Le "modèle abstrait" du corpus Bouvard : première approche

Stéphanie Dord-Crouslé, Emmanuelle Morlock-Gerstenkorn

Le « modèle abstrait » du corpus Bouvard - première approche

Journée d'études « Constitution et exploitation de corpus issus de manuscrits » – Grenoble – 12 mars 2009

Stéphanie Dord-Crouslé et Emmanuelle Morlock-Gerstenkorn

Introduction

Notre intervention porte sur le nécessaire travail de conceptualisation et de modélisation, préalable à l'encodage TEI du corpus Bouvard[1]. Elle vise également à décrire une démarche progressive d'appropriation de la TEI. En effet, comme pour tous les projets de ce type, l'encodage doit être formalisé. Mais il n'est ni possible ni souhaitable de tout décrire ; il faut d'abord identifier les traits importants, puis la manière et la précision avec laquelle on souhaite utiliser les éléments proposés par la TEI, en fonction des objectifs et des contraintes du projet.

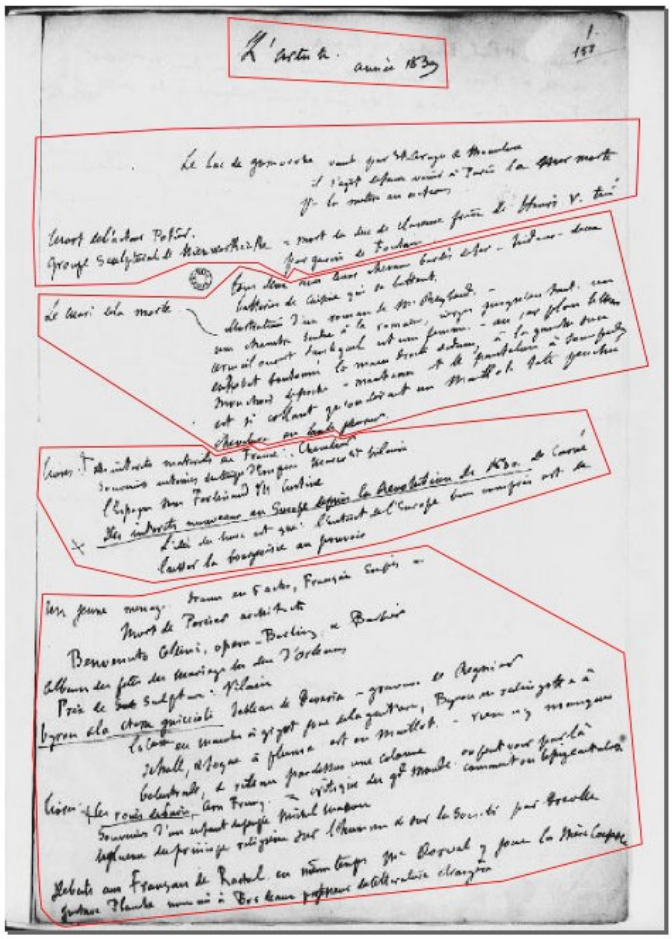
Or, dans le projet Bouvard, la transcription et l'indexation du corpus sont réalisées en deux temps distincts. À l'heure actuelle, l'indexation consiste à associer à chaque page de manuscrit, dans la base de données relationnelle du site de travail, diverses métadonnées (comme des classements typologiques et chronologiques, le nom des scripteurs, des relations intertextuelles, etc.). Plus tard seulement, l'intégralité du corpus transcrit sera balisé en TEI. De plus, des « régions d'intérêt » - on parlera également de « fragments » - seront délimitées visuellement sur l'image du manuscrit et dans le flux de la transcription TEI (voir la figure 1).

1 Voir la présentation du projet sur le site : http://dossiers-flaubert.ish-lyon.cnrs.fr/. Les manuscrits concernés sont conservés à la Bibliothèque municipale de Rouen sous la cote g 226 volumes 1 à 8.

Régions-fragments

Folio g226 vol. 1 f° 158 recto

En rouge : délimitation de zones polygonales dans l'image pour récupérer des fragments (outil Inkscape)



(Figure 1)

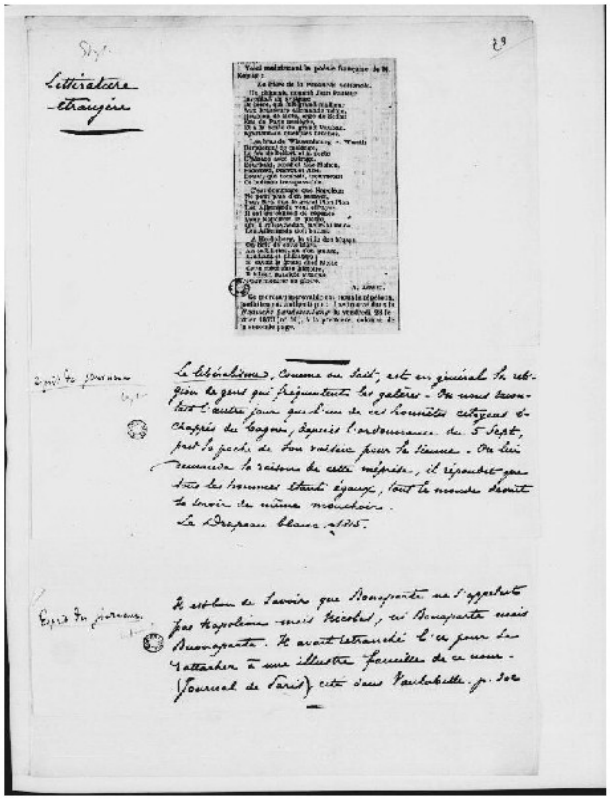
Entre mai et août 2008, un premier guide d'encodage TEI du corpus (comprenant un choix de balises, une liste d'attributs et une structure globale) a été rédigé dans le cadre d'un stage[2]. Fin 2008, la réalisation des premiers tests d'encodage nous a confrontées à un certain nombre de difficultés et de questions qui ont rendu nécessaire l'approfondissement de ce travail :

- Comment encoder en TEI la disposition topologique des composants de mises en page, sachant que celles-ci sont diverses ?
- Comment prendre en compte les chevauchements d'éléments textuels (et donc logiques) d'une page sur l'autre ?
- Comment traiter les pages qui se présentent tête-bêche à la lecture du volume[3] ?

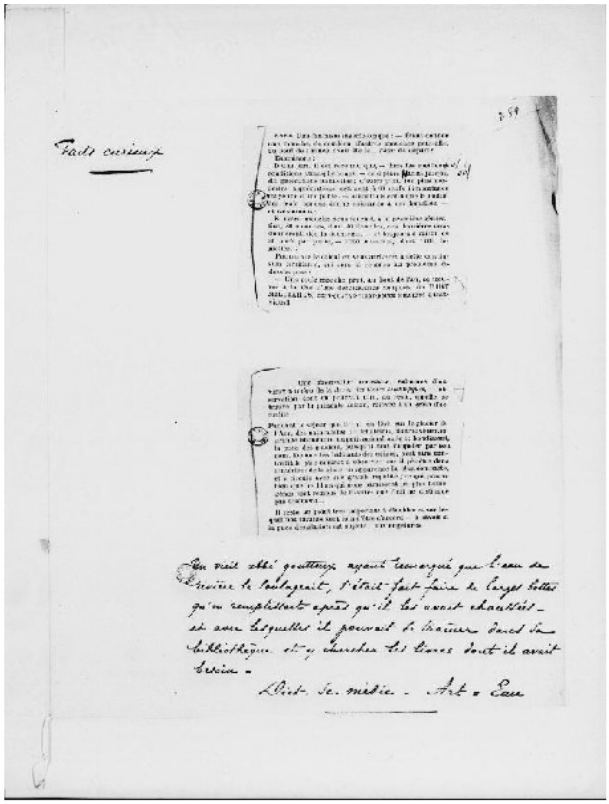
2 Vanessa Le Rolle, *Les Dossiers préparatoires de* Bouvard et Pécuchet *de Flaubert - Approche critique d'une structuration du corpus avec la TEI (Text Encoding Initiative)*, Rapport de stage de Master 2 professionnel « Patrimoine écrit : histoire et pratiques de l'édition », Université François Rabelais de Tours et Institut des Sciences de l'Homme de Lyon, 2008. Ce rapport de stage a permis la rédaction ultérieure, toujours par Vanessa Le Rolle, d'un *Manuel d'encodage XML/TEI des Dossiers préparatoires de* Bouvard et Pécuchet *de Flaubert* (document interne en trois volets), version 1, août 2008.

3 Flaubert utilisait le verso de ses pages après les avoir retournées. Sur le site de travail, les pages ont toutes été « remises à l'endroit » pour faciliter le déchiffrage et la transcription. Mais ne faudrait-il pas rétablir l'orientation des images concernées par ce traitement dans une optique de « fidélité » au manuscrit conservé à la bibliothèque municipale de Rouen ? Pour le moins, ne serait-il pas souhaitable de décrire ce trait dans l'encodage TEI ?

- Comment remettre en ordre les folios qui ont été reliés sans que soit respecté l'ordre de la prise de notes de Flaubert[4] ?
- Comment juger de la pertinence de certains éléments visibles sur la page ? Prenons l'exemple des lignes verticales : certaines sont le simple produit mécanique des collages effectués par Flaubert (voir la figure 2a) ; d'autres sont des marques d'insistance et de mise en valeur de certains éléments textuels, volontairement tracées par Flaubert (voir la figure 2b).



(Figure 2a)



(Figure 2b)

Dans toutes les éditions de corpus littéraires articulant l'image du manuscrit en fac-similé et la transcription du contenu textuel, il existe une tension – pour ne pas dire une contradiction – entre plusieurs structures d'information. Le manuscrit à encoder est d'abord un objet physique. Or la TEI privilégie la logique du texte, sa structure interne (un ouvrage est divisé en grandes sections, elles-mêmes divisées en chapitres ; chaque chapitre est composé d'un titre et d'une suite cursive de paragraphes...).

Dans le projet Bouvard, cette tension intervient à de nombreux niveaux et il est particulièrement difficile d'identifier une hiérarchie unique valable pour la totalité du corpus en

4 Un des objectifs du projet est de permettre le rétablissement de cet ordre génétique et logique.

raison de diverses caractéristiques :

- Le corpus est lacunaire et l'œuvre inachevée[5].
- Le corpus se présente physiquement comme un montage d'éléments disparates.
- Son intertextualité repose sur un recopiage à l'identique de nombreux éléments.
- Les dossiers ont été conçus comme un outil de travail, ils forment un réservoir documentaire gardant la mémoire de plusieurs classements successifs, de sélections et de recopiages.
- Enfin, la présentation physique que l'on connaît (huit volumes reliés, folios montés sur onglets, tampons posés sur chaque élément collé, etc.) est le fruit d'un traitement patrimonial qui a défait l'état originel essentiellement mobile de ces dossiers de travail.

Autre difficulté et non la moindre, le passage de l'unité-page au fragment oblige à repenser la définition et l'affectation des métadonnées jusque-là attachées à la seule page.

1- Modélisation : comment nous avons procédé

Pour plus de clarté, il fallait remettre à plat le système de métadonnées et représenter toutes les structures à formaliser. L'enjeu d'une modélisation de corpus est d'identifier tous les composants principaux et de décrire leurs relations. Dans ce but, nous avons utilisé deux « outils » d'analyse.

Le premier est une typologie du document récemment présentée par Bruno Bachimont dans le cadre d'un séminaire de l'INRIA sur les métadonnées et la mutation numérique[6]. Cette typologie du document vise à articuler le contenu, le support d'inscription et les procédés d'inscription. On n'en parlera que brièvement car le travail n'est pas achevé et il intervient à un niveau plus fin par rapport au travail d'encodage. Cette typologie distingue plusieurs « couches » dans un même document :

- le support d'inscription (pour une édition imprimée, il s'agit du papier ; pour une édition électronique, du fichier XML) ;
- la forme d'inscription (le codage, c'est-à-dire la mise en page manuscrite ou typographique et l'encodage de cette mise en page) ;
- le support d'appropriation (le papier ou l'écran) ;
- la forme sémiotique d'appropriation (autrement dit la forme intelligible, textuelle ; par exemple : l'article, la lettre ou le poème) ;

5 Sur tous ces aspects, voir Stéphanie Dord-Crouslé, Bouvard et Pécuchet *de Flaubert, une « encyclopédie critique en farce »*, Paris, Belin, coll. « Belin-Sup Lettres », 2000.

6 Bruno Bachimont, « Audiovisuel et numérique. La reconstruction éditoriale des contenus », in *Métadonnées : mutations et perspectives - Séminaire INRIA 29 septembre - 3 octobre 2008*, sous la dir. de Lisette Calderan, Bernard Hidoine et Jacques Millet, Paris, ADBS Éditions, 2008, p. 195-221.

- enfin, dernière couche, la modalité d'appropriation (s'il s'agit d'une lecture sur papier ou d'une navigation sur un site web).

Dans le document papier, les couches se superposent. Là réside la difficulté du balisage : pour pouvoir encoder, il faut d'abord procéder à une opération de décomposition, de discrimination de ce qui appartient à telle ou telle couche. L'encodage est d'abord un dés-encodage.

Intégrant une distinction entre support et processus d'inscription, cette typologie nous a paru particulièrement intéressante pour réorganiser le système de métadonnées en fonction des contraintes du projet. Au niveau fin du manuel TEI, elle permet de sélectionner les éléments, de définir les attributs requis avec toutes leurs valeurs possibles ; au niveau de l'édition, elle permet de répartir les métadonnées entre base de données et transcription TEI, limitant ainsi les risques de désynchronisation à chaque nouvelle mise à jour.

Le second « outil » est un modèle de représentation de la macro-structure du corpus à encoder, tel que l'ont proposé Richard Furuta et Neal Adenaert[7]. Dans leur communication, les deux auteurs remarquent que c'est la plupart du temps l'interface d'accès qui conditionne l'organisation de l'édition électronique - et non les spécificités de l'œuvre. Selon eux, passer par une étape de modélisation abstraite permet de s'affranchir des biais induits par les interfaces de gestion ou de visualisation. Ils proposent comme point d'entrée principal un outil de modélisation d'éditions électroniques autour duquel se déploient différents niveaux de description scientifique. Pour illustrer leur démonstration, ils ont construit sous forme de schéma le modèle abstrait d'une édition de poèmes en deux volumes[8]. Ce diagramme montre l'ensemble des unités minimales nécessaires dans les deux niveaux de structure de l'édition : volume, pages et lignes pour le niveau physique ; et poèmes, titres, épigraphes et strophes pour le niveau de la logique textuelle. Il montre également la hiérarchie de chaque structure, ainsi que les relations entre les deux structures, leurs convergences ou leurs divergences. Mettant en évidence les divergences entre les niveaux de structure physique et logique, il confirme l'idée exposée par Aurèle Crasson et Jean-Daniel Fekete selon laquelle, « contrairement aux autres documents numériques connus, un seul niveau de description ne suffit pas pour rendre compte d'un manuscrit[9]. »

---

7 Richard Furuta et Neal Adenaert (Texas A&M University), « Annotated Facsimile Editions : Defining a Macro-level Structure for Image-Based Electronic Editions », in *Digital Humanities 2008 Conference*, Oulu, Finlande, June 25-29, 2008, p. 47-50. Disponible sur : <http://www.ekl.oulu.fi/dh2008/Digital%20Humanities%202008%20Book%20of%20Abstracts.pdf>.

8 *Ibid.*, p. 48.

9 Aurèle Crasson et Jean-Daniel Fekete, « Structuration des manuscrits : Du Corpus à la région », *Proceedings of CIFED 2004*, La Rochelle, France, p. 162-168. Disponible sur : <http://hal.archives-ouvertes.fr/docs/00/06/24/97/PDF/sic_00001210.pdf>.

2- Élaboration d'une représentation schématique du corpus

Nous avons essayé d'élaborer notre propre schéma en commençant par analyser plusieurs « gabarits-types » de mises en page. Ce travail nous a permis de voir qu'au-delà de l'hétérogénéité typologique des documents contenus dans le corpus, une même macro-structure textuelle se répète. Pour un même texte, on trouve éventuellement une zone de titre, suivi d'un corps de texte composé de sous-éléments, qui sont la plupart du temps des paragraphes. Cependant, une autre structure se dégage des pages de notes documentaires. Celles-ci peuvent en effet être appréhendées comme des listes de notes, qu'il s'agisse de pages de notes bibliographiques, de listes de faits ou de dates, de pages de notes de lecture voire de pages de « notes de notes ».

2.1 Structure globale du corpus Bouvard : niveaux physique et logique

Après cette étape préparatoire, nous avons dessiné un premier schéma (voir la figure 3).

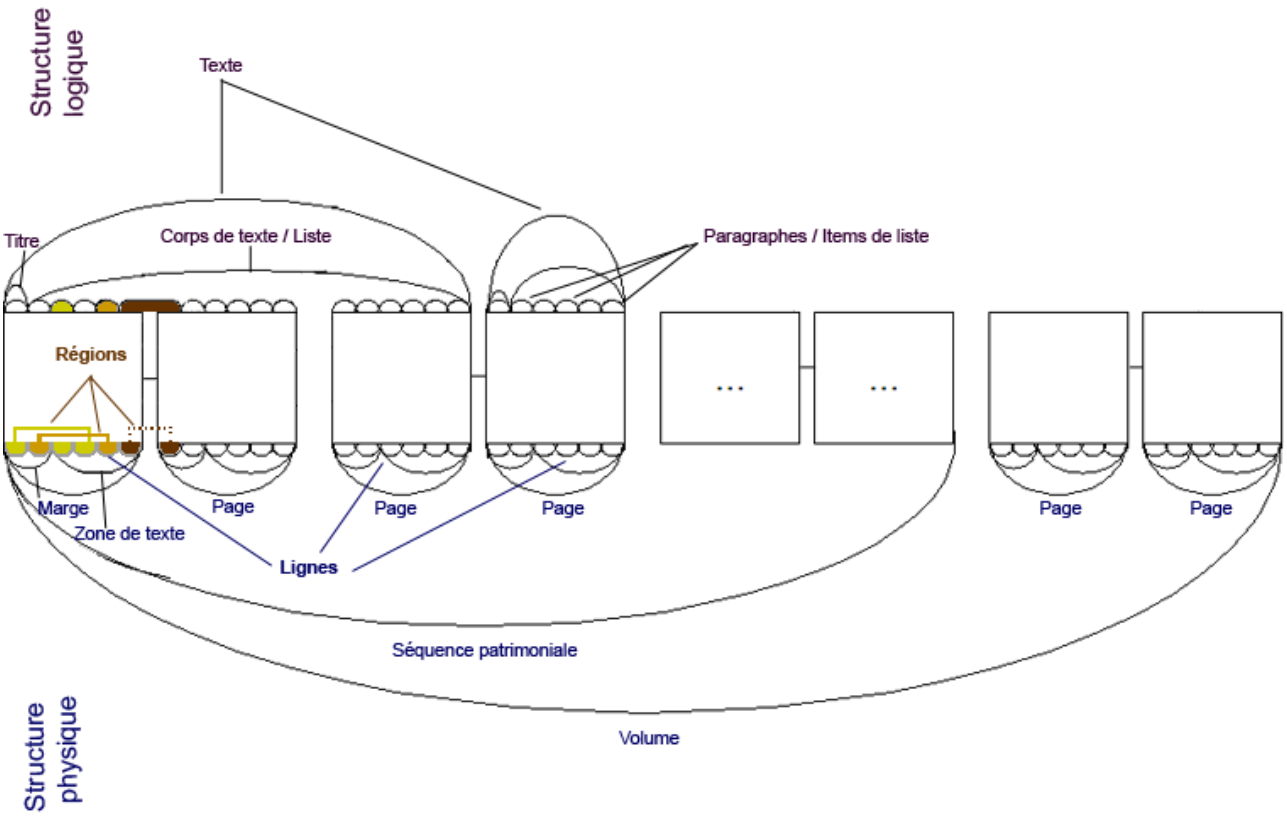


Figure 3 : La structure globale.

Cette représentation montre d'abord que les unités fondamentales de la macro-structure restent assez simples à manipuler. D'autre part, l'articulation de l'aspect physique et de la logique du contenu s'opère à deux niveaux : au niveau de la notion de « texte » (au sens d'« ensemble cohérent formant une unité de taille variable ») et au niveau de la région-fragment.

2.2 Positionnement des éléments TEI sur la structure globale

Nous avons ensuite tenté de mettre ce schéma en correspondance avec les éléments TEI (voir la figure 4).

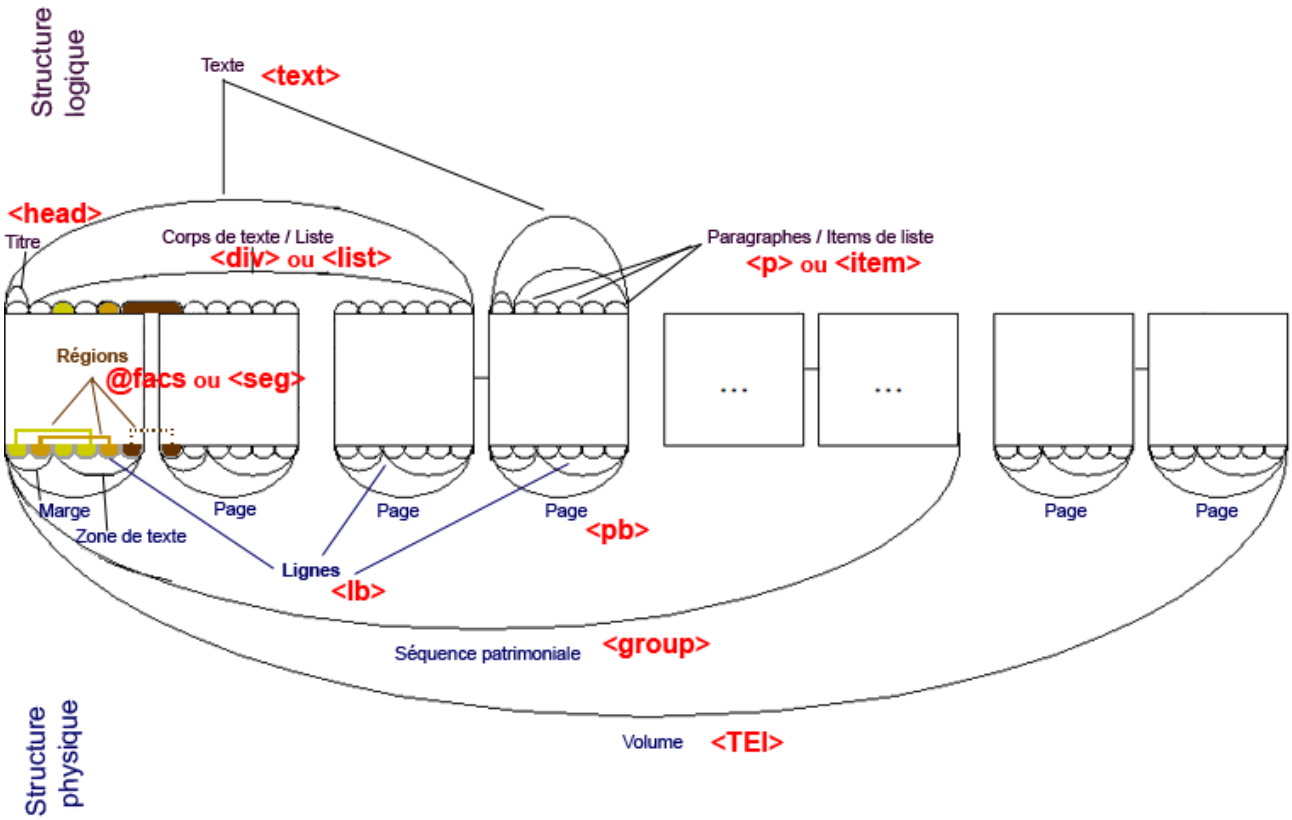


Figure 4 : La structure TEI.

Il apparaît alors qu'une arborescence de base complète semble pouvoir être définie en exploitant les éléments TEI dédiés à la structuration de corpus et de textes composites (éléments <teiCorpus> et <group>). D'autre part, on voit que la structure TEI se distribue entre les niveaux logique et physique. Le schéma permet ainsi de résoudre la contradiction pointée plus haut. Les niveaux logiques s'expriment dans le choix des éléments TEI et leur positionnement dans

l'arborescence. Les distributions topologiques (l'aspect physique) seront rendues par des attributs. Par exemple, une annotation de Flaubert de type « vedette » en marge sera un <label> de l'item de liste de notes, tandis qu'un attribut pourra indiquer un positionnement en marge.

2.3 Workflow d'encodage

Nous avons ensuite tenté d'exploiter encore mieux cette représentation. Sachant que le processus d'encodage devait nécessairement être organisé en plusieurs étapes, nous avons essayé de les positionner sur le schéma afin d'élaborer le processus complet en le visualisant (voir la figure 5).

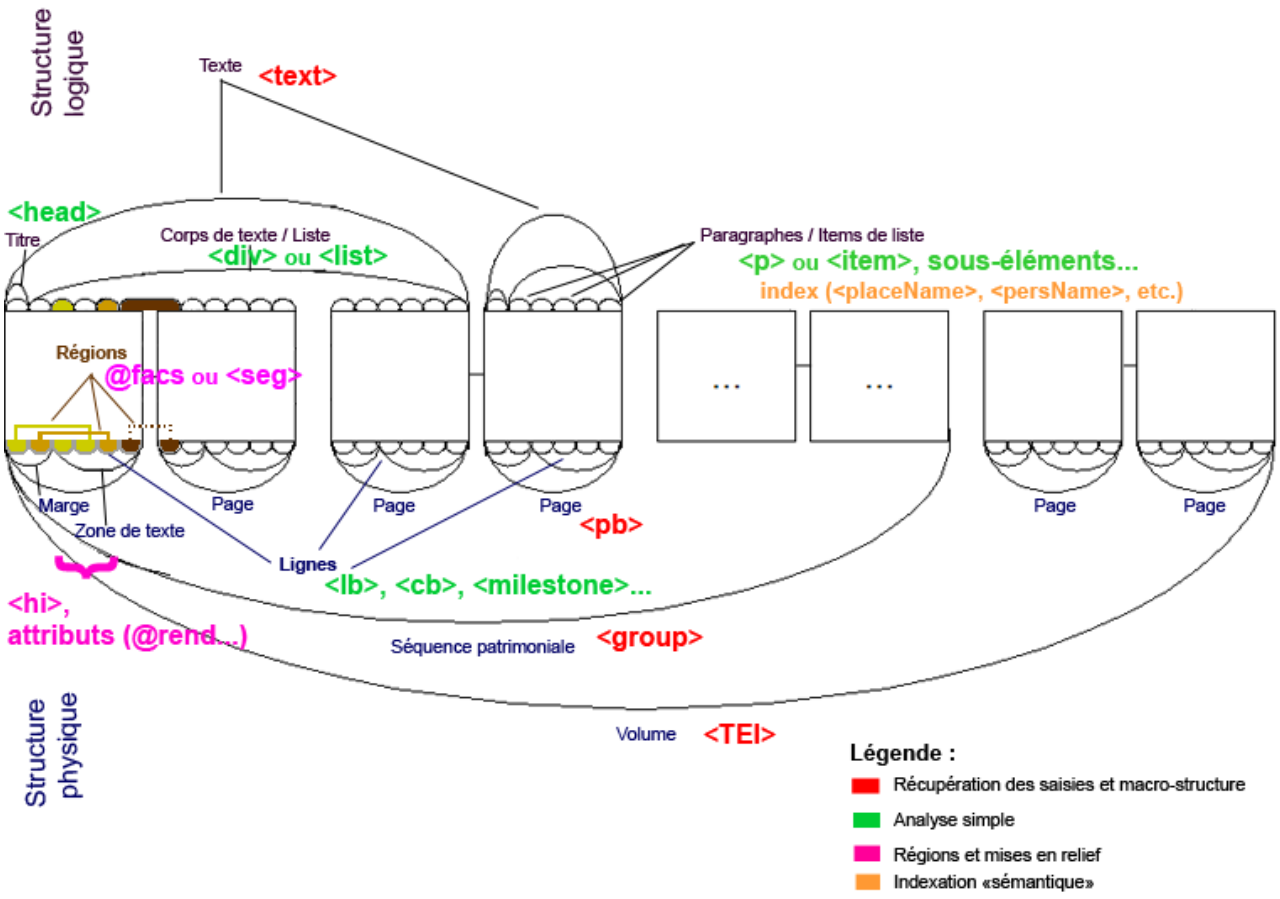


Figure 5 : Le workflow d'encodage.

L'opération pourra être organisée en quatre étapes :

1. Récupération des textes déjà saisis numériquement et encodage de la macro-structure ;

2. Analyse simple ;

3. Délimitation des régions en mettant en relation les zones dessinées sur l'image et la transcription TEI ;

4. Indexation sémantique.

3- Premiers résultats

Ce travail n'est pas achevé - comme on l'a indiqué (restent à finaliser le choix des métadonnées et leur « traduction » en TEI). Mais il est déjà possible de lister quelques-uns de ses apports pour le projet. Le premier est la représentation visuelle des deux niveaux de description principaux, le niveau physique et le niveau logique. La représentation de la macro-structure sous forme schématique pourra servir d'outil de repérage dans l'arborescence TEI pour l'encodeur. Le travail de balisage consiste en effet souvent à se demander à quel élément telle caractéristique du folio se rapporte. Par exemple, un trait horizontal de séparation entre deux notes dans une liste n'est pas un sous élément de la note ou un frère, mais un jalon. Il doit donc être encodé comme un <milestone> au même titre que les sauts de ligne ou de colonne.

En outre, le modèle permet d'opérer des choix plus précis et plus justes. Il permet de traiter les cas de chevauchement d'éléments d'une page sur l'autre. Il permet d'exploiter la description patrimoniale qui a été faite du manuscrit et qui isole des regroupements de folios pertinents. Il met en valeur le niveau logique « texte » qui est un point d'entrée important dans le corpus. Enfin, il permet de délimiter les régions dans l'arborescence TEI en utilisant chaque fois que c'est possible[10] les éléments disponibles dans cette « grammaire ».

Conclusion

Telle est la démarche que nous avons suivie pour identifier et organiser la description d'un manuscrit, en vue d'une édition électronique articulant images en facsimilé et transcriptions, dans le cadre d'une plateforme de publication web mixte relationnelle et XML/TEI.

Réalisé en dehors d'une interface ou d'un format, cet effort de modélisation vise d'abord à permettre d'opérer les choix techniques les plus pertinents et à proposer une représentation claire du projet, partageable par tous les partenaires scientifiques et plus particulièrement par ceux qui sont chargés de la transcription et de l'encodage. Mais il permet aussi de pouvoir revenir en arrière ou de réexaminer plus tard ces choix, lors d'un « repentir », d'une mise à jour ou d'une migration. Et ce d'autant plus que les entêtes TEI et le module de description des manuscrits offrent toute la syntaxe

10 Quand la région délimitée par le chercheur ne correspond à aucun élément dans l'arborescence TEI, il semble plus opportun d'utiliser l'élément précis « <seg>1 ou <ab>» plutôt que l'élément générique « ». <seg> fonctionne avec des pointeurs ; par conséquent, il correspond bien à notre besoin d'indexation de « circulations » certaines ou probables, inter ou extra corpus. Le modèle pourrait aussi permettre de prendre en compte une dimension qui avait été éludée par la structure de la base de données, celle de la distinction recto / verso (et qui serait peut-être une manière de traiter le cas des « images tête-bêche ».

nécessaire à la description de cette partie du travail d'édition. Peu importe la méthode utilisée ou le format d'encodage choisi, l'important est de garder une trace du « désencodage », c'est -à-dire des analyses et des choix réalisés. C'est tout l'enjeu de l'étape prochaine du travail qui va consister à « traduire » ce modèle et les consignes d'encodage sous la forme d'un guide d'encodage ODD, produit via l'interface ROMA.