Les *Guidelines* TEI à l'épreuve de la complexité textuelle et graphique médiévale

Marta Materni

Università degli Studi di Padova

marta.materni@gmail.com
https://orcid.org/0000-0002-1953-709X

Received: 22/02/2020; accepted 04/04/2021
DOI: https://doi.org/10.7203/MCLM.7.16724

The TEI *Guidelines* tested in the context of Medieval graphic and textual complexity

Abstract

The TEI *Guidelines* represent today an essential standard in the world of textual encoding for digital editions. However, some problems concerning TEI encoding, both theoretical and practical, are far from being definitively resolved. In this context we firstly discuss a theoretical question: what is the nature of the encoding? Are we interested in the Text or in its documentary image? What is the difference between a descriptive encoding and an editorial one? Secondly, we analyse the specific case of the TEI *Guidelines* application to the Medieval textual and writing reality: indeed, this application risks being anything but automatic, as TEI tags were set in the context of contemporary printed books.

Keywords

XML-TEI encoding; textual encoding; digital editions; Medieval writing

Magnificat Cultura i Literatura Medievals 8, 2021, 1-32.
http://ojs.uv.es/index.php/MCLM

ISSN 2386-8295

RESUMÉ

Les *Guidelines* TEI constituent aujourd'hui un standard incontournable dans le monde de l'encodage textuel préalable à l'édition numérique. Toutefois, certains problèmes aussi bien théoriques que pratiques, concernant l'encodage TEI, sont loin d'être définitivement résolus. Dans ce contexte on essaiera de discuter d'abord une question théorique : quelle est la nature de l'encodage ? Nous intéressons-nous au Texte ou à son image documentaire ? Quelle est la différence entre un encodage descriptif et un encodage éditorial ? Ensuite, on analysera le cas spécifique de l'adaptation des *Guidelines* TEI à la réalité du texte et de l'écriture médiévale : il s'agit en fait d'une application qui risque d'être tout sauf automatique, car les balises TEI naissent dans le contexte du livre imprimé moderne.

MOTS-CLEFS

Encodage XML-TEI; encodage textuel; édition numérique; écriture médiévale

Marta Materni. 2021. 'Les *Guidelines* TEI à l'épreuve de la complexité textuelle et graphique médiévale', *Magnificat Cultura i Literatura Medievals,* 8: 1-32, DOI: https://doi.org/10.7203/MCLM.7.16724

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under the Marie Sklodowslka-Curie grant agreement N° 886478.

TABLE DE MATIÈRES

1 Le contexte textuel et documentaire

La réalisation d'un projet d'édition numérique (édition diplomatique et interprétative) de plusieurs témoins d'un roman français du XII[e] siècle, le *Roman de Florimont* d'Aimon de Varennes,[1] a représenté une occasion de réflexion critique concernant l'efficacité sémantique des balises TEI actuellement disponibles face aux exigences d'un éditeur de textes médiévaux en langue vulgaire. Cette dernière précision n'est pas superflue par rapport à l'économie du discours. S'il est évident que tous les textes manuscrits, quelle que soit leur époque de composition, offrent au transcripteur un éventail de problématiques communes, il est également évident que les facteurs chronologiques et linguistiques ajoutent des déclinaisons individuelles à cette base commune. Finalité de l'écriture manuscrite (écriture personnelle ou écriture 'de copie'), gestion de l'espace page (la mise en page médiévale ou la topographie incontrôlable des dossiers génétiques),[2] état de cristallisation orthographique de la langue (de la phase de formation, encore sans l'appui théorique d'une grammaire, à la phase où l'on élabore la notion de faute d'orthographe), etc.: tous ces éléments nous obligent à chaque fois à une contextualisation de l'écriture manuscrite. Même le concept de transcription est à mon avis à contextualiser, à partir de la définition générale, inattaquable, de Peter Robinson selon qui la transcription est une traduction d'un système sémiotique à un autre:

> Transcription for the computer is a fundamentally interpretative activity, composed of a series of acts of translation from one system of signs (that of the manuscript) to another (that of the computer) [...]. Transcription is both decoding and encoding the text in the computer system. Will not be the same as the text of the primary source. Accordingly, transcription of a primary textual source cannot be regarded as an act of substitution, but a series of acts of translation from one semiotic system (that of the primary source) to another semiotic system (that of the computer). Like all acts of translation, it must be seen as fundamentally incomplete and fundamentally interpretative. (Robinson-Solopova 1993: 19, 21)

Dans la même ligne s'inscrivent les réflexions récentes de Pierre-Yves Testenoire qui, dans un article très dense et très clair, met bien en évidence la fécondité de la rencontre entre philologie et école génétique française en termes de réflexion autour du vocabulaire de la transcription:

> Toute transcription de manuscrits, indépendamment du cadre théorique dans lequel elle s'inscrit, est soumise à la tension résultant de la poursuite de deux objectifs contradictoires: un objectif de *vi-lisibilité* et un objectif de *lisibilité*. La *vi-lisibilité* est un néologisme forgé par Jacques Anis (1983) pour désigner les propriétés visuelles d'un texte qui participent de sa signifiance. L'exigence de *vi-lisibilité* est celle qui tend à restituer la disposition et les caractéristiques visuelles de l'écrit manuscrit. L'exigence de *lisibilité* tend inversement à présenter au lecteur le texte le plus clair et accessible [...]. C'est à cette double fidélité qu'est assigné le transcripteur dans son opération de transcodage: fidélité au système sémiotique de départ (l'écriture manuscrite) et fidélité au système sémiotique d'arrivée (l'écriture typographique). (Testenoire 2017: §11-12)

1. *DigiFlorimont. Archive numérique du "Roman de Florimont" d'Aimon de Varennes (Lyon, 1188)* http://digiflorimont.huma-num.fr et Materni 2020a et b.

2. D'Iorio 2010.

Dominique Stutzmann offre une définition différente de la 'transcription' qui, à mon avis, risque toutefois d'effacer complètement déjà au niveau théorique (la pratique les efface souvent *de facto*) les bornes entre 'transcription' et 'édition'/'encodage':

> Il ne faut pas, selon nous, réfléchir à la transcription [...] comme à la restitution où à la 'traduction' d'un texte, mais plutôt comme à sa description. Lors de la transcription, un premier document textuel donne naissance à un second ainsi qu'à de multiples descriptions. [...] Une structure de balisage telle que celle proposée par la TEI permet de décrire un 'texte' en créant un nouveau document où l'opération descriptive va plus loin que la transcription, en explicitant des implicites du texte. [...] Cette possibilité nouvelle d'établir une description formalisée d'un texte à des échelles plus grandes que l'échelle 1, où le document descriptif dit davantage que l'original décrit, modifie profondément l'univers documentaire actuel et la conception des opérations documentaire : *transcrire* et *encoder* ne sont pas des opérations de traduction, mais des opérations descriptives et d'explicitation. (Stutzmann 2011: 249)

Je préfère garder la distance entre les deux moments, en définissant la transcription comme une 'traduction' et l'édition/encodage comme une 'description'. Ou, encore mieux, la transcription comme une 'transposition graphique' et l'édition/encodage comme une 'représentation sémantique'.[3]

2 Les choix de principe de l'encodage *DigiFlorimont*

2.1 *Le choix du conformisme formel*

Malgré les aspects critiques qu'on essaiera de mettre en lumière et malgré l'autorisation de principe de la TEI à la personnalisation (*customization*),[4] le choix de principe du projet *DigiFlorimont* a été de ne pas sortir des limites officielles des balises TEI en proposant de nouvelles balises, car les problèmes affrontés n'étaient pas propres à ces textes spécifiques mais communs à tous les textes manuscrits médiévaux.[5] J'ai donc essayé de proposer des alternatives avec les balises à notre disposition, exploitant surtout les attributs. La modélisation de l'encodage *DigiFlorimont*[6] s'appuie sur un travail de 'recherche archéologique' qui a fouillé la documentation et les fichiers d'environ une dizaine de projets d'édition numérique de textes médiévaux, étalés dans le temps:

3. Sur la notion de transcription je renvoie aussi à Huitfeldt-Sperberg McQueen 2008, 2017 et 2018.

4. "Because the TEI *Guidelines* must cover such a broad domain and user community, it is essential that they be customizable: both to permit the creation of manageable subsets that serve particular purposes, and also to permit usage in areas that the TEI has not yet envisioned. Customization is a central aspect of TEI usage and the *Guidelines* are designed with customization in mind". (TEI Customization)

5. J'ai retrouvé la même attitude face à ce problème chez les responsables du projet DSL-*Diplomatarium Danicum* (https://diplomatarium.dk/): "because particular extensions to the standard are less likely to be universally deployed, we have exercised some self-constraint not to deviate from the standard. While thus is mainly to minimize the risk of rendering any effort obsolete, it also reflects a modest hope that such larger concentrations of standard-conformant material might stimulate further development of tools and technique". (Hansen 2012: § 11)

6. L'encodage *DigiFlorimont* est décrit dans Materni 2020b et, plus en détail, dans le mauel d'encodage en ligne en voie de publication (http://digiflorimont.huma-num.fr/ suivant le parcours Le site/Les textes/Manuel d'encodage).

1. BFM/BFM-MSS/QUESTE : *Base de Français médiévale* et *Queste del Saint Graal. Édition numérique interactive du manuscrit de Lyon (Bibliothèque municipale, P.A. 77)* (Heiden 2010, Lavrentiev 2008, Marchello-Nizia 2013), http://bfm.ens-lyon.fr
2. BVH : *Les Bibliothèques Virtuelles Humanistes* (BHV 2008), http://www.bvh.univ-tours.fr
3. CHARRETTE : *The Princeton Charrette Project* (Lavrentiev 2002a et b), http://www.princeton.edu/~lancelot/ss/
4. ESPANNA : *Estoria de Espanna Digital* (Espanna 2016), https://blog.bham.ac.uk/estoriadigital/
5. FROISSART : *The Online Froissart. A Digital Edition of the Chronicles of Jean Froissart* (Croenen-Romanova 2010), http://pcwww.liv.ac.uk/~gcroenen/Guidelines.pdf
6. LIBGLOSS : *The Liber Glossarum. A Digital Edition* (LibGloss 2011), http://liber-glossarum.huma-num.fr
7. MENOTA : *Medieval Nordic Text Archive* (Haugen 2019), https://www.menota.org
8. ORIFLAMMS : *Ontology Research, Image Features, Letterform Analysis on Multilingual Medieval Scripts* (Lavrentiev-Leydier-Stutzmann 2016), https://oriflamms.hypotheses.org
9. OTINEL : *Geste. Un corpus de chansons de geste* (Camps 2016), https://dev.chartes.psl.eu/elec/geste/
10. PIZAN : *Christine de Pizan. The Making of the Queen's Manuscript (London, British Library, Harley MS 4431)* (Pizan 2010), http://www.pizan.lib.ed.ac.uk
11. TITULUS : *Titulus. Corpus des inscriptions de la France médiévale* (Renedo Mirambell 2018), http://titulus.huma-num.fr

Ce travail a permis de faire ressortir l'existence d'aspects critiques récurrents, de hiatus évidents entre le modèle textuel à la base de la nomenclature TEI et la réalité de nos textes, d'exigences communes auxquelles on a dû donner, paradoxalement, des réponses individuelles et personnalisées.

2.2 `<choice>` *or* `<not choice>` *? Description sémantique ou pragmatisme éditorial ?*

Le deuxième choix de principe de l'encodage *DigiFlorimont* dérive d'une longue réflexion sur le sens profond de la définition 'balisage sémantique'. Comme résultat de cette réflexion, mis à part le cas où cette balise fonctionne en tant que conteneur du couple `<sic><corr>` (à défaut, malheureusement, d'une autre balise sémantiquement plus significative), la balise `<choice>` a été soigneusement évitée, et cela pour deux raisons.

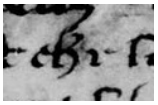
La première est d'ordre strictement théorique et elle répond à une vision précise de l'encodage textuel, tout à fait critiquable. Si la nature profonde de l'encodage proposé par la TEI est descriptive et sémantique, en utilisant la balise `<choice>` souvent nous ne décrivons pas sémantiquement notre texte mais nous le dédoublons (ou le triplons) graphiquement en proposant des alternatives éditoriales 'prêtes-à-afficher' selon les contextes, par ex. formes graphiques de l'édition diplomatique *vs*. formes graphiques de l'édition interprétative. Ce que j'ai essayé de faire était donc de réduire au minimum cette 'hybridité' dont parle explicitement par ex. Dominique Stutzmann, avec sa définition "hybrid or multi-layer digital transcription" fondée sur l'encodage des "alternative versions" du texte à travers la balise `<choice>` (Stutzmann 2015, Lavrentiev-Leydier-Stutzmann 2015), ou indirectement par ex. Elena Pierazzo, avec sa "paradigmatic edition":

> Perhaps these types of editions could be called paradigmatic editions, as they embed many alternative options for the same string of text in a nonlinear way, as opposed to editions that can only display the text in one format (such as printed editions, among others), which could instead be called syntagmatic editions. [...] In fact, one can store as much information or alternative points of view as one wishes for each string of text, following the paradigmatic axis, in a nonlinear way, and processed differently according to the editorial vision. (Pierazzo 2014: 4-5)

Ce "store" est un véritable stockage physique des données, voire, en réalité, une accumulation des couches éditoriales autour d'un même segment textuel. L'application extrême de ce principe a été la 'représentation diffractée' inaugurée par MENOTA et adoptée par BFM-MS/QUESTE:

> Nous utilisons le terme de représentation diffractée pour désigner la séparation systématique des niveaux de transcription. C'est à partir de cette représentation que sont générées les différentes formes d'éditions (visualisations) et c'est à cette représentation que sont rattachées les annotations linguistiques. (Lavrentiev 2008: 12)

Exemple de representation diffractée de l'occurrence du mot *cheualier* sous forme d'abréviation:



```
<w>
  <choice>
    <me:norm>chevaliers</me:norm>
    <me:dipl>che<ex>eualie</ex>rs</me:dipl>
    <me:facs><bfm:mdvAbbr>ch&apos;&rrot;</bfm:mdvAbbr>s</me:facs>
  </choice>
</w>
```

Il est évident que, si l'on multiplie cet encodage pour chaque mot composant un texte, le sucharge de verbosité devient un poids presque insoutenable.

De plus, si nous n'ajoutons pas d'attributs spécifiques par ex. au couple `<orig><reg>`, nous uniformisons de façon opaque toutes les interventions éditoriales au détriment de toute recherche quantitative sur le texte (c'est-à-dire une des raisons qui justifient justement le choix du numérique en matière d'édition). En présence de ces formules d'encodage:

```
<choice>
  <orig>cheualier</orig>
  <reg>chevalier</reg>
</choice>

<choice>
  <orig>lancelot</orig>
  <reg>Lancelot</reg>
</choice>
```

avons-nous décrit le texte? Quelle est l'information sémantique qu'on peut tirer de cet encodage? Quelle interrogation pourrait-on effectuer sur cet encodage? De quelle façon pourrait-on remarquer la différence entre, dans ce cas, l'opération de régularisation des majuscules selon l'usage moderne et la modernisation massive de la représentation des lettres ramistes? Il me semble,

je le répète, que, en l'absence d'attributs spécifiques, la généralisation du couple <orig><reg> risque de faire plonger dans l'anonymat indistinct les différentes opérations éditoriales. Pour rendre plus prégnant ce code on devrait au moins ajouter un attribut correspondant à chaque typologie d'intervention: par ex., modernisation de la représentation des lettres ramistes.

En d'autres termes, ce que nous classons et encodons est-il donc la description du texte ou l'acte éditorial? Et du point de vue numérique, notre but sera-t-il de générer les éditions à partir du texte de la source décrit sémantiquement ou nous limiterons-nous à utiliser le numérique pour choisir (*choice*) l'option, déjà explicitée, dont nous avons besoin selon le contexte? Ce texte encodé, que représente-t-il alors? Une accumulation de couches éditoriales?

À côté de cette raison théorique, il y a des raisons pratiques au refus d'une utilisation généralisée et massive de la balise <choice>. Dédoublant le texte, dans certains cas (qui dans le cas spécifique du texte médiéval représentent la majorité des occurrences) nous ne profitons pas des avantages offerts par le numérique en matière d'automatisation des actions et, surtout en présence de phénomènes de masse comme par ex. la modernisation de l'orthographe des lettres ramistes, nous accroissons de façon exponentielle la verbosité congénitale du langage XML.

À ce propos, il me semble qu'une tendance se manifeste, plus ou moins explicitement: celle à la simplification, au moins là où elle est possible. C'est à cette instance que répond par ex. la proposition d'un attribut @reg pour la balise <w>, avancée par le SIG (*Special Interest Group*) For Linguists:

> The TEI has at its disposal the powerful mechanism of //choice/reg|orig for such cases. However, with the usage of //choice/reg|orig for the annotation of regularized word forms, text annotation in specialized corpora may need to be extended significantly since there may already exist further (elaborate) annotation on or below the token level, which gets automatically multiplied with the use of <choice>. If the <choice> machinery is applied, these existing annotations will end up being represented twice: once in <reg> and once in <orig>. Thus texts might be significantly extended due to such unnecessary repetition, which may also negatively affect the sustainability of the resource. (SIG:LING 2017)

Aussi la proposition de Dominique Stutzmann et Alexei Lavrentiev (2015: 22), concernant l'encodage des allographes à travers la balise <g>, parle explicitement d'excès de verbosité par rapport à la solution classique de la triplette <choice><orig><reg>: "Very verbose. 46 code characters for 1 letter in the source!". Et la lourdeur du balisage et sa tendance à répéter plusieurs fois la même information étaient parmi les raisons à la base de la proposition alternative de Nicolas Mazziotta (2015) concernant les abréviations du français médiéval, dont on parlera plus loin mais qu'on peut déjà résumer dans le passage du système TEI <choice><abbr><am><expan><ex> au système simplifié <ex> @am.

En 2006 (2, 4) Nicolas Mazziotta a clairement indiqué les limites de l'approche diffractée[7] ou "approche fondée sur la séparation des vues [où] l'éditeur construit plusieurs éditions parallèles" (que Mazziotta considère "assez proche de la démarche traditionnelle", voire de l'édition-papier): "le constant aller-retour entre les différentes éditions en cas de correction ou d'amélioration de celles-ci; [...] la multiplication *ad infinitum* du nombre d'éditions distinctes, en fonction des intérêts des lecteurs potentiels; [...] la diffraction conceptuelle du texte". En revanche, cette solution est très facile à gérer: "On obtient facilement la représentation souhaitée (vue) en employant des feuilles de style XSLT". La deuxième approche est définie par Mazziotta (2006: 2, 3) "approche fondée sur la fusion des vues" ou "édition monolithique": "L'éditeur construit une seule édition,

7. Je renvoie aussi à Mazziotta 2004.

'farcie' d'une multitude d'informations combinant la description et l'élaboration des données". Pour ce qui est des avantages et des désavantages: "L'intérêt principal [*de cette approche*] est de maintenir l'unité conceptuelle du texte. [...] D'un point de vue technique, la conversion vers une visualisation particulière est plus difficile à gérer que dans le cas d'un encodage séparant les vues". Mazziotta (2016: 4, 6) propose finalement une troisième approche, très séduisante du point de vue théorique (je la définirais comme le modèle idéal d'encodage) mais, actuellement, assez difficile à gérer aussi bien au niveau de la production du code que de la visualisation. Il s'agit de l'approche basée sur les "pointeurs" (*pointer*), voire sur la séparation entre le texte et l'annotation: "Chaque caractère devant être décrit ou élaboré peut être délimité et identifié [*à travers un* `@id`]. L'ensemble des informations contenues dans d'autres attributs ou éléments peut donc être reporté hors du texte. [...] La description et l'élaboration éditoriale des caractères identifiés par un `@id` sont regroupées à l'extérieur du texte, dans un élément nommé graphData".

Suivant les suggestions derivant de ce genre d'approche, le contenu du fichier XML-TEI à la base des éditions *DigiFlorimont* a été conçu comme une traduction formalisée du texte (l'encodage est au niveau des mots, chacun identifié par un `@xml:id`) fondée sur une transcription graphémique linéarisée,[8] annotée sémantiquement de façon à permettre de produire automatiquement, à partir de ce texte formalisé et encodé, une édition diplomatique et une édition interprétative; il s'agit d'un modèle inspiré de l'édition monolithique proposée par Mazziotta mais dans la perspective de son approche par pointeurs. En d'autres termes, l'effort de l'encodeur est finalisé à la nette (ou, au moins, aussi nette que possible) séparation entre le texte et le discours autour du texte, les données et les méta-données, le niveau-transcription et le niveau-description.

Pour ce qui concerne la visualisation, les espaces blancs (ou leur absence), les diacritiques modernes, la régularisation des lettres ramistes et beaucoup d'autres interventions éditoriales ont été générées automatiquement à partir de la sémantique de l'encodage du fichier-source, sans mélanger, dans les limites du possible, le code descriptif du texte avec un code éditorial. Ce même principe a permis de gérer automatiquement la production des majuscules selon les normes de l'orthographe moderne: après un élément de ponctuation fort, au début du discours direct, en relation avec un nom propre.

8. Je fais ici référence à la nomenclature proposée par Robinson-Solopova (1993: 22): "graphic: every mark in the manuscript, every space, is represented in the transcription, even to the point of decomposition of letter forms into discrete marks [...] graphetic: every distinct letter-type is distinguished [...] graphemic: every manuscript spelling is preserved without distinction of separate letter forms as in a graphetic transcription [...] regularized: all manuscript spellings are regularized to a particular norm". Ou, en version française, selon André-Jimenes (2013: 117): "Transcription graphétique, distinguant précisément chaque variante graphique [...]. Transcription graphémique, respectant la graphie (*spelling*) de chaque mot [...] mais sans tenir compte des éventuelles variations graphiques de ces différentes lettres [...]. Transcription régularisée, qui normaliserait la graphie des mots, développerait les abréviations". Lavrentiev-Leydier-Stutzmann (2015) identifient la transcription graphétique avec la "paleographic (allographetic) transcription": "[*they*] are also called 'imitative' but we prefer the term of 'allographetic' that corresponds to the linguistic status of letters variants in the frameworks of the theory of graphemics. This kind of transcription represents very faithfully the aspect of the source document writing. It distinguishes major letter variants [...] reproduces abbreviations markers, original punctuation, and word segmentation. It's a rule, allographetic transcriptions reproduce (or represent somehow) the layout of the original document (including line breaks, scribal correction markers, marginalia, etc.)".

3 L'édition diplomatique comme banc d'essai de l'encodage numérique

Le banc d'essai, ou l'écueil, d'un balisage TEI face à un texte médiéval est naturellement la création d'une édition diplomatique. On constate que, malgré les discussions et la pratique séculaire, le champ terminologique lié aux éditions centrées sur une source primaire présente encore des marges d'incertitude, de superposition, d'ambiguïté, en plus des différences liées aux traditions philologiques nationales.[9] J'ai donc pris comme point de référence la définition classique de François Masai (1950: 185): l'édition diplomatique est un "relevé archéologique des textes, tels qu'ils sont transmis par les manuscrits existants". En conclusion de ce bref article, d'une clarté et d'une logique désarmantes qui ne laissent place à aucune contestation, Masai énumère les éléments manuscrits à transférer dans une édition diplomatique, nous invitant toujours, plus ou moins implicitement, à nous montrer prudents face au risque du fétichisme documentaire. On notera donc "l'aspect du document", tout en nous souvenant que "l'essentiel de l'édition diplomatique ne réside pas en cette conformité cependant" (Masai 1950: 188). On notera les abréviations: "la seule chose qui importe est que le lecteur n'en ignore pas la présence [...] en cas de solution, il est indispensable de noter les lettres qui y représentent l'interprétation de l'éditeur" (*ibid.*). Pour la *scriptio* du manuscrit: "Pareille pratique est, en droit, parfaitement inattaquable, mais est-il opportun de donner au lecteur moderne ce surcroît de difficulté [...]? Certes, lorsque le texte est susceptible d'être coupé de diverses façons ou qu'une erreur de scribe est intimement liée à la *scriptura continua*, il faudra maintenir celle-ci" (*ibid.*). Mais "l'effort principal doit porter sur le relevé méticuleux des altérations subies par les textes" (*ibid.*).

Pour ce qui est de la segmentation de l'écriture, la réserve de Masai était donc liée exclusivement à des raisons qui, en réalité, ont à faire avec l'"économie": si l'on est obligé d'imprimer physiquement une seule version, on est obligé aussi d'accepter quelques compromis permettant d'atteindre un public plus large. Mais ces raisons s'effacent complètement en présence du support numérique, qui engage bien sûr d'autres frais mais qui, au moins, ne connaît pas le prix du papier et de l'encre. La possibilité d'offrir une version interprétative à côté de la version diplomatique, et même une triple version diplomatique/semi-diplomatique/interprétative, satisfera enfin à la fois les exigences d'une lecture plus aisée et celles d'une analyse plus pointue tout en élargissant le public de ce travail scientifique.

> Par rapport au manuscrit, une édition de type non diplomatique fait donc toujours gagner de l'information, en même temps qu'elle en fait perdre. Elle augmente pour le lecteur moderne la lisibilité du texte, mais au détriment d'une certaine fidélité au document. [...] les manuscrits offrent un terrain d'enquête incomparablement plus riche que n'en fournissent les éditions modernes et ce à tous les niveaux de la langue, de la phonétique à la syntaxe en passant par la morphologie. [...] S'agissant de la morphologie, seul le retour au document permet de voir les phénomènes d'agglutination et déglutination [...] Au niveau de la syntaxe enfin, la ponctuation médiévale, susceptible de renseigner sur la 'phrase médiévale', restera elle aussi cachée. Plus précisément, nos manuscrits sont aussi une mine d'information parce que l'on peut étudier grâce à eux les pratiques orthographiques des copistes. C'est à dire qu'ils nous renseignent sur un état de langue tel que le percevait tel ou tel copiste contemporain, et permettent de mesurer comment les différents usages orthographiques tiennent compte des réalités linguistiques. (Cazal *et al.* 2003: 2, 4)

9. Meyer 1910, Roques 1926, Ménard 1990, Duval 2012, Lavrentiev 2017 et 2019. Foulet-Speer 1979 ont été les premiers à aborder le problème de la segmentation graphique.

3.1 *La* scriptio *du manuscrit: la segmentation irrégulière*

Après la première difficulté représentée par le déchiffrement de l'écriture, l'édition d'un texte médiéval comporte un second défi: la segmentation correcte de la chaîne de l'écriture selon les paramètres modernes d'individualisation de chaque mot.[10] Si dans la plupart des cas cette segmentation ne présente pas de difficultés, dans certaines situations il y a quand même une marge d'ambiguïté qui appelle à l'interprétation. La fidélité à la *scriptio* du manuscrit au niveau de l'édition diplomatique correspond donc à la vision de cette dernière en tant que "relevé archéologique du texte" proposée par Masai. Mais en plus de cette fonction 'documentaire', la segmentation originale du texte manuscrit nous fournit des indices importants concernant le niveau de la conscience grammaticale naissante, à une étape d'évolution des langues romanes qui ne connaît pas encore la systématicité d'une théorie grammaticale bien enracinée.

Si dans le cas de la *scriptio continua* de l'Antiquité nous sommes en présence d'une forme particulière de la mise en écriture du texte, abolissant complètement les espaces blancs en tant que séparateurs formels entre les mots,[11] la *scriptio* du manuscrit médiéval est plutôt une segmentation irrégulière mais pas du tout liée au hasard. De nombreuses études ont amplement montré l'existence de lignes de tendance bien précises, en particulier dans les phénomènes d'agglutination, et ont commencé à éclairer les mécanismes linguistiques à l'origine de ces agglutinations constantes, manifestations tangibles d'un apprentissage où la langue de la communication est la langue romane et la langue de la grammaire est le latin.[12] Concernant les implications de l'étude des séquences graphiques d'un point de vue d'histoire linguistique, je renvoie directement aux mots de Nelly Andrieux-Reix et Simone Monsonégo (1998: 34-36):

> Ces tendances [*dans la composition des séquences*] comme ces impossibilités mettent au jour une forme d'analyse syntaxique sous-jacente à la production de séquences comme, complémentairement, à la segmentation de la chaîne graphique; là où (et c'est fréquent) ces séquences ne correspondent pas à des syntagmes entiers, la nature et les fondements de cette analyse restent à découvrir. [...] Tout cela invite à interroger ce qui nous est parvenu de l'enseignement médiéval tant en matière d'écriture elle-même que de grammaire. [...] En termes plus simples, quels sont les rapports entre la séquence, une donnée de l'écrit et le 'mot' entendu comme notion métalinguistique construite?

Les nombreuses et récentes études d'Elena Llamas-Pombo ont attiré ultérieurement l'attention sur cet aspect de la textualité manuscrite médiévale, en plus du système de ponctuation.[13] Le fait d'accroître le corpus textuel à disposition dans ce domaine est d'autant plus important que les éditions respectant la segmentation et la graphie originale du manuscrit représentent depuis

---

10. "L'actuel 'mot graphique' y [*dans les écritures médiévales*] était pratiqué [...] et ce depuis les plus anciennes représentations écrites à nous être parvenues du français; mais ce 'mot graphique' a toujours été en cooccurrence avec d'autres séquences l'excédant, constituées par ce qui nous apparaît maintenant comme l'agglutination de généralement deux ou trois 'mots', rarement plus. Si donc l'espace du français écrit a toujours été ponctué de blancs, les séquences graphiques ainsi délimitées étaient particulièrement variées au Moyen Age". (Andrieux Reix-Monsonégo 1998: 30-31).

11. Sur la notion de 'mot' on peut consulter la synthèse offerte par Catach 1998.

12. Rickard 1982, Andrieux Reix-Monsonégo 1997, 1998 et 2000, Hasenhor 1998, Baddeley-Biedermann Pasques 2004, Balon 2016, pour n'en citer que quelques-unes. Sur le rapport avec le latin, je renvoie aux mots de Baddeley-Biedermann Pasques (2004: 188): "Les scribes ayant appris le latin par l'enseignement et par la fréquentation de textes et d'auteurs latins [...] recopient des manuscrits latins et conservent pour la mise en écriture de textes en langue vernaculaire, qui n'ont pas encore de longue tradition graphique, les processus d'écriture ou 'habitudes d'écriture' de la langue latine; en même temps ils empruntent à la tradition latine le matériel graphique et les règles graphiques pour la notation des textes. Cet aspect cognitif n'est pas à négliger dans l'histoire de l'orthographe de cette période".

13. Llamas Pombo 2001, 2009, 2016 et 2017.

toujours une nette minorité, pour des raisons économiques de production mais aussi, il ne faut pas l'oublier, par négligence intellectuelle: "pendant longtemps la pratique des copistes concernant la 'syntaxe graphique' des manuscrits a été renvoyée à l'aléatoire de la 'variation graphique' sans plus de recherches systématiques ou perçue comme un geste vain ne méritant pas notre attention" (Balon 2016: 305). J'ajouterais que, au-delà de la linguistique, dans une perspective strictement philologique, la segmentation irrégulière peut être à l'origine de variantes ou d'erreurs de copie, et qu'elle n'est pas à négliger au moment de la collation de passages textuels particulièrement difficiles.

La version actuelle des *Guidelines* ne nous fournit pas d'instruments satisfaisants et surtout sémantiquement significatifs. Le problème a été en effet résolu de façon différente par plusieurs projets, ce qui signifie en dernière analyse:

1. en positif, que le problème existe et qu'il exige donc une solution univoque partagée;
2. en négatif, que les fichiers TEI-XML actuellement en circulation proposent des solutions personnelles pour un problème commun affectant l'ensemble de la communauté des philologues romanes.

3.1.1 Panorama des solutions proposées

CHARRETTE. Pour CHARRETTE le respect de la segmentation de l'écriture manuscrite représente un des traits marquants de l'édition diplomatique. L'encodage des manuscrits du *Roman de la Charrette* a subi une double évolution: de SGML à TEI-XML,[14] et ensuite, une fois dans le contexte TEI, de 'texte libre' à texte encodé au niveau des mots. Par conséquent, deux solutions ont été apportées à ce problème.

Premier cas, 'texte libre':

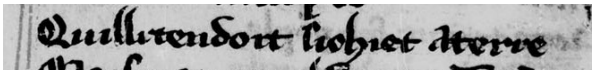
- balise `<chr_sb/>`[15] pour encoder l'espace blanc qui brise un mot déglutiné;
- balise `<chr_aggl>` pour encoder le segment textuel aggluti né; à l'intérieur de ce segment chaque mot est identifié par une balise `<chr_ap1>`, `<chr_ap2>`, etc.; la nature du segment est explicitée par l'attribut `@type` avec les valeurs *`simple`* (agglutination entre deux mots, sans élision), *`complex`* (agglutination entre trois ou plusieurs mots), *`elision`*; l'incertitude dans la détermination de l'absence effective d'un espace est marquée par l'attribut `@cert`, avec la valeur *`yes/no`*; pour marquer l'élision à l'intérieur d'un segment d'agglutination complexe, on introduit un troisième attribut, `@elision`, avec la valeur *`yes/no`*.

Exemple d'encodage selon le modèle-CHARRETTE:[16]

14. SGML (*Standard Generalized Markup Language*): il s'agit du métalangage pour l'élaboration de langages de *markup* réconnu standard ISO en 1986 et déscendant de GML (*Generalized Markup Language*), créé au sein de l'IBM par Charles Goldfarb, Edward Mosher et Raymond Lorie aux années 1960, avec l'objectif de séparer les données et la répresentation des données et, en même temps, rendre les données *machine-readable*. Le langage HTML (*HyperText Markup Langage*) est une application de SGML tandis que le métalangage XML (*eXtensible Markup Language*) est son successeur.

15. La sigle `chr` (CHARRETTE), de même que, dans les exemples qui suivent, `bfm` (BFM), `me` (MENOTA), `ori` (ORIFLAMMS), est un *namespace* indiquant une balise introduite par un projet et ne faisant pas partie du *schema* TEI.

16. Toutes les images sont tirées du ms. Tours, Bibliothèque Municipale, 941 (XIII^{e} s.), contenant le *Roman de Florimont.*



```
<chr_aggl type="complex" cert="yes">
  <chr_ap1 elision="yes">Qu</chr_ap1>
  <chr_ap2>il</chr_ap2>
  <chr_ap3>li</chr_ap3>
  <chr_ap4>tendoit</chr_ap4>
</chr_aggl>
<chr_aggl type="simple" cert="yes">
  <chr_ap1>si</chr_ap1>
  <chr_ap2>chiet</chr_ap2>
</chr_aggl>
```

Deuxième cas, encodage au niveau des mots: l'encodage est très simplifié parce que la balise `<w>` est accompagnée par l'attribut `@aggl` avec les valeurs *`empty`*, *`simple`* et *`elision`*. Voilà donc l'encodage du même segment selon le deuxième modèle-CHARRETTE:

```
<w aggl="elision">Qu</w>
<w aggl="simple">il</w>
<w aggl="simple">li</w>
<w aggl="empty">tendoit</w>
<w aggl="simple">si</w>
<w aggl="empty">chiet</w>
<w aggl="simple">a</w>
<w aggl="empty">terre</w>
```

MENOTA. Ce projet aussi, dont la granularité d'encodage est au niveau des mots (tous lemmatisés), opte pour l'idée de 'segment', avec la balise `<seg>` (on est désormais en pleine époque TEI). La déglutination n'est pas marquée et dans ce cas le mot, avec son espace blanc, est tout simplement enveloppé par la balise `<w>`:

```
<w lemma="chevalier">ch eualiers</w>
```

L'agglutination est encodée à travers la balise `<seg>` et l'attribut `@type` avec la valeur *`nb`* (*no break*):

```
<seg type="nb">
  <w lemma="son">si</w>
  <w lemma="cheoir">chiet</w>
</seg>
<seg type="nb">
  <w lemma="a">a</w>
  <w lemma="terre">terre</w>
</seg>
```

On prévoit en plus le cas de l'enclise, et dans ce cas la valeur de `@type` sera *`encl`* (`<seg type="encl">`) tandis que le cas de l'élision n'est pas considéré. On peut remarquer finalement que les valeurs de l'attribut `@type` oscillent entre la dimension physique (*no break*) et la dimension grammaticale (*encl*), et que le phénomène de la déglutination, *de facto*, n'est pas signalé.

BFM-MSS/QUESTE. Selon les pratiques d'encodage BFM-MSS, le respect de la segmentation originale de l'écriture manuscrite est un critère éditorial pertinent pour l'édition facsimilaire. On se souviendra du fait que les éditions lyonnaises sont fondées sur la représentation diffractée inaugurée par MENOTA: les balises concernant la segmentation s'inscrivent donc au niveau de la couche éditoriale représentée par `<me:facs>`.

La déglutination est marquée au niveau de l'espace de déglutination par la balise `<bfm:sb/>` (accompagnée de l'attribut `@cert`):

```
<w>
  <choice>
    <me:norm>chevalier</me:norm>
    <me:dipl>cheualier</me:dipl>
    <me:facs>ch<bfm:sb cert="yes"/>eualier</me:facs>
  </choice>
</w>
```

Dans le manuel de la BFM-MSS l'agglutination est signalée par l'attribut `@bfm:aggl` appliqué à chaque `<w>`, avec les valeurs *`simple/elision`*, et accompagné du deuxième attribut `@bfm:agglCert`, avec les valeurs *`y/n`*.

```
<w bfm:aggl="elision" bfm:agglCert="y">
  <choice>
    <me:norm>Qu</me:norm>
    <me:dipl>Qu</me:dipl>
    <me:facs>Qu</me:facs>
  </choice>
</w>
<w bfm:aggl="simple" bfm:agglCert="y">
  <choice>
    <me:norm>il</me:norm>
    <me:dipl>il</me:dipl>
    <me:facs>il</me:facs>
  </choice>
</w>
<w bfm:aggl="simple" bfm:agglCert="y">
  <choice>
    <me:norm>li</me:norm>
    <me:dipl>li</me:dipl>
    <me:facs>li</me:facs>
  </choice>
</w>
<w>
  <choice>
    <me:norm>tendoit</me:norm>
    <me:dipl>tendoit</me:dipl>
    <me:facs>tendoit</me:facs>
  </choice>
</w>
```

Dans les fichiers QUESTE on trouve un autre vocabulaire: la structure de l'encondage reste inchangée mais l'attribut `@bfm:aggl` est substitué par l'attribut `@rend` (avec les valeurs *`aggl/elision`*).

ORIFLAMMS-OTINEL. Le projet ORIFLAMMS, sous-entendant lui aussi une granularité de l'encodage au niveau des mots, change toutefois complètement les règles du jeu en proposant encore une fois l'emploi de l'attribut `@rend` pour indiquer l'entité de l'espace blanc après le mot, mais en connexion avec les valeurs suivantes: *`normal/long/short/none`*. Pour la déglutination on continue, par contre, à s'appuyer sur la balise `<space/>`, donc à marquer l'espace anomal de déglutination. En s'intéressant à la 'dimension de l'espace', ce dernier projet, né dans le contexte d'une recherche paléographique, déplace donc de façon décisive la perspective par rapport à la description du texte: du texte en tant qu'objet linguistique on passe au texte en tant qu'objet graphique, système d'enchaînement des signes de l'écriture, du moment que "le texte est une image comme les autres" (Stutzmann 2011: 249). Jean Baptiste Camps, dans OTINEL, reprend cet attribut, `@rend`, mais lui assigne à nouveau une connotation sémantique à travers les deux valeurs *`aggl`* et *`elision`*.

```
<w rend="none">Qu</w>
<w rend="none">il</w>
<w rend="none">li</w>
<w rend="normal">tendoit</w>
<w rend="none">si</w>
<w rend="normal">chiet</w>
<w rend="none">a</w>
<w rend="normal">terre</w>
```

3.1.2 La proposition *DigiFlorimont*

La matérialité du phénomène est aussi à la base de l'attribut TEI `@join` qui a été récemment ajouté à la balise `<w>`[17] et qui n'était pas disponible au moment où le modèle d'encodage *DigiFlorimont* a été élaboré. Mais cet attribut n'aurait pas toutefois exprimé correctement la philosophie d'encodage que je voulais proposer: de mon point de vue le problème n'est pas d'indiquer si deux mots sont physiquement joints (*joined*) mais quelle est la nature de cette jonction afin de pouvoir ensuite interroger de façon ponctuelle le texte. Les deux chaînes graphiques "*lautre*" et "*lecheualier*" n'ont pas le même poids en termes d'analyse: car, dans le premier cas, la jonction physique est due à l'élision de la voyelle de l'article et à l'absence de diacritiques dans le système graphique médiévale; dans le deuxième cas, il s'agit d'une véritable agglutination. Cette taxonomie des liens entre les mots – agglutination, enclise, élision – permet en outre de gérer automatiquement, au niveau de la visualisation des données, aussi bien la production des espaces blancs que la production des signes diacritiques selon les règles de la *scriptio* moderne.

Pour ce qui concerne l'encodage *DigiFlorimont*, j'ai donc d'abord essayé d'utiliser la balise `<seg>`, qui semblait la plus fonctionnelle en présence de segments complexes mélangeant agglutination et élision. Mais rapidement je me suis heurtée au problème de l'*overlapping* dont voici un exemple (avec un code simplifié):

Par ex. *leroi demacedoine*

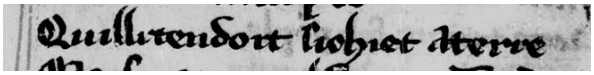
```
<seg>
   <w>le</w>
```

17. Cet attribut fait partie du *set* att.linguistic (https://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-att.linguistic.html), avec `@lemma`, `@lemmaRef`, `@pos` et `@msd`.

```
    <rs>
      <w>roi</w>
</seg>
<seg>
    <w>de</w>
    <w>macedoine</w>
</seg>
    </rs>
```

J'ai donc opté pour la caractérisation de chaque mot concerné par une altération et j'ai utilisé l'attribut (peut-être un attribut passe-partout) `@ana`, avec les valeurs:

- `degl` (déglutination) + `<space type="#degl"/>`
- `encl` (enclise)
- `elis` (élision)
- `aggl_s` (agglutination simple, deux mots)
- `aggl_c` (agglutination complexe, trois éléments ou plus)
- `aggl_s_unc` et `aggl_c_unc` (*uncertain*, pour indiquer les cas ambiguës).



```
<w ana="#elis">Qu</w>
<w ana="#aggl_c">il</w>
<w ana="#aggl">li</w>
<w>tendoit</w>
<w ana="#aggl_s">si</w>
<w>chiet</w>
<w ana="#aggl_s">a</w>
<w>terre</w>
```

Le CSS gère la production (ou l'absence) des espaces blancs et, en plus, celle des diacritiques dans la version interprétative. L'accent est donc mis sur la nature de la jonction (ou de la disjonction) et pas sur sa matérialité. Le système est satisfaisant sauf: quand il s'agit d'une séquence mixte qui mélange agglutination et élision car dans ce cas nous perdons un peu l'unité *de facto* du segment en brisant cette chaîne graphique unique en sous-segments agglutinés et sous-segments liés par une élision; et par rapport au dernier mot de la séquence d'agglutination, qui n'est pas caracterisé.

Au moment de publier cet article, une année environ après sa rédaction, l'encodage *DigiFlorimont* a encore évolué jusqu'à résoudre les deux problèmes que je viens de mentionner. À côté du travail philologique, le projet *DigiFlorimont* a representé en fait l'occasion pour entamer une réflexion autour de l'ergonomie des outils de travail à disposition des humanistes engagés dans un projet d'édition numérique. Le résultat de cette réflexion a été la création d'un *editor* TEI, TEI-Medit, capable de produire le code XML-TEI de façon automatisée à partir d'un fichier .txt (le fonctionnement de cet outil est décrit dans les détails dans Materni 2020c). La deuxième version de l'editor TEI-Medit permet désormais de gérer de façon très simple l'écriture d'un *stand-off markup*, un modèle d'encodage que j'avais d'abord utilisé de façon limitée, en connexion avec les discours directs/monologues. Une fois le système d'écriture du code évolué et standardisé, j'ai pu appliquer le même principe d'encodage au phénomène de l'agglutination, en revenant ainsi à l'idéé

initiale de "segment", beaucoup plus correcte de mon point de vue interprétatif, tout en evitant le piège de l'*overlapping*. Il s'agit encore une fois de séparer les plans: d'un côté les mots composant le texte et de l'autre côté les descriptions, traduites sous forme d'encodage, de ces mots. L'encodage de l'agglutination n'est rien d'autre, alors, que la description de la "texture physique" créée par les mots sur la ligne de l'écriture, une texture physique qui toutefois ne se reduit pas à une question d'espace (absent, présent, long, court, etc.) mais qui constitue, par contre, la manifestation phénoménique du niveau de conscience linguistique atteint. Qui plus est, l'agglutination ne concerne pas en réalité le mot lui-même mais la relation entre chaque mot et ses "voisins": en d'autres termes, il ne s'agit pas d'une propriété intrinsèque, comme le lemme ou la déclinaison, mais d'une propriété contextuelle fondée sur la notion de rélation (avec les autres).

Les fichiers *DigiFlorimont* mis à jour offrent donc cette nouvelle solution d'encodage:

- Au niveau des mots, `<w>`, on garde l'attribut `@ana` avec les valeurs `degl/encl/elis`;
- Le segment agglutiné est encodé à travers la balise `<span>` (avec les attributs `@type agglutination/agglutination_uncertain @from @to`) et tous les *spans* sont rassemblés au fond du document dans un `<div type="agglutination">`:

```
<w xml:id="w01" ana="#elis">Qu</w>
<w xml:id="w02">il</w>
<w xml:id="w03">li</w>
<w xml:id="w04">tendoit</w>
<w xml:id="w05">si</w>
<w xml:id="w06">chiet</w>
<w xml:id="w07">a</w>
<w xml:id="w08">terre</w>

<div type="agglutination">
  <span type="agglutination" from="w01" to="w04"/>
  <span type="agglutination" from="w05" to="w06"/>
  <span type="agglutination" from="w07" to="w08"/>
</div>
```

3.2 *Les interventions massives (1): les lettres ramistes*

Jusqu'à la date de 1563, quand Ramus résout enfin le problème tout en ouvrant une nouvelle saison orthographique, l'alphabet latin ne disposait pas de graphèmes distincts pour indiquer les phonèmes suivants: voyelles /u/ et /i/, consonne /v/ et semi-consonne /j/. Les deux lettres graphiques u/i (V/J), appelées ensuite "lettres ramistes", couvraient toute la casuistique. Il ne s'agit pas d'une question banale, parce que, si l'équivalence phonique est claire dans la majorité des occurrences, selon les époques et les domaines linguistiques nous gardons quand même une marge d'incertitude qui appelle à l'interprétation. C'est le cas typiquement, par ex., des deux phases évolutives du futur du verbe *avoir* et *savoir* en français (Baker 1937). Les *Conseils pour l'édition des textes médiévaux* de l'École des Chartes (2014: 23-24, 25) nous donnent des indications précieuses pour la discussion:

> [*distinction de -i- et de -j-:*] on transcrit par "j" la lettre "i" ayant valeur de consonne indépendamment de la graphie [...] employée par le scripteur. [...] Il faut donc tenir compte de la nature du phonème correspondant à la graphie "i" [...] On doit proscrire la transcription en "j" du "i" plongeant en fin de mot, cette dernière forme étant purement graphique.

> [*distinction de -u- et de -v-:*] en français, un doute peut exister sur la valeur, consonantique ou vocalique, de la graphie "u". En onomastique l'aboutissement phonétique du nom peut donner des indications [...] Pour les formes verbales, en particulier pour les futurs et les formes en *-roie* des verbes "avoir" et "savoir", on imprime généralement "v" pour les textes d'ancien français et "u" pour les textes de moyen français.

Le trait caractéristique d'une édition diplomatique est le respect de la graphie du manuscrit et par conséquent, évidemment, le maintien des lettres ramistes dans leur état original: la question se pose donc de façon inévitable. Et les termes de cette discussion sont tous présents dans les lignes citées, là où l'on parle de: "graphie", "forme purement graphique", "phonème correspondant à la graphie". Ce phénomène superpose en fait deux niveaux d'analyse: le niveau paléographique et le niveau linguistique, qui ne coïncident pas totalement ni en termes de résultats ni, surtout, à l'échelle de la terminologie. La discussion qui suit s'inspire de ce cas particulier, mais l'intention réelle est de soulever une question beaucoup plus générale: celle qui concerne l'absence totale, parmi les balises TEI, d'un système de notation pour décrire la phonétique.

La seule discussion dédiée au problème des lettres ramistes au niveau de l'encodage est, à ma connaissance, la proposition faite par Alexei Lavrentiev et Dominique Stutzmann (2015) d'encoder les allographes – catégorie dans laquelle retomberaient nos lettres ramistes, mais aussi les lettres majuscules – avec la balise `<g>` au lieu du système habituel de la triplette `<choice><orig><reg>`. La balise `<g>` serait sémantiquement renforcée par deux attributs: `@ref`, pour le côté graphique et descriptif (un pointeur pour un `<charDecl>`); et `@ana`, avec les valeurs `dipl/norm` (*diplomatic/normalized*) pour donner des indications éditoriales alternatives à la forme encodée:

```
1. <g ana="ori:dipl-small">D</g>ieu
2. <g ana="ori:norm-u ori:norm-caps">v</g>ne
```

En d'autres termes, l'encodeur peut décider de réaliser la transcription à n'importe quel niveau (diplomatique ou déjà normalisé), en donnant en même temps des indications concernant l'"autre' niveau à travers les valeurs de l'attribut `@ana`. On doit donc interpréter les exemples cités ci-dessus selon ce schéma:

1. la transcription est déjà normalisée et l'encodage nous signale que la lettre majuscule était minuscule (`small`) dans le document (`dipl`): Dieu<dieu;
2. la transcription est diplomatique et l'encodage nous signale que l'allographe "v" correspond, au niveau normalisé (`norm`) à "U" (`u, caps`): vne>Une.

Encore une fois ce qu'on met en évidence dans la discussion Lavrentiev-Stutzmann est la verbosité quelquefois exaspérante de XML:

```
<choice>
  <orig><c>V</c></orig>
  <reg><c>u</c></orig>
</choice>[18]
```

18. Ou la répétition du mot entier: `<choice><orig>`cheualier`</orig><reg>`chevalier`</reg></choice>`, la solution qui correspond de façon TEI conforme à l'encodage MENOTA/QUESTE.

d'autant plus exaspérante quand on se confronte à de véritables phénomènes de masse du point de vue graphique.[19] Il me semble pouvoir interpréter de cette façon la conclusion des deux auteurs: "Questions to the TEI. What annotation mechanism to use for 'simple' normalization cases?" (Lavrentiev-Stutzmann 2015: 29).

Le discours qui suit exclut la référence aux majuscules que j'appellerais 'éditoriales' (en connexion avec un nom propre, au début du discours direct, après un signe de ponctuation fort): on a dit que le modèle d'encodage *DigiFlorimont* permet de produire ces majuscules de façon automatique sur la base de la sémantique des balises et des attributs, car le principe-guide est de ne pas encoder les interventions éditoriales. Pour ce qui est des allographes, je propose d'abord de faire une distinction entre les allographes du point de vue paléographique et les allographes du point de vue linguistique. La terminologie graphème/allographe dans les deux domaines, en fait, ne se superpose pas complètement.

Peter Stokes, dans le contexte du projet *DigiPal* (*Digital Resource and Database of Palaeography Manuscript Studies and Diplomatic* http://www.digipal.eu), a récemment abordé de façon systématique ce sujet terminologique du point de vue paléographique, proposant une série de définitions qui ont été ensuite traduites en français par Dominique Stutzmann (2013: §28), ce qui a permis de créer une sorte de vocabulaire bilingue:

> Le modèle prévoit, du côté des signes graphiques, l'idée d'un *alphabet* (en anglais, "alphabet"), composé d'un ensemble de *signes* (en anglais, "grapheme", deux signes différents servant à marquer des sens différents, comme dans "dis" et "bis"), de *caractères* (un même signe, avec des différences graphiques porteuses d'une différence sémantique, comme A majuscule et a minuscule), d'*allographes* (chaque caractère peut se manifester sous différentes graphies canoniques et également valables dans le système graphique), d'*idiographes* (en anglais, "idiograph", image mentale de chaque allographe que peut concevoir un scribe) et de *graphies* (en anglais, "graph", chacune des matérialisation du signe).

Sur le mot *grapheme*, que le paléographe français traduit par 'signe', se joue l'intersection avec la linguistique, où le graphème est la transcription d'un phonème; par conséquent, au lieu de l'unicité du graphème paléographique, par ex. 's' qui s'articule en plusieurs allographes, 's rond' et 's long', en linguistique on peut avoir, dans certains cas, une pluralité de graphèmes, voire des allographes, qui expriment un même phonème, par ex. les graphèmes 'o', 'au' et 'eau', qui sont des allographes du phonème /o/. Évoquer la tripartition de Nina Catach en phonogrammes, morphogrammes et logogrammes suffit à suggérer la complexité de la question.[20]

De mon point de vue, la différence entre les allographes 's rond'/'s long' d'un côté (ou les allographes de 'r' et de 'd') et les allographes 'u'/'v'/'V' - 'i'/'j'/'J', réside dans la coprésence, dans ces derniers, d'un niveau d'articulation graphique et d'un niveau d'articulation phonétique. La balise <g> proposée serait alors parfaitement fonctionnelle pour la description paléographique : 'u', 'v' et 'V' en tant qu'allographes du signe/*grapheme* 'u' (*grapheme* du point de vue paléographique). Mais pour ce qui est de leur nature de lettres ramistes, on aurait en plus besoin d'une deuxième

19. À ce propos, je voudrais ouvrir une discussion concernant la sémantique de la balise <reg> (*regularization*). Qu'est-ce que cela signifie exactement du point de vue éditorial, une régularisation? Pouvons-nous faire une différence entre 'régulariser' et 'moderniser'? Le système graphique médiéval ne connaît pas les signes diacritiques: ajouter un accent aigu pour marquer une *-e* tonique final est-ce une opération de régularisation ou plutôt de modernisation? Par contre, la réduction des anomalies graphiques, comme par ex. la réduction à la forme *ca* des graphies occasionnelles *qua*, peut être interprétée comme une régularisation.

20. Catach 1973 et 1979.

balise exprimant à chaque fois la valeur consonantique, /v/, ou vocalique, /u/, de ces signes, c'est à dire la valeur phonétique du graphème (graphème du point de vue linguistique).

Il s'agirait en conclusion de fournir un parallèle phonétique à la balise <c>, une unité minimale de l'énonciation à côté d'une unité minimale de l'écriture. Je pense par ex. aux difficultés posées par les phonèmes /b/ et /v/ dans la langue espagnole. Ou aux altérations graphiques des textes franco-italiens: l'emploi, par ex., du signe 'ç' dans le ms. I du *Roman de Florimont* (Venezia, Biblioteca nazionale Marciana, fr. Z XXII) couvre plusieurs réalités phonétiques, *borçois/saçant/douç/ça/françois/greçois*, sous l'influence de la *scriptio* d'aire vénitienne.

À défaut de cette, souhaitable, balise TEI, l'encodage *DigiFlorimont* propose alors de décrire la lettre:

1. en l'isolant avec la balise <c>,
2. en déclarant qu'il s'agit d'une lettre ramiste: `@ana="#lram"`,
3. en précisant si elle a une valeur vocalique ou consonantique: `lram-c`, *`lram-v`*.

```
che<c ana="#lram-c">u</c>alier
<c ana="#lram-v">V</c>ne
```

À partir de cet encodage, qui transcrit le niveau diplomatique du texte, on peut générer de façon automatique les deux visualisations: *dipl.* cheualier / Vne; *norm.* chevalier / une.

Tous les signes diacritiques (de 'modernisation' plutôt que de 'régularisation') caractérisant la version normalisée ou interprétative du texte ont été produits avec le même principe: *e-ton* (/e/ tonique) pour e/é; *c-affr* (/ts/) pour c/ç; *hiat* (élément de hiatus) pour ¨, etc.

3.4 *Les interventions massives (2): les abréviations*

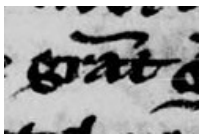
Un élément qui s'associe de façon presque automatique à l'action de transcription et d'édition d'un texte manuscrit médiéval est naturellement l'abréviation. En présence de ce phénomène marquant de l'écriture, on doit se confronter à trois problèmes:

1. signaler l'existence de l'abréviation,
2. signaler la typologie de l'abréviation,
3. signaler les lettres restituées par l'éditeur.

Selon les critères éditoriaux choisis, la réponse graphique à ces problèmes est différente mais elle dépend aussi des limites imposées par le modèle économique de l'édition papier; le *medium* numérique, encore une fois, nous permet de sauvegarder une quantité majeure de données.

Les *Guidelines* TEI proposent à ce sujet une solution d'encodage très verbeux qui dans sa forme la plus complète se présente de cette façon:

```
<choice>
  <abbr><am><g/></am></abbr>
  <expan><ex></ex></expan>
</choice>
```



```
<choice>
  <abbr>gra<am><g ref="#tilde"/></am>t</abbr>
  <expan>gra<ex>n</ex>t</expan>
</choice>
```

Qui plus est, selon les indications des *Guidelines*, les balises <abbr> et <expan> devraient à la rigueur s'appliquer à l'ensemble du mot intéressé par une abréviation, à la fois dans le cas où l'abréviation concerne seulement une lettre, par ex. le tilde de nasalité, et dans le cas où nous avons une abréviation par contraction du mot entier, par ex. *chr*.

3.3.1 Panorama des solutions proposées

BFM/BFM-MSS/GRAAL. L'équipe lyonnaise a élaboré plusieurs solutions au fil des années et des projets. Le manuel d'encodage BFM[21] nous présente ce jeu de balises:

- l'élément <abbr> est utilisé exclusivement pour les abréviations suivies d'un point;
- l'élément <expan> marque la résolution des abréviations et s'applique au mot entier;
- l'élément <ex> signale les caractères ajoutés à la place d'une abréviation:

  ```
  <expan>gra<ex>n</ex>t</expan>
  ```

Le projet BFM-MSS (et par conséquent GRAAL), orienté vers la représentation diplomatique, nous offre un panorama plus complexe et souligne de façon indirecte un 'défaut' potentiel du modèle TEI, qui associe les balises <abbr> et <expan> avec le mot entier tandis qu'il serait préférable de pouvoir les associer plus précisement avec les segments textuels concernés (qui dans certains cas effectivement peuvent coïncider avec le mot: par ex. *mlt/moult*, *chr/chevalier*, etc.). Sur ce point, la pratique de la plupart des projets révèle une 'insubordination' à la règle qui sollicite peut-être une révision de la règle même. Et c'est précisément, par contre, pour la respecter que BFM-MSS propose la balise <bfm:mdvAbbr>:

> Abbréviation médiévale. Cette balise marque le segment graphique "affecté" par l'abréviation. Selon le type d'abréviation, il peut s'étendre de la marque d'abréviation proprement dite à une ou plusieurs lettres associées à la marque, voire à un mot entier. Le fait que les limites de l'abréviation ne coincident pas forcément avec celles d'un mot nous empêche d'utiliser l'élément <abbr> de la TEI.[22]

Cette balise se place au niveau de la représentation facsimilaire, <me:facs>, et contient la balise <am>. La balise <ex>, entourant les lettres restituées dans la résolution, se place au niveau de la représentation diplomatique, <me:dipl>.

```
<w>
  <choice>
```

21. Heiden 2010.
22. Lavrentiev 2008: 20.

```
    <me:facs>gra<bfm:mdvAbbr><am>&tilde;</am></bfm:mdvAbbr>t</me:facs>
    <me:dipl>gra<ex>n</ex>t</me:dipl>
    <me:norm>grant</me:norm>
  </choice>
</w>
```

BHV. Une distinction est faite entre les abréviations par suppression de lettres et les abréviations par brévigraphe: dans le premier cas on utilise un encodage `<choice><abbr><expan>`, dans le deuxième `<choice><orig><reg>` (dans les deux cas les balises font toujours référence à l'ensemble du mot).

```
<choice>
<abbr>Fl.</abbr>
<expan>Florimont</expan>
</choice>

<choice>
<orig>gra¯t</orig>
<reg>grant</reg>
</choice>
```

ESPANNA. Le code juxtapose simplement les deux balises `<am><ex>`.

```
gra<am>¯</am><ex>n</ex>t
```

ORIFLAMMS. Ce projet a repris le modèle TEI dans sa formulation la plus complète et canonique : `<choice><abbr><am><expan><ex>`. Il a aussi introduit la pratique de remplacer le code concernant les abréviations, dans le corps du texte, par des *entities*, par ex. *&ei-tilde-enim;* ou *&con-tironien-v-suscrit-conspectu;*:

> Nous proposons de traiter les abréviations par des entités donnant la prééminence à la forme résolue par des raisons d'ergonomie [...] Dans le cadre d'*Oriflamms* la tendance actuelle se fait jour, avec l'emploi d'entités nommées d'après la substance graphique dont elles rendent compte: par exemple *&e-tilde-est;*. (Stutzmann 2013b: 86)

Mais les dimensions de la liste créée (https://oriflamms.hypotheses.org/1618) montrent qu'à un certain point le système risque peut-être l'implosion.

```
<w>gra&tilde-n;t</w>
<!ENTITY "<choice><expan><ex>n</ex></expan><abbr><am>˜</am></abbr></
choice>">
```

OTINEL. Jean-Baptiste Camps se conforme complètement au système ORIFLAMMS dans la pratique d'utiliser des *entities* tout au long du code et propose un développement très complexe de l'abréviation:

```
&tilde-nasale-an;
  <choice>
    <expan><ex>an</ex></expan>
    <abbr>&#x0303</abbr>
  </choice>
```

&chevalier;

```
<choice>
  <expan>ch<ex>eva</ex>l<ex>ie</ex>er</expan>
  <abbr>ch<c rend="apos">&#x0303;</c>&#xA75B;</abbr>
</choice>
```

TITULUS. Le projet nous offre à ce jour la forme d'encodage la plus complète, dans le même style que le modèle ORIFLAMMS mais en ajoutant l'indication de la typologie d'abréviation. En réalité ce dernier ajout a un précédent, l'encodage de CHARRETTE, dont on parlera ci-dessous. La formule d'encodage TITULUS est donc la suivante:

```
<choice>
  <abbr type="contraction" type="suscription">
  q<am><hi rend="suscription">i</hi></am>
  </abbr>
  <expan>q<ex>u</ex>i</expan>
</choice>
```

Rappelons brièvement le classement des abréviations du projet CHARRETTE, qui avait introduit la balise `<chr_abbr>`:

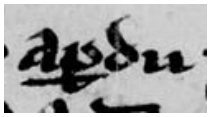
`@type`

`reg`	*regular* (*nasal tildes*, *superscript letters*, etc.)
`frq`	*frequent world* (*chr*, *mlt*)
`pre`	*proper name*
`tit`	*title*
`geo`	*geographical name*
`sac`	*sacred word*

`@class`

`ss`	*special symbol*
`ml`	*modified letter*
`dc`	*diacritic* (*horizontal bar*, etc.)
`sc`	*superscription*
`ct`	*contraction*
`in`	*initial*

Mais en réalité, les *Guidelines* TEI nous proposent une deuxième possibilité en termes d'encodage, beaucoup moins redondante du point de vue des balises et de la répétition des données. Il s'agit de la formule centrée sur l'utilisation de la balise `<g>` *glyph* ("a non-standard character"):



```
<abbr><g ref="#b-per">per</g>du</abbr>
```

ou

```
<abbr><g ref="#b-per"/>du</abbr>
```

Naturellement, ce système comporte un `<charDecl>` où chaque élément d'abréviation est analysé dans le détail, jusqu'à inclure l'image de l'abréviation. À ma connaissance, le seul projet médiéval utilisant un système semblable, dans le principe d'inspiration, à celui que je viens de citer est le projet CHRISTINE DE PISAN, bien que le jeu des balises soit différent, `<abbr> @type`:

```
<abbr type="per">du
```

Il me semble que les propositions de Nicolas Mazziotta (2008) s'insèrent à peu près dans cette ligne, que j'appellerais le 'modèle-glyphe'. Sans entrer dans les détails de sa complexe description de l'abréviation,[23] je viens à ses conclusions concernant l'encodage:

> Nous proposons donc: 1/ de nous débarrasser simplement des éléments *choice*, *expan* et *abbr*, qui deviennent redondants du fait de la mise en relation explicite des éléments *am* et *ex*, 2/ de créer un attribut spécialisé dans la mise en correspondance de ces deux éléments; il apparaîtrait par exemple dans l'élément *ex*, porterait le nom *@am* et contiendrait un URI. (Mazziotta 2008: § 78)

Voici des exemples:

chrs

```
ch<ex am="#a1">eva</ex>l<ex am="#a1">ie</ex>r
<am xml:id="a1"><g ref="#titulus"/></am>s
```

conqⁱs

```
conq<ex am="#a1">u</ex><c xml:id="#c1">i</c>
<am xml:id="a1">
  <rendition ref="c1">superscript</rendition>
</am>s
```

Le risque évident est de passer d'une complexité redondante à une autre typologie de complexité, surtout en considération de la fréquence de ce phénomène qui peut devenir quelques fois le cœur de l'encodage. Mais le principe qui me semble à retenir est sûrement celui de la mise en relation, à travers un attribut, du signe abréviatif avec l'abréviation résolue, sans dédoubler l'information. Je trouve que le style d'encodage TEI 'modèle-glyphe' se fonde exactement sur ce principe. On pourrait alors proposer cette formule très élégante:

```
<g type="abbr" ref="#chr"/>
<g type="abbr" ref="#tilde-nas"/>
```

23. "Nous donnons aux unités discrètes de la langue écrite le nom de *grammène*, classe que nous pouvons subdiviser en deux sous-classes: 1/ celle des *caténogrammes*, comprenant les grammèmes non signifiants organisés suivant le même axe que les unités signifiantes, en *chaîne*; 2/ celle des *logogrammes*, grammèmes signifiants dont les parties constitutives ne sont pas forcément organisées suivant le même axe que celui où s'enchaînent les unités signifiantes. [...] La relation d'abréviation lie un caténogramme, significatif, à un logogramme synonyme. Nous pouvons qualifier le rapport paradigmatique par le verbe *porter*. Nous dirons que la *portée* des logogrammes impliqués dans un mécanisme d'abréviation consiste en un caténogramme. [...] Il est important de noter que la notion d'*abréviation* devient dès lors relationnelle: une abréviation s'envisage toujours *par rapport* à une autre unité qui n'est pas abrégée. D'autre part, la relation de *portée* est potentielle et non obligatoire". (Mazziotta 2008: § 20-22)

Mais dans le contexte des langues romanes le même signe abréviatif peut avoir des résolutions différentes, par ex. *n* ou *m* dans le cas du tilde de nasalité, *con* ou *com* ou *cun* pour la note tironienne, d'où la nécessité d'indiquer la résolution contextuelle. Les exemples TEI cités montrent que la balise <g> peut contenir sa 'traduction' alphabétique; on aura donc, par ex.:

```
mai<g type="abbr" ref="#tilde-nas">n</g>tes
```

Toutefois, à l'état actuel des règles d'emboîtement des *Guidelines*, la balise <g> ne peut pas contenir d'autres balises. Par conséquent, l'encodage suivant ne serait pas conforme à la TEI:

```
<g type="abbr" ref="#ctr-chr">ch<ex>evalie</ex>r</g>
```

Mais face à ce problème du contenu potentiel de <g> la question devient plus théorique et elle requiert une réflexion approfondie et une décision partagée: est-ce qu'il est sémantiquement approprié d'autoriser un contenu pour la balise <g>?

3.3.2 La proposition *DigiFlorimont*

Pour le moment, avec les balises, les attributs et les règles TEI disponibles, j'ai donc proposé une solution qui considère le problème en inversant les termes de la question: je ne déclare pas qu'un glyphe abréviatif correspond à l'expansion qu'il contient, mais, suivant l'exemple de Mazziotta, je déclare qu'une expansion correspond à un glyphe abréviatif déterminé, et en même temps j'entoure avec la balise <expan> la seule expansion de l'abréviation et pas tout le mot:[24]

```
chr
  <expan corresp="#abr-ctr-chr">ch<ex>evalie</ex>r</expan>
pi
  <expan corresp="#abr-sus-pi">p<ex>r</ex>i</expan>mier
mai~tes
   mai<expan corresp="#abr-tilde-nas"><ex>n</ex></expan>tes
```

Tous les signes abréviatifs sont décrits dans le <charDecl> et associés avec leur image, ce qui a permis, dans la version diplomatique, de faire apparaître directement l'image de l'abréviation en passant avec la souris sur le texte: une façon de fondre édition ultradiplomatique ou facsimilaire, diplomatique et semi-diplomatique sans multiplier les visualisations. L'attribut @corresp a été utilisé à défaut de la disponibilité de l'attribut @ref pour la balise <expan>.[25]

Toute proposition, surtout si elle est le fruit d'une réflexion individuelle, est contestable, mais il me semble que la variété des solutions proposées par chaque projet est objectivement le signal

24. Il s'agit d'une pratique courante, partagée par ex. par Jean-Baptiste Camps et Alexei Lavrentiev.

25. Le choix de l'attribut @corresp a été fait malgré les perplexités avancées par Mazziotta (2008: §76-77) et que je partage partiellement, mais il s'agissait du seul attribut à disposition si l'on voulait rester dans le périmètre de la TEI: "En outre, si la sémantique de l'élément *choice* est relativement claire, celle de l'attribut @corresp est on ne peut plus floue. On peut se poser la question de la transparence de l'encodage. Comme la TEI ne propose pas une ontologie à strictement parler, la permissivité et la flexibilité extrême de certaines recommandations sont susceptibles de poser des grandes difficultés d'échange de données. Comment savoir si l'attribut @corresp a la même valeur d'un document à l'autre, voire d'une occurrence à l'autre? Un humain aura tôt fait de comprendre la relation, mais que dire d'un automate?".

d'une insatisfaction latente concernant cette question, qui appelle une solution plus ergonomique pour tous.

4 Un élément de la mise en texte: l'initiale

Les noms (et la valeur sémantique) des balises TEI décrivant les éléments de la structuration visuelle d'un document écrit en tant que traduction physique de la structuration logique du texte (au delà de l'écriture) laissent entrevoir clairement leur origine enracinée dans la structure du livre imprimé: par ex., l'utilisation du mot 'paragraphe' pour désigner les blocs textuels dans un manuscrit médiéval en prose (quand ces blocs existent) est-elle correcte?

Mais ce que je voudrais souligner ici est le fait que les *Guidelines* ne contemplent pas du tout un élément dont on connaît, par contre, très bien l'importance, en plus de son évidence physique, quelques fois très évidente: l'initiale.

La définition du *Dictionnaire codicologique* de Muzerelle explique synthétiquement sa valeur: "Lettre commençant un ouvrage, une partie du texte, un paragraphe [...] mise en valeur par un procédé quelconque (enluminure, etc.)". Mais à côté de ce type d'initiales nous avons aussi celle qui est définie "initiale secondaire": "Initiale incluse dans le corps du texte, sans alinéa", donc une initiale qui n'ouvre pas une section nouvelle mais qui signale quand même quelque chose de significatif qui 'arrive' au texte à ce point-là. À travers la modulation de la décoration on assiste enfin à la création d'un système hiérarchique de relations entre les sections du document, et donc d'articulation du texte.[26] Parfois sa fonction peut se réduire au rang de simple élément décoratif au début du texte, et on pourrait alors l'encoder en tant qu'élément de la décoration. Mais dans la majorité des occurrences la dimension esthétique procède à côté de la dimension fonctionnelle (à la structuration). Il est donc évident qu'on ne peut pas ignorer cet élément au niveau de l'encodage, mais de quelle façon l'encoder?


début de section


début de lasse


début de vers

Le problème a été encore une fois identifié, et des balises, encore une fois personnalisées, ont été proposées:

- BFM-MSS: `<bfm:lettrine> @color @colorSuppl @size @decoration @ref`
- MENOTA: `<c> @type (littNot/initial) @rend`
- FROISSART: `<hi> @rend (illuminate initial/campi/color-stroked red/pen-flourished initial`, etc.)
- ORIFLAMMS: `<hi> @rend (initiale)` **ou** `<g> @type (initiale)`
- OTINEL: `<hi> @rend (initiale)`

26. Bastlein 1991, Cavallo 1994, Le Saux 2004, Fuksas 2014, Géhin 2017: 149-151.

On peut voir clairement que la dimension physique a été la seule dimension observée et décrite. Pour ne pas proposer une balise personnalisée, l'encodage *DigiFlorimont* s'est conformé à la pratique courante d'indiquer les initiales à travers la balise `<hi>` et de les classer ensuite à travers les valeurs associées à l'attribut `@rend`, suivant le modèle FROISSART: une solution plutôt anonyme et passe-partout, pas du tout satisfaisante et plutôt discutable sur le plan sémantique du moment que l'initiale médiévale est loin d'être une simple 'mise en évidence'. C'est un peu comme si le titre d'un roman était encodé de cette façon : `<hi rend="bold-italic">`Les Misérables`</hi>`.

5 *Et caetera*

Les éléments que l'on vient de discuter ne représentent qu'une partie des défis lancés au système TEI par la textualité médiévale littéraire. Pour en rester au niveau du document, nous n'avons pas un système de balises bien développé pour décrire de façon satisfaisante le 'peritexte': instruction pour les illuminateurs, notes de lecture, gloses marginales et interlinéaires, etc. Officiellement, les *Guidelines* nous offrent exclusivement la balise `<note>`, qui est chargée d'exprimer, seule, l'encodage de cette réalité complexe et multiforme, avec le seul appui de l'attribut `@type`. Mais surtout, ce qui me semble très insatisfaisant pour notre contexte spécifique est le fait que, selon la perspective officielle TEI, cette même balise est proposée pour encoder aussi bien les notes pertinentes au document que les notes dérivant de l'activité éditoriale. La généricité extrême du contenu autorisé pour cette balise – "A note is any additional comment found in a text, marked in some way as being out of the main textual stream" – est donc accentuée par les corollaires successifs: "Notes may be in a different hand or typeface, may be authorial or editorial, and may have been added later" (*Guidelines*: 3.8.1 Notes). Pour remonter du document au niveau du texte, la discussion concernant par ex. l'encodage des variantes est bien loin d'être archivée et d'être parvenue à des résultats satisfaisants, aussi bien en termes de code que d'interface, d'écriture et de visualisation.

Pour conclure, la réflexion que je voulais proposer concerne la durabilité de la généricité du standard TEI, celle généricité qui signifie universalité et adaptabilité extrêmes, bien exprimée par les premières phrases de l'Introduction aux *Guidelines*: "These *Guidelines* apply to texts in any natural langage, of any date, in any literary genre or text type, without restriction on form or content".

Mais l'écriture médiévale, avec son alphabet latin peuplé d'allographes, de lettres ramistes et d'abréviations, est un système sémiotique bien différent de celui par ex. des lettres autographes de Van Gogh; par conséquent, les modalités de la traduction sémiotique ainsi que ses enjeux sont très différents. À mon avis, la voie la plus fructueuse pour une application de la TEI sans compromis est donc celle inaugurée par les communautés EpiDoc (*Epigraphic Documents in TEI XML* https://sourceforge.net/p/epidoc/wiki/Home/) ou MEI (*Music Encoding Initative* https://music-encoding.org), c'est-à-dire des sous-communautés TEI 'contextualisées'; ce qui signifie aussi des sous-standards beaucoup plus formalisés et rigides garantissant l'homogénéité de l'encodage au sein de la même communauté textuelle.

Au moment où l'encodage WYSIWYM[27] ne propose pas un système de balises orienté vers la traduction typographique mais, à l'opposé, un système orienté vers l'analyse (d'où la différence

27. L'anagramme WYSIWYM (*What You See Is What You Mean*, 'Ce que vous voyez est c'est que vous voulez dire') fait référence aux systèmes d'écriture, par ex. le langage de markup XML-TEI ou le langage de mise en page LaTex, qui séparent le fond de la forme dans la création des documents; en d'autre termes, ces langages séparent la donnée

profonde entre le standard XML-TEI et LaTex), la généricité, évidemment, peut se transformer en limite. Au contraire, l'analyse, ou au moins l'analyse capable de fournir des résultats significatifs par rapport à un objet déterminé, a besoin nécessairement de s'appuyer sur des données spécifiques, bien individualisées, voire spécialisées. C'est pour cette raison que l'expérience du projet EpiDoc me semble une des interprétations les plus réussies et efficaces des principes d'encodage exprimés par la TEI: le texte épigraphique a ses règles, ses formes et ses dynamiques qui, au delà du trait commun sémiotique représenté par le fait d'être des objets textuels verbaux, n'ont (presque) rien à partager avec les règles, les formes et les dynamiques d'un manuscrit du XIII^{e} siècle ou avec les pages des brouillons de Proust. De cette façon la rigidité du standard, très pointu et largement partagé par une communauté spécifique, s'impose en tant que laissez-passer pour la 'libre circulation' des données de la recherche.

En conclusion, je cite directement les mots de présentation du projet MEI, qui, de mon point de vue, définissent un véritable modèle méthodologique:

> The *Music Encoding Initiative* (MEI) is a community-driven effort to define a system for encoding musical documents in a machine-readable structure. MEI brings together specialists from various music research communities, including technologists, librarians, historians, and theorists in a common effort to define best practices for representing a broad range of musical documents and structures. The results of these discussions are formalized in the MEI schema, a core set of rules for recording physical and intellectual characteristics of music notation document expressed as an *eXtensible Markup Language* (XML) schema. It is complemented by the MEI *Guidelines*, which provide detailed explanations of the components of the MEI model ad best practices suggestions.

textuelle, en tant que donnée, et sa représentation. Par contre, l'anagramme WYSIWYG (*What You See Is What You Get*, 'Ce que vous voyez est ce que vous obtenez'), fait référence aux systèmes, par ex. le text-editor Microsoft Word, qui permettent à l'utilisateur de créer directement à l'écran la mise en page du texte en melangeant dans la fichier-source les données textuelles et les instructions graphiques.

6 Ouvrages cités

André, Jacques; Jimenes, Rémi. 2013. 'Transcription et codage des imprimés de la Renaissance: Réflexions pour un inventaire des caractères anciens', *Document Numérique*, 16.3: 113-39 <https://go.uv.es/6jmKn24>

Andrieux-Reix, Nelly; Monsonégo, Simone. 1997. 'Écrire des phrases au Moyen Age: Matériaux et premières réflexions pour une étude des segments graphiques observés dans des manuscrits français médiévaux', *Romania*, 115.459-460: 289-336 <https://www.persee.fr/doc/roma_0035-8029_1997_num_115_459_2243>

Andrieux-Reix, Nelly; Monsonégo, Simone. 1998. 'Les Unités graphiques du français médiéval: Mots et syntagmes, des représentations mouvantes et problématiques', *Langue Française*, 119 (= *Segments graphiques du français: Pratiques et normalisation dans l'histoire*): 30-51 <https://doi.org/10.3406/lfr.1998.6258>

Andrieux-Reix, Nelly; Monsonégo, Simone. 2000. 'Transcription, lisibilité, transgression: Quelques problèmes posés par les éditions de textes médiévaux', in Claude Buridant (ed.), *Le Moyen français: Le Traitement du texte (édition, apparat critique, glossaire, traitement électronique), actes du IX^e colloque international sur le moyen français* (Strasbourg: Presses Universitaires de Strasbourg), pp. 55-63

Anis, Jacques. 1983. 'Vilisibilité du texte poétique', *Langue Française*, 59: 88-102 <https://doi.org/10.3406/lfr.1983.5167>

Baddeley, Susan; Biedermann-Pasques, Liselotte. 2004 'Histoires des systèmes graphiques du français à travers des manuscrits et des incunables (IX^e-XV^e siècle): Segmentation graphique et faits de la langue', *Revue de Linguistique Romane*, 68: 181-201

Baker, A. T. 1937. 'Le Futur des verbes avoir et savoir', *Romania*, 63.249: 1-30 <https://doi.org/10.3406/roma.1937.3825>

Balon, Laurent. 2016. 'Les Unités graphiques de l'ancien et du moyen français: Un lieu d'observation privilégié pour une meilleure approche des phénomènes de lexicalisation et de grammaticalisation du français?', *Les Dossiers d'Histoire, Épistémologie, Langage*, 9 (= *Écriture(s) et représentations du language et des langues*): 304-16 <https://hal.archives-ouvertes.fr/hal-01305404>

BHV. 2008. *Manuel d'encodage XML-TEI Renaissance et temps modernes: Imprimés-Manuscrits*, ed. by Nicole Dufournaud *et al.*, latest version 17/12/2019 <http://www.bvh.univ-tours.fr/XML-TEI/ManuelWeb/Manuel_TEI_BVH.html> [accessed: 24-03-2021]

Camps, Jean-Baptiste. 2016. *La 'Chanson d'Otinel': Édition complète du corpus manuscrit et prolégomènes à l'édition critique – digital appendices*, unpublished Ph.D. dissertation, Paris-Sorbonne <https://doi.org/10.5281/zenodo.1116735>

Camps, Jean-Bapiste, 'La *Chanson d'Otinel*: Édition complète du corpus manuscrit et prolégomènes à l'édition critique' [abstract of Camps 2016], *Perspectives Médiévales*, 38, <https://doi.org/10.4000/peme.13004>

Catach, Nina. 1973. 'Que Faut-il Entendre par Système graphique du français?', *Langue Française*, 20 (= *L'Ortographe*): 30-44 <https://doi.org/10.3406/lfr.1973.5652> [accessed: 24-03-2021]

Catach, Nina. 1979. 'Le Graphème', *Pratiques: Linguistiques, Littérature, Didactique*, 25 (= *Ortographe*): 21-32 <https://doi.org/10.3406/prati.1979.1122>

Catach, Nina. 1998. 'Les signes graphiques du mot à travers l'histoire', *Langue Française*, 119 (= *Segments graphiques du français*): 10-23 <https://doi.org/10.3406/lfr.1998.6256>

Cavallo, Guglielmo. 1996. 'Iniziali, scritture distintive, fregi. Morfologia e funzioni', in *Libri e documenti d'Italia: dai Longobardi alla rinascita delle città, atti del convegno nazionale dell'Associazione Italiana di Paleografi e Diplomatisti, Cividale del Friuli 1994*, ed. by Cesare Scalon (Udine: Forum), pp. 15-33

Cazal, Yvonne; Parussa, Gabriella; Pignatelli, Cinzia *et al.* (éds.). 2003. 'L'orthographe: du manuscrit médiéval à la linguistique moderne', *Médiévales*, 45 (= *Grammaire du vulgaire*): 99-108 <http://medievales.revues.org/969>

Conseils. 2014. *Conseils pour l'édition des textes médiévaux*, Fascicule I, *Conseils généraux*, ed. by François Vieilliard et Olivier Guyotjeannin (Paris: Éditions du CTHS et École Nationale des Chartes)

Croenen, Godfried; Romanova, Natasha. 2010. *The Online Froissart Project: Manual for Transcription and Markup,* Version 1.2, July 2010 <http://pcwww.liv.ac.uk/~gcroenen/Guidelines.pdf> [accessed: 24-03-2021]

D'Iorio, Paolo. 2010. 'Qu'est-ce qu'une édition génétique numérique', *Genesis*, 30: 49-53 <http://doi.org/10.4000/genesis.116>

Duval, Frédéric. 2012. 'Transcrire le Français médiéval: de l'*Instruction* de Paul Meyer à la description linguistique contemporaine', *Bibliothèque de l'École des Chartes*, 170.2: 321-42 <https://doi.org/10.3406/bec.2012.464252>

Foulet, Alfred; Speer, Mary Blakely. 1979. *On Editing Old French Texts* (Lawrence: The Regents Press of Kansas)

Fuksas, Anatole Pierre. 2014. 'The *Divisio Operis* of Chrétien's Romances and the Paratextual System of the Guiot Manuscript (Paris, BNF, fr. 794)', *Segno e Testo*, 12: 309-25 <https://go.uv.es/7YQ2B5g>

Hansen, Thomas. 2012. 'TEI - Keeping It Simple', *Digital Medievalist,* 7 <http://doi.org/10.16995/dm.40> [accessed: 24-03-2021]

Hasenhor, Géneviève. 1998. 'Abréviations et frontières de mots', *Langue Française*, 119 (= *Segments graphiques du français*): 24-29 <https://doi.org/10.3406/lfr.1998.6257>

Haugen, Odd Einar (ed.). 2019, *The Menota Handbook: Guidelines for the Electronic Encoding of Medieval Nordic Primary Sources,* Version 3.0 (Bergen: Medieval Nordic Text Archive) <http://www.menota.org/handbook.xml> [accessed: 24-03-2021]

Heiden, Serge; Guillot, Céline; Lavrentiev, Alexei *et al.* (éds.). 2010. *Manuel d'encodage XML-TEI des textes de la Base de Français Médiéval,* Projet BFM, UMR5191 ICAR, CNRS/ENS de Lyon, Version 4.0 (Aout 2010) <http://bfm.ens-lyon.fr/IMG/pdf/Manuel_Encodage_TEI.pdf> [accessed: 24-03-2021]

Hilka, Alfons (ed.). 1933. Aimon von Varennes, '*Florimont', ein altfranzösischer Abenteuerroman zum ersten Mal mit Einleitung, Anmerkungen, Namenverzeichnis und Glossar unter Benützung der von Alfred Risop gesammelten handschriftlichen Materialien* (Halle: Niemeyer)

Huitfeldt, Claus; Sperberg-McQueen, C. Michael. 2008. 'What is Transcription?', *Literary and Linguistic Computing*, 23.3: 295-310 <https://doi.org/10.1093/llc/fqn013>

Huitfeldt, Claus; Sperberg-McQueen, C. Michael. 2017. 'Transcriptional Implicature: Using a Transcript to Reason About an Exemplar', *Abstracts of 12th Annual International Conference of the Alliance of Digital Humanities Organizations (Montréal, 2017)* <https://dh2017.adho.org/abstracts/235/235.pdf> [accessed: 24-03-2021]

Huitfeldt, Claus; Sperberg-McQueen, C. Michael. 2018. 'Interpreting Difference Among Transcripts', Abstracts of *13th Annual International Conference of the Alliance of Digital Humanities Organizations (Mexico City, June 26-29, 2018)*: 287-91 <https://dh2018.adho.org/en/interpreting-difference-among-transcripts/> [accessed: 24-03-2021]

Lavrentiev, Alexei. 2002a. *Proposal for XML markup of Old French Text Corpora* <http://www.princeton.edu/~lancelot/ss/materials.shtml> [accessed: 24-03-2021]

Lavrentiev, Alexei. 2002b. *Charrette Project and XML* <http://www.princeton.edu/~lancelot/ss/materials.shtml> [accessed: 24-03-2021]

Lavrentiev, Alexei. 2008. *Manuel d'encodage XML-TEI étendu des transcriptions de manuscrits dans le projet BFM-Manuscrits,* CNRS/ENS-LSH, UMR 5191 ICAR, Version 2.1 (Juin 2008) <http://ccfm.ens-lyon.fr/IMG/pdf/BFM-Mss_Encodage-XML.pdf> [accessed: 24-03-2021]

Lavrentiev, Alexei (ed.). 2017. *Systèmes graphiques de manuscrits médiévaux et incunables français: Ponctuation, segmentation, graphies, actes de la Journée d'étude de Lyon, ENS LSH, 2005* (Chambéry: Université de Savoie)

Lavrentiev, Alexei. 2019. 'Traitement de la ponctuation dans les éditions de textes en français médiéval: Normes, introductions, pratiques', in *Les Introductions linguistiques aux éditions de textes*, ed. by Frédéric Duval, Céline Guillont-Barbance and Fabio Zinelli, (Paris: Classiques Garnier), pp. 275-94

Lavrentiev, Alexei; Stutzmann, Dominique. 2015. *Encoding Allographs: Using the <g> Element*, presentation within *Connect, Animate, Innovate: TEI Conference and Members Meeting, 2015 Lyon* <https://halshs.archives-ouvertes.fr/halshs-01318710> [accessed: 24-03-2021]

Lavrentiev, Alexei; Leydier, Yann; Stutzmann, Dominique. 2015. 'Specifying a TEI-XML Based Format for Aligning Texts to Image at Character Level', *Proceedings of the Symposium Cultural Heritage Markup*, 16 <https://halshs.archives-ouvertes.fr/halshs-01318701/document> [accessed: 24-03-2021]

Lavrentiev, Alexei; Leydier, Yann; Stutzmann, Dominique. 2016. 'Spécification du format XML-TEI pour l'alignement texte-image, 2: Bonnes Pratiques d'encodage', in *Écriture médiévale & numérique*, 12-09-2016 <https://oriflamms.hypotheses.org/1510> [accessed: 24-03-2021]

Le Saux, Françoise. 2004. 'On Capitalization in Some Early Manuscripts of Wace's *Roman de Brut*', *Arthurian Studies in Honour of P.J.C. Field*, ed. by Bonnie Wheeler (Woodbridge: Brewer), pp. 29-47

LibGloss. 2011. 'Encodage du *Liber glossarum*' <http://liber-glossarum.huma-num.fr/encodage.html> [accessed: 24-03-2021]

Llamas Pombo, Elena. 2001. 'La Construction visuelle de la parole dans le livre médiéval', *Diogène*, 196.4: 40-52 <https://doi.org/10.3917/dio.196.0040> [accessed: 24-03-2021]

Llamas Pombo, Elena. 2009. 'Variación gráfica y secuenciación de la palabra en manuscritos medievales hispánicos', in *Los códices literarios de la Edad Media: interpretación, historia, técnicas y catalogación*, ed. by Pedro Cátedra, Eva Belén Carro Carbajal and Javier Durán Barceló (San Millán de la Cogolla: Cilengua), pp. 225-57

Llamas Pombo, Elena. 2016. 'Ponctuation médiévale, pragmatique et énonciation: Lire l'*Ovide moralisé* au XIV^e^ siècle', *Linx*, 73 (= *Énonciation et marques d'oralité dans l'évolution du français*) <https://doi.org/10.4000/linx.1638> [accessed: 24-03-2021]

Llamas Pombo, Elena. 2017. 'Graphie et ponctuation du français médiéval: Système et variation', in *Enregistrer la Parole, écrire la langue dans la diachronie du français*, ed. by Gabirella Parussa, Maria Colombo Timelli and Elena Llamas Pombo (Tübingen: Günter Narr), pp. 39-88

Marchello-Nizia, Christiane (ed.). 2013. *Queste del saint Graal: Manuscrit Lyon BM P.A. 77*, dernière revision 19-07-2013, avec la collaboration d'Alexei Lavrentiev <http://txm.ish-lyon.cnrs.fr/bfm/pdf/qgraal_cm_2013-07.pdf> [accessed: 24-03-2021]

Materni, Marta (ed.). 2020a. *Autour du 'Roman de Florimont': Approches multidisciplinaires à la complexité textuelle médiévale*, Quaderni di Francigena, 2 (Padova: Università degli Studi di Padova) <http://doi.org/10.25430/2724-0975/2>

Materni, Marta. 2020b. '*DigiFlorimont*: une occasion de réflexion philologique et numérique autour de la représentation de la complexité textuelle médiévale', in Materni 2020a.

Materni, Marta. 2020c. 'Complessità della codifica ed ergonomia strumentale nel contesto XML-TEI: dove siamo?', *Umanistica Digitale*, 8: 123-43 <http://doi.org/10.6092/issn.2532-8816/9976>

Mazziotta, Nicolas. 2004. 'Le Texte dans tous ses états: Philosophie d'encodage du projet Khartês', in *Le Poids des mots, actes des 7^es^ JADT, Journées internationales d'analyse statistique des données textuelles, Louvain-la-Neuve, 2004*, ed. by Anne Dister, Cédrick Fairon and Gérald Purnelle, (Louvain: Presses Universitaires de Louvain), pp. 793-803 <http://hdl.handle.net/2268/15850>

Mazziotta, Nicolas. 2006. 'Propositions pour l'édition des sources primaires CCDM', in *Actes des 2^es^ Journées du Consortium pour les Corpus du Français Médiéval, 2006, ENS Lyon* <http://ccfm.ens-lsh.fr/IMG/pdf/CCFM-codage-manuscrits_ULG.pdf> [accessed: 24-03-2021]

Mazziotta, Nicolas. 2008. 'Traiter les Abréviations du français médiéval: Théorie de l'écriture et pratique d'encodage', *Corpus*, 7 <http://journals.openedition.org/corpus/1517> [accessed: 24-03-2021]

Ménard, Philippe. 1990. 'Problèmes de paléographie et de philologie dans l'édition des textes français du Moyen Age', in *The Editor and the Text: In Honour of Professor Anthony J. Holden*, ed. by Philip E. Bennett, Graham A. Runnalis and Anthony J. Holden, (Edinburgh: Edinburgh University Press), pp. 1-10

Meyer, Paul. 1910. 'Instructions pour la publication des anciens textes français', *Bibliothèque de l'École des Chartes*, 71 (1910): 224-33 [reprint from 1909 *Bulletin de la Société des Anciens Textes Français*, 35.1: 64-79] <https://doi.org/10.3406/bec.1910.460996>

Pierazzo, Elena. 2014. 'Digital Documentary Editions and the Others', *Scholarly Editing: The Annual of the Association for Documentary Editing*, 35: 1-23 <http://www.scholarlyediting.org/2014/essays/essay.pierazzo.html> [accessed: 24-03-2021]

PIZAN. 2010. Christine de Pizan 'The Making of the Queen's Manuscript (British Library, Harley MS 4431): Readers' Guide', 06-07-2010 <http://www.pizan.lib.ed.ac.uk/xml.html> [accessed: 24-03-2021]

Renedo Mirambell, Clara. 2018. 'Manuel d'encodage pour les notices épigraphiques du projet TITULUS conforme à la TEI P5 et à EpiDoc 8.22, Version 1.1', 04-02-2018 version <http://titulus.huma-num.fr/exist/apps/titulus/resources/manuel-encodage-TITULUS_1.1.pdf> [accessed: 24-03-2021]

Rickard, Peter. 1982. 'Système ou arbitraire? Quelques réflexions sur la soudure des mots dans les manuscrits français du Moyen Age', *Romania*, 103.4: 470-512 <https://doi.org/10.3406/roma.1982.2125>

Robinson, Peter; Solopova, Elizabeth. 1993. 'Guidelines for Transcription of the Manuscripts of the *Wife of Bath*'s Prologue', in *The 'Canterbury Tales' Project Occasional Papers*, ed. by Norman Blake and Peter Robinson (Oxford: Office for Humanities Communication Publications), I: 19-52 <http://canterburytalesproject.com/pubs/transguide-MI.pdf>

Roques, Mario. 1926. 'Établissement de règles pratiques pour l'édition des anciens textes français et provençaux', *Romania*, 52: 243-49 <http://www.persee.fr/doc/roma_0035-8029_1926_num_52_205_6906>

SIG-LING. 2017. 'att.linguistic for <w> and <pc>', ticket #1670 open on 22-07-2017, in *Text Encoding Initiative Repository* <https://github.com/TEIC/TEI/issues/1670> [accessed: 24-03-2021]

Stutzmann, Dominique. 2011. 'Paléographie statistique pour décrire, identifier, dater... Normaliser pour coopérer et aller plus loin?', in *Kodikologie und Paläographie im digitalen Zeitalter*, 2 (= *Codicology and Palaeography in the Digital Age*, 2) (Norderstedt: Books on Demand), pp. 247-77 <http://kups.ub.uni-koeln.de/4337/>

Stutzmann, Dominique. 2013a. 'Paléographie latine et vernaculaire (livres et documents): Conférences de l'année 2011-2012', *Annuaire de l'École Pratique des Hautes Études - Sciences Historiques et Philologiques*, 144: 115-28 <http://journals.openedition.org/ashp/1485> [accessed: 24-03-2021]

Stutzmann, Dominique. 2013b. 'Ontologie des formes et encodage des textes manuscrits médiévaux: Le projet ORIFLAMMS', *Document Numérique*, 16.3: 81-95 <https://doi.org/10.3166/dn.16.3.81-95>

TEI Customization. <https://tei-c.org/guidelines/customization/> [accessed: 24-03-2021]

Testenoire, Pierre-Yves. 2017. 'Transcrire des Écrits scolaires: Entre Philologie et génétique textuelle', *Corpus*, 16 <https://doi.org/10.4000/corpus.2762> [accessed: 24-03-2021]