avec cette méthode. Dans le cas des méthodes du simple et du double attachement, on a vu que l'apparat pouvait être directement mêlé au texte ou maintenu à part (à la suite du texte dans un même fichier, dans un fichier différent, etc.).

Au total, le choix d'une des trois méthodes dépend à la fois du degré d'automatisation souhaité (si l'on souhaite un maximum d'automatisation, mieux vaut éviter la méthode de l'attachement simple qui est trop imprécise) et de la complexité du texte à éditer (dans le cas d'un texte complexe, la méthode de segmentation en parallèle n'est pas utilisable).

4. Extensions souhaitables

Le chapitre de la TEI spécifiquement dédié au codage des apparats critiques offre la possibilité de réaliser des éditions critiques au sens classique du terme.

Dans certains cas, il est cependant souhaitable de disposer de mécanismes d'annotation plus complexes, soit pour associer au texte édité des interprétations de nature linguistique, soit pour donner pleinement aux manuscrits leur dimension de «textes à intérêt graphique» [3, 5].

4.1. Associer des interprétations de nature linguistique

La nécessité d'enrichir l'édition critique par l'ajout d'interprétations de nature linguistique paraît particulièrement pressante dans le cas de l'édition des manuscrits modernes où, pour décrire la genèse des textes, il est nécessaire de s'appuyer sur les traces des processus linguistiques mis en œuvre.

La TEI comporte d'ailleurs des sections sur le codage des analyses linguistiques qui présentent un intérêt dans cette perspective. Les chercheurs en traitement automatique du langage ont en effet à résoudre des problèmes qui ressemblent à la gestion des variantes d'un texte. Ne citons ici que les problèmes posés par l'alignement de corpus bilingues [4], la représentation d'ambiguïtés lexicales ou syntaxiques, la normalisation des textes préalablement à leur analyse automatique⁸.

Il est clair que si des codages de cette nature étaient utilisés dans le cadre d'éditions critiques, ils pourraient contribuer à donner à ces dernières une dimension linguistique plus riche. Comme nous l'avons déjà signalé, ceci serait particulièrement précieux dans le cas des éditions de textes modernes où cette dimension est indispensable.

4.2. Créer des liens normalisés entre le texte et l'image des manuscrits

Dans l'esprit de la norme SGML, sur laquelle reposent les recommandations de la TEI, les codages décrits jusqu'ici mettent l'accent sur la structure logique des documents. Mais il y a des types d'écrits où l'aspect physique est important ; notamment les manuscrits.

En s'en tenant à la version actuelle du chapitre de la TEI consacré à l'édition critique, il est certes déjà possible de coder certains aspects relatifs à l'aspect physique des textes.

L'élément `<witDetail>` permet par exemple d'attacher un commentaire à une leçon particulière d'un manuscrit donné. On peut ainsi par exemple indiquer

8. Nous donnons ci-dessous un exemple de normalisation lexicale des textes. Si l'on souhaite analyser automatiquement la séquence «*d'au moins*», il peut être utile, pour faciliter la consultation plus aisée des dictionnaires, de la ramener à la séquence «*de le a moins*», tout en gardant un lien avec la forme lexicale d'origine.

par un commentaire libre que le fragment de texte correspondant à une leçon débute par une lettrine.

L'élément `<witList>` peut servir à recenser tous les manuscrits utilisés dans l'édition critique. Cet élément est composé d'une suite d'au moins un élément `<witness>`, chaque élément `<witness>` correspondant à un manuscrit et ayant un attribut obligatoire *sigil* déclarant le sigle du manuscrit sous forme d'un identifiant SGML (ID)⁹. En plus de l'attribut *sigil*, il est possible de fournir à cet endroit une description du manuscrit (commentaires libres, description de type codicologique, etc.)¹⁰.

Cependant aucune description textuelle ne peut remplacer un renvoi à une reproduction de l'original. Ceci est particulièrement vrai pour les manuscrits enluminés ou les manuscrits littéraires modernes à la topologie souvent très complexe. Il serait donc particulièrement intéressant de coder de façon normalisée les liens entre le texte édité conformément à la TEI et les images des manuscrits utilisés dans l'édition critique. La TEI propose déjà un mécanisme de pointeurs étendus qui vont au delà des pointeurs SGML traditionnels dont la portée est restreinte à un seul fichier¹¹.

5. Conclusion

Le chapitre des recommandations consacré au codage des appareils critiques propose des mécanismes adaptés au cas des éditions philologiques traditionnelles (textes antiques et textes médiévaux), qu'il s'agisse d'obtenir des versions électroniques de ces éditions ou simplement de produire de façon plus efficace des versions imprimées.

Pour effectuer des opérations plus complexes comme l'association d'interprétations linguistiques fines à l'édition critique proprement dite ou le codage normalisé de liens entre textes et images, il est nécessaire de recourir à d'autres parties de la TEI, ou de proposer des extensions à cette dernière.

9. Les valeurs des attributs *wit* qui, comme on l'a vu en section 3.2, sont associées aux leçons sont des IDREF qui font référence à ces sigles. Le mécanisme SGML ID//IDREF permet ainsi d'être sûr que tous les sigles référencés dans l'apparat ont bien été déclarés dans la liste des manuscrits.

10. Ceci pose d'ailleurs le problème de la description des manuscrits dans les catalogues informatisés. Le format MARC utilisé dans la plupart des applications de bibliothèques sont conçus pour décrire des imprimés et non des manuscrits. Des projets de notices pour les manuscrits sont cependant en cours aux USA (*Research Libraries Group*) et en Allemagne (sous l'égide de la *Deutsche Forschungsgemeinschaft*). Il serait intéressant d'intégrer dans l'élément `<witness>` les éléments de données définis dans le cadre de ces projets.

11. En cette matière, il faut également être attentif aux mécanismes proposés par la norme HyTime [10].

De telles extensions seraient très précieuses pour des documents comme les manuscrits modernes ou les textes dont l'aspect graphique a une grande importance.

Bibliographie

- [1] Jacques ANDRÉ, «Traitement de texte et histoire des textes», *Le médiéviste et l'ordinateur*, XVI, automne 1986, 16–33.
- [2] Jacques ANDRÉ et Hélène RICHY, «Édition structurée et indexation hypertextuelle d'actes médiévaux», *Histoire et informatique*, textes réunis par Martine COCAUD, PUR, Rennes, 1995, p. 79–87.
- [3] Jacques ANDRÉ, Jean-Daniel FEKETE et Hélène RICHY, «Traitement mixte image/texte de documents anciens», *Cahiers GUTenberg*, n° 21, juin 1995, p. 75–85.
- [4] Patrice BONHOMME, Florence BRUNESSEUX et Laurent ROMARY, «Codage, documentation et diffusion de ressources textuelles», *Cahiers GUTenberg*, n° 24 (ce cahier), juin 1996.
- [5] Andrea BOZZI et Antonio SAPUPPO, «Computer-aided preservation and transcription of Ancient Manuscripts», *Ercim news*, 1994, n° 19, p. 27–28.
- [6] Michel GOOSSENS, «Introduction pratique à SGML», *Cahiers GUTenberg*, n° 19, janvier 1995, p. 27–58.
- [7] Nancy IDE et Jean VERONIS, «Présentation de la TEI: *Text Encoding Initiative*», *Cahiers GUTenberg*, n° 24 (ce cahier), juin 1996.
- [8] *Guidelines for electronic Text Encoding and Interchange*, Oxford–Chicago, mai 1994. ...
- [9] International Organisation for Standardization, *Langage normalisé de balises généralisé (SGML)*, ISO 8879-1986 (F), Genève, 1986.
- [10] International Organisation for Standardization, *Information technology: Hypermedia/Time-based Structuring Language (HyTime)*, ISO/IEC 10744-1992, Genève, 1992.
- [11] Reinhard WONNEBERGER, «Chapter mottos and optional semi-parameters», *TUGBoat* 1986, vol. 7, n° 3, 177–185.