Recherches & Travaux

72 | 2008
De l'hypertexte au manuscrit

Manuscrits de Stendhal

De la base de données à la base documentaire : le projet CLELIA

Thomas Lebarbé, Alexia Blanchard et Cécile Meynard

Édition électronique
URL : http://journals.openedition.org/recherchestravaux/94
ISSN : 1969-6434

Éditeur
UGA Éditions/Université Grenoble Alpes

Édition imprimée
Date de publication : 15 juin 2008
Pagination : 97-117
ISBN : 978-2-84310-125-0
ISSN : 0151-1874

Référence électronique
Thomas Lebarbé, Alexia Blanchard et Cécile Meynard, « Manuscrits de Stendhal », *Recherches & Travaux* [En ligne], 72 | 2008, mis en ligne le 15 décembre 2009, consulté le 30 avril 2019. URL : http://journals.openedition.org/recherchestravaux/94

Thomas LEBARBÉ
Alexia BLANCHARD
Cécile MEYNARD
Université Stendhal - Grenoble 3

Manuscrits de Stendhal
De la base de données à la base documentaire : le projet CLELIA

Introduction

Les manuscrits de Stendhal, conservés en très grande majorité par la Bibliothèque municipale de Grenoble, représentent un patrimoine culturel fondamental qu'il est nécessaire de préserver mais aussi de mettre à la disposition du public scientifique comme du grand public[1].

D'autre part, il convient de rééditer de façon scientifique et exhaustive l'ensemble de ce fonds qui, jusqu'à présent, a toujours été publié au coup par coup et dans lequel il subsiste quelques documents inédits. De plus, il semble particulièrement pertinent de fournir un travail scientifique sur l'intertexte et sur l'acte de création littéraire (d'un point de vue littéraire et linguistique). Étant donné le nombre et la complexité des documents à traiter, seule la mise en place d'un corpus numérique peut s'avérer satisfaisante, rendre possible l'analyse littéraire et linguistique des écrits, et à terme permettre la réalisation d'éditions critiques électroniques et/ou papier beaucoup plus satisfaisantes.

Gérald Rannaud, Serge Linkès et Jean-Yves Reysset avaient créé en 2001 un prototype de base de données numérique, qui devait à terme devenir le Catalogue informatisé du fonds Stendhal. Le projet de base documentaire des

1. En complément de cet article, un diaporama a été mis en ligne sur le site de la MSH-Alpes : http://www.msh-alpes.prd.fr/Actualites/Manuscrits.htm.

manuscrits de Stendhal CLELIA (Corpus littéraire et linguistique assisté par des outils d'informatique avancée) a ainsi pour objectif de reprendre ce travail colossal en le remaniant et en l'améliorant, la perspective de mettre en place un support plus adéquat à la mise à disposition sur Internet. Deux partenaires appartenant à l'Université Stendhal - Grenoble 3 sont ainsi associés dans ce projet, l'équipe Traverses 19-21 (composante Centre d'études stendhaliennes et romantiques) qui fournit le travail scientifique sur les pages, et le laboratoire LIDILEM qui se charge de la partie informatique et linguistique. Cette plate-forme collaborative réunit une vingtaine de personnes (chercheurs littéraires, informaticiens et linguistes) coordonnées par Cécile Meynard, en partenariat avec la Bibliothèque municipale (chargée de la numérisation des manuscrits).

L'objectif général de ces travaux reste la constitution d'un corpus littéraire et linguistique, accompagné des outils nécessaires à sa manipulation. Dans cette perspective, la base documentaire a été définie de manière à intégrer aisément des outils de traitement automatique des langues et de consultation avancée des données sous forme de modules logiciels.

Une base de données existante

Concevoir une base de données consiste à définir un modèle de l'objet (le manuscrit et sa transcription en l'occurrence) dans la perspective de l'usage ou des usages que peuvent en faire les utilisateurs. Le modèle développé par Gérald Rannaud, Serge Linkès et Jean-Yves Reysset répondait à deux exigences :

– représenter un certain nombre d'informations sur chaque page : transcription, qui se voulait une simple aide à la lecture ; et description physique de la page et du document auquel cette page appartient ;

– fournir une interface logicielle simple pour le travail de transcription et la consultation.

Informatiquement, ce modèle se matérialise en une table se subdivisant implicitement en quatre sous-parties :

– *identification de la page* : cote, volume, foliotage de la bibliothèque ;

– *description de la page* : description physique (dimension, encre, type de papier) ; description logique (outil d'écriture, scripteur, présence d'annotations et de corrections, dates indiquées, lieux indiqués, graphismes…), et observations sur la page ;

– *analyse du document* : rattachement de la page à une collection logique indépendante des rassemblements effectués par les bibliothèques, identifiable par son support, son unité logique, sa provenance, ses dates de début et de fin de rédaction, son lieu de rédaction ;

– *aide à la lecture* : la transcription en elle-même ainsi que des informations relatives à la transcription, notamment des mots-clés (par exemple «Daru», quand Stendhal écrit «Z») rendus invisible par l'usage du blanc comme couleur de police[2].

Cette base de données était construite à l'aide du système de gestion de bases de données FileMaker® Pro. Ce logiciel avait à l'époque l'avantage d'offrir à des utilisateurs néophytes les moyens de mettre en service rapidement et simplement une base de données. La copie d'écran ci-après (fig. 1) montre l'interface produite par Gérald Rannaud, Serge Linkès et Jean-Yves Reysset affichant l'une des 853 pages transcrites.

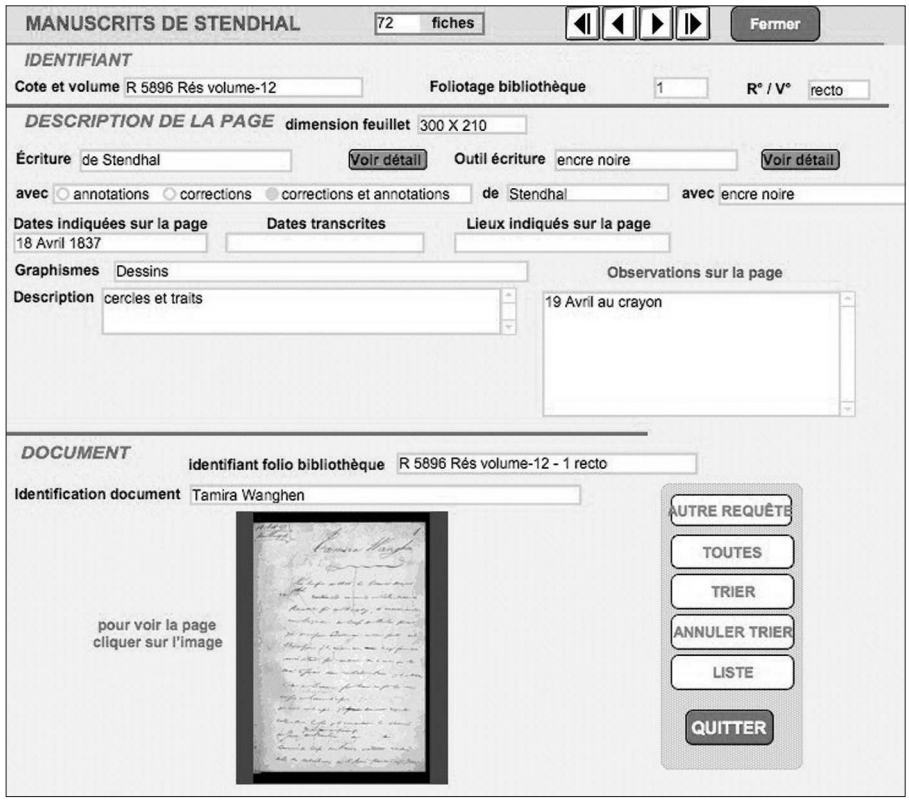


Fig. 1. Modèle de la base de données de 2001. L'original est en couleur.

2. L'usage de la couleur de police blanche a été nécessaire car, FileMaker® ne permettant pas de mettre aisément en place un système de balisage, ces mots-clés étaient inscrits directement à la suite de la transcription, et ne devaient donc pas être visibles pour l'utilisateur. Cette limite du prototype est l'une des raisons de son abandon.

Toutefois, l'utilisation de FileMaker® Pro présente un certain nombre d'inconvénients pour un usage plus large de la base de données. L'usage ayant été prévu comme individuel, différentes versions des données se sont retrouvées sur différents ordinateurs, versions difficiles à rassembler dans une même source de données. Le logiciel se prête peu à la mise en ligne des données et implique des investissements importants puisqu'il s'agit d'un logiciel propriétaire plus adapté aux ordinateurs Apple qu'aux PC (le transfert des informations et de la base de données des uns vers les autres étant complexe).

Augmentation de la base de données

Bien qu'une refonte profonde ait été nécessaire afin d'éviter les aléas de la base de données faite avec FileMaker® Pro, l'énorme travail effectué devait être pérennisé, afin de tirer profit des avancées techniques en matière de bases de données et d'augmenter le potentiel déjà important de la base actuelle.

Notre projet repose sur des exigences plus complexes que la base initiale :

– la valorisation du patrimoine des manuscrits par une mise en ligne libre sur Internet ;

– la possibilité donnée aux transcripteurs d'alimenter cette base à distance depuis leur ordinateur, indépendamment de la plateforme utilisée (Mac, PC…) ;

– la centralisation de l'ensemble des données afin d'éviter les doublons ou les problèmes de rassemblement des données ;

– l'intégration d'outils de recherche simple pour le grand public et d'outils de recherche avancée pour les experts ;

– la restructuration dynamique des données de transcription en fonction de critères définis par les utilisateurs afin de produire des éditions électroniques virtuelles ;

– l'introduction d'une dimension évolutive permettant d'adapter à l'avenir cet outil à de nouveaux besoins.

Par ailleurs, il nous a semblé opportun de profiter de ce travail de refonte pour modéliser à nouveau la base de données. En effet, certaines données, bien que présentes, étaient mélangées indistinctement à d'autres, comme les commentaires de transcription par exemple. D'autres données n'étaient pas représentées dans leur complexité : ainsi une page sur laquelle sont intervenus deux scripteurs distincts pouvait être transcrite en indiquant les deux scripteurs concernés mais ne permettait pas de délimiter les parties d'écriture respectives. De même, une page comportant deux documents distincts (par exemple un fragment de théâtre et un paragraphe à caractère diariste) n'était référencée que pour le document jugé principal, le document secondaire étant indiqué dans des commentaires.

1. Restructuration des données

Nous avons donc opéré une restructuration de la base de données comme le montre la figure 2, ci-après. Au centre aussi bien de la figure que du modèle, la page, objet de la transcription. Par conséquent, elle est définie notamment par ses propriétés physiques et par les observations du transcripteur relatives à l'objet matériel. Cette page se rattache à une collection, permettant une restructuration dynamique du corpus[3].

Des dates, dont la forme graphique sur la page n'est pas conforme à une interprétation numérique (notamment les dates du calendrier révolutionnaire), sont présentes sur certaines pages et sont donc mises en relation avec celles-ci, leur nombre étant potentiellement infini. Ceci permet une meilleure gestion de l'aspect temporel de la rédaction, information capitale pour l'analyse génétique tout comme pour l'analyse linguistique diachronique.

Des lieux sont indiqués sur les pages, à titre informatif pour indiquer le lieu de rédaction ou à titre rédactionnel (références à des lieux dans le discours de l'auteur). Dans les deux cas, les noms de lieu peuvent être mal orthographiés ou délibérément codés par le scripteur. Ainsi la ville de « Milan » peut être désignée par Stendhal sous la forme « 1000 ans », et un logiciel ne peut reconnaître la ville derrière ce pseudonyme – d'où la nécessité pour le transcripteur de signaler la forme équivalente à cette forme écrite.

La page est constituée d'unités d'écriture, de graphismes et de foliotages, chacun avec des outils d'écriture éventuellement différents (crayon, encre rouge, encre noire…) et surtout par des scripteurs (ou auteurs) différents. À ce jour un certain nombre de scripteurs différents ont pu être identifiés[4] sur les pages analysées et retranscrites. Cette complexité est représentée par la jointure « QuiEcritQuoi » entre les tables « Page » et « Auteur ».

Enfin, chaque transcription est effectuée par un transcripteur. La reconnaissance et la valorisation de ce travail sont capitales et motivantes pour les membres de l'équipe. Pour chaque page transcrite, le transcripteur est affiché. Ceci présente aussi l'avantage de permettre d'attribuer des tâches différentes aux membres de l'équipe au niveau de la saisie, évitant ainsi des redondances.

Par ailleurs, ce que la figure 2 ne montre pas pour plus de lisibilité, un historique des modifications des données de transcription est préservé. Il est

3. Le travail collaboratif et éminemment pluridisciplinaire avec C. Meynard et l'ensemble de l'équipe « Manuscrits de Stendhal » nous a permis depuis de mettre en évidence que certaines pages contiennent des parties de texte dépendant de collections différentes. Ce lien entre page et collection a donc été remanié.

4. Liste provisoire des scripteurs déjà identifiés par G. Rannaud et son équipe : Stendhal, Colomb, Crozet, Fougeol, Delbono, Uralez, copiste inconnu, commis, secrétaire, Monsieur Azile, Bonavie, copiste de Rome, Corbeau, Vismarra, Borsieri, bibliothécaire, Lambert, Félix Faure, Pauline Beyle.

ainsi possible d'accéder à la fiche transcrite mais aussi à sa version modifiée à la suite de l'avis émis par le comité de validation.

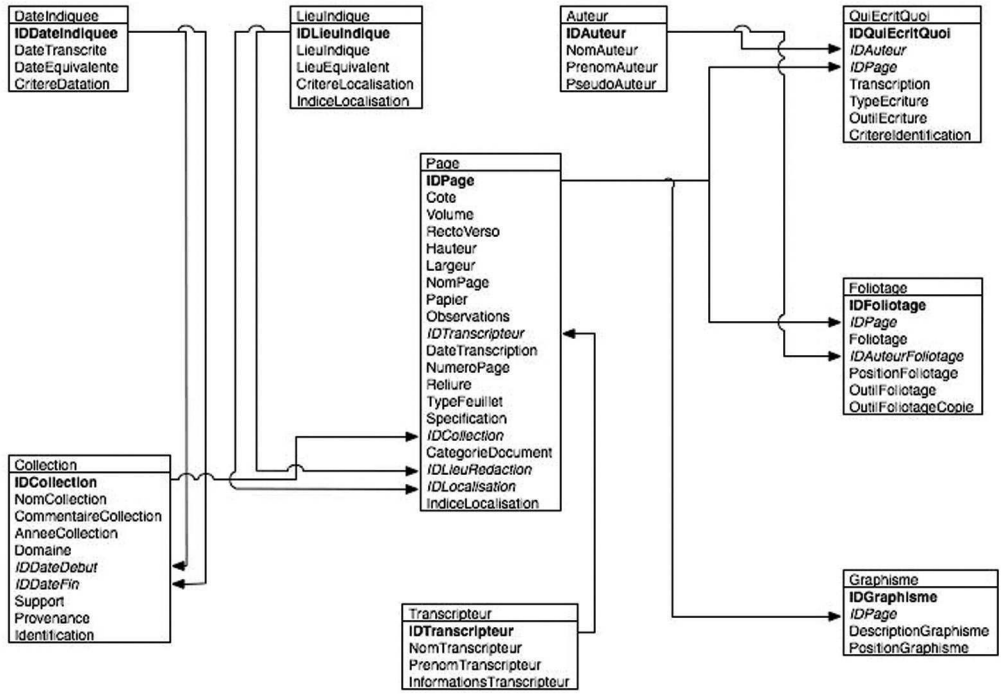


Fig. 2. Modèle relationnel de la nouvelle base de données.

2. Implantation informatique

Cette modélisation peut donner une impression de complexité inutilement exagérée. La modularisation des données de transcription présente toutefois l'avantage d'ouvrir les possibilités de description et d'éliminer les redondances : le transcripteur peut associer chaque partie de texte à son scripteur, et n'est plus contraint de reproduire la description d'une collection[5] pour chacune des pages s'y rattachant. Par ailleurs, la modularisation des données facilite et allège les requêtes du moteur de recherche : ainsi, plutôt que de parcourir l'ensemble des données de transcription pour trouver les pages correspondant à un critère en particulier, le moteur de recherche n'aura qu'à parcourir un sous-ensemble du descriptif. Pour une centaine de pages consultées sur un

5. Cette notion de « collection » a depuis été étendue à une triple attribution : document, corpus, et domaine. Par exemple « Tamira Wanghen » appartient au corpus du « Rose et Vert » qui appartient lui-même au domaine « Romans et nouvelles ».

ordinateur de bureau, la différence de temps de recherche entre l'ancienne base de données et la nouvelle reste négligeable ; en revanche pour les 32 000 feuillets prévus à terme sur un serveur Internet consulté simultanément par plusieurs utilisateurs, la différence de temps de recherche peut varier d'un coefficient de 10 à 100.

Le modèle de base de données a été implanté en MySQL – norme libre, gratuite, reconnue au sein de la communauté informatique et régulièrement mise à jour. L'interface de gestion des données (consultation, ajout, modification) a été implantée dans le langage PHP, lui aussi libre, gratuit, reconnu et régulièrement mis à jour. Les données de l'ancienne base ont été converties et intégrées dans la nouvelle. Certaines informations ont toutefois été perdues dans la conversion, en particulier les informations de mise en forme (texte biffé par exemple), car encodées dans le format propriétaire de FileMaker® Pro.

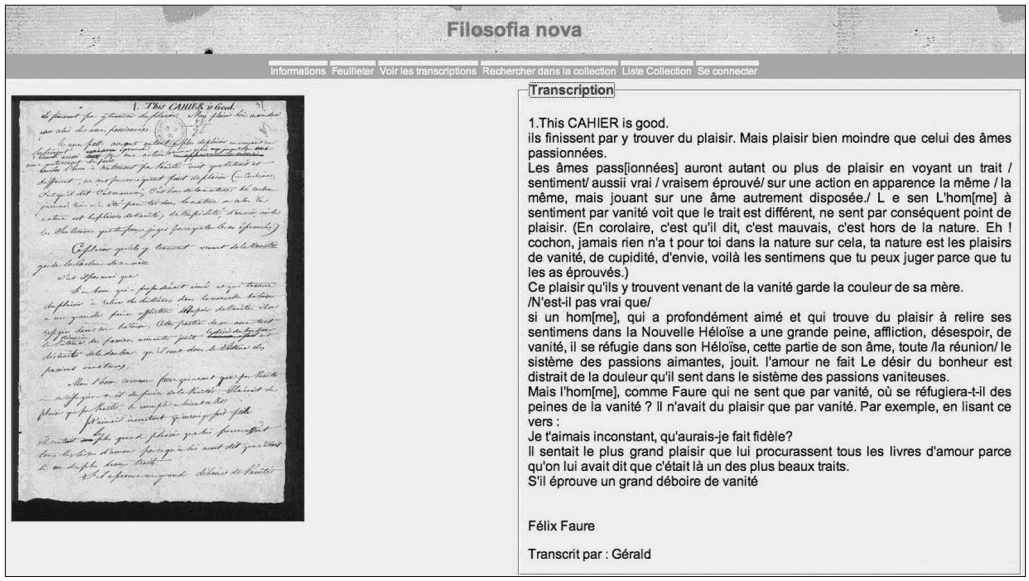


Fig. 3. Exemple d'affichage d'aide à la lecture d'une page intégrée dans la base de données MySQL. L'original est en couleur.

Vers une base documentaire en ligne

Le prototype de base de données a été développé afin de mettre à disposition les transcriptions existantes. Nous avons toutefois constaté que l'utilisation d'une interface Web pour la description des pages pouvait s'avérer fastidieuse. Qui plus est, nous n'avions pas prévu de modalités pratiques pour décrire la mise en forme (surcharges, biffes, interlignes…).

1. De la transcription au document structuré

Dans la définition d'un protocole pour la description de contenu des pages, nous avons fait le constat suivant :

– les éléments à décrire sont des blocs de texte ayant des propriétés les distinguant du bloc englobant ;

– ces blocs s'enchâssent les uns dans les autres ;

– les propriétés distinctives peuvent être d'ordre graphique (par exemple l'outil d'écriture utilisé) ou scientifique (par exemple une aide à l'interprétation, un commentaire d'édition).

Ces particularités correspondent au standard d'encodage de l'information XML (*eXtensible Mark-up Language*) du consortium W3C (*World-Wide Web consortium*). XML est un langage de description de contenu totalement ouvert fondé sur un principe de balises de délimitation (telles les balises HMTL). Un document XML peut être assorti d'une feuille de style permettant de le mettre en page (CSS : *Cascading StyleSheet*), d'une grammaire contraignant l'usage des balises au sein du document (DTD : *Document Type Definition*) et de routines de manipulations afin de transformer un document XML en un autre type de document (XSLt : *eXtensible StyleSheet Language Transformations*).

Le diagramme de la figure 4 ci-après représente une partie du potentiel de description des données de transcription possible dans le modèle XML défini en collaboration avec les membres de l'équipe Traverses 19-21. XML permet de définir des éléments avec certaines propriétés (attributs). Chaque élément contient lui-même des éléments. Dans notre cas, l'élément de départ (la racine du document XML) est la page (un recto ou un verso de feuillet) contenant un descriptif (les propriétés de la page) et un contenu (le contenu de transcription). La particularité d'un tel mode de description est d'assurer l'héritage des propriétés (attributs). À titre d'exemple, un des attributs de l'élément « page » est le « scripteur ». Tous les éléments contenus dans la page, quelle que soit la profondeur de l'arbre, héritent donc de cette propriété, sauf si stipulé autrement. Ainsi, il n'est pas nécessaire de spécifier le scripteur pour chacun des éléments mais seulement pour la page puis pour les éléments pour lesquels le scripteur diffère (par exemple une annotation de Stendhal sur une page rédigée par un de ses secrétaires). Il en est de même pour l'outil d'écriture : l'ensemble de la page peut être écrit à l'encre noire (attribut de « page »), mais il est possible de signaler une biffe faite au crayon (attribut de « biffe »)[6].

6. Si le modèle de description permet de stipuler qu'une biffe a été effectuée au crayon sur un mot écrit à l'encre noire, nous atteignons toutefois les limites des capacités de représentation graphique. L'information est donc présente et peut être utilisée dans les recherches, mais n'est pas représentée à l'écran. Il en va de même pour l'emplacement d'une marginale par exemple, qui ne sera pas représentée à sa place exacte à l'écran, même si cette dernière est signalée dans les données de transcription.

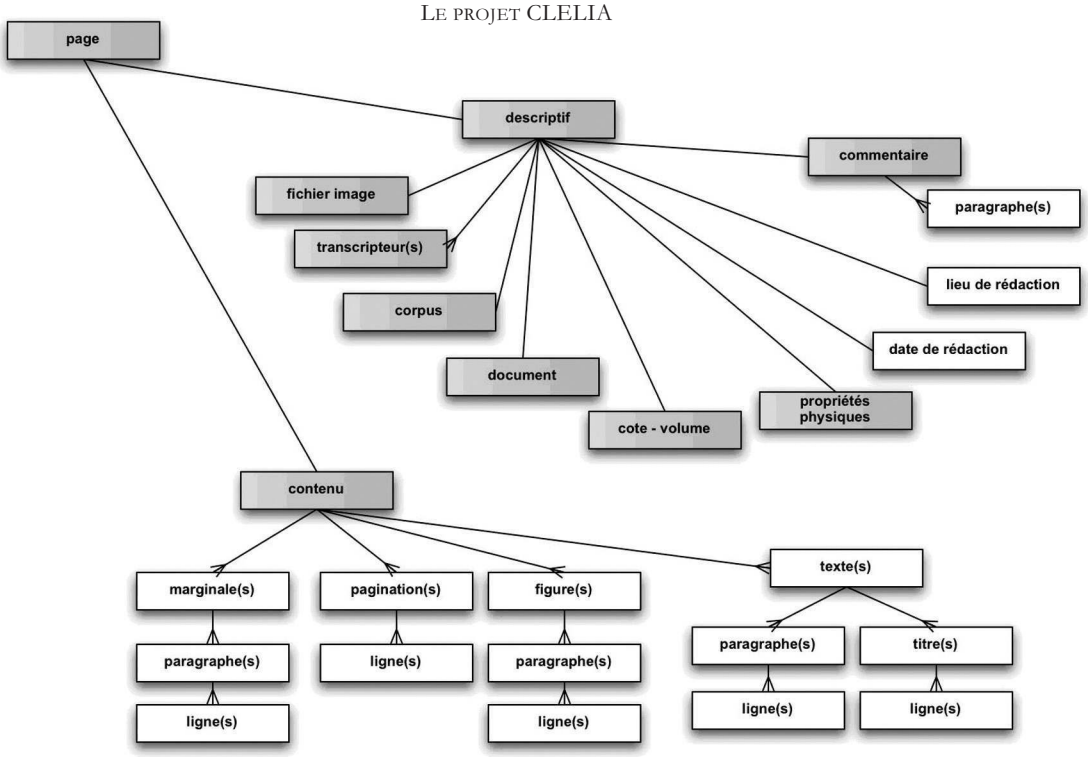


Fig. 4. Représentation XML de l'arborescence de description.

L'inconvénient d'un tel langage reste toutefois sa manipulation. Il n'est pas judicieux d'attendre d'un néophyte qu'il encode l'information «en brut», c'est-à-dire en délimitant manuellement les unités. Ainsi, comme le montre la figure 6, l'encodage d'un simple paragraphe exige l'utilisation de balises entre chevrons (< et >), certaines spécifiées par des attributs (outil="encre_noire"), toute erreur de graphie dans ces balises ayant des conséquences funestes sur la reconnaissance du document (la grammaire indiquant ce qui peut ou ne peut pas être utilisé comme balises et attributs, le document sera déclaré non conforme).

2. Outillage informatique

Pour simplifier au maximum la tâche des transcripteurs, chercheurs littéraires pour la plupart, nous avons donc mis en service un logiciel libre d'aide à la rédaction de documents XML : Morphon XML Editor. Cet outil, développé par *Lunatech Research*, présente de nombreux avantages non négligeables pour les utilisateurs concernés :

– le logiciel fonctionne sur toutes les plateformes informatiques quel que soit le système d'exploitation (PC Windows®, PC Linux, Apple®…);

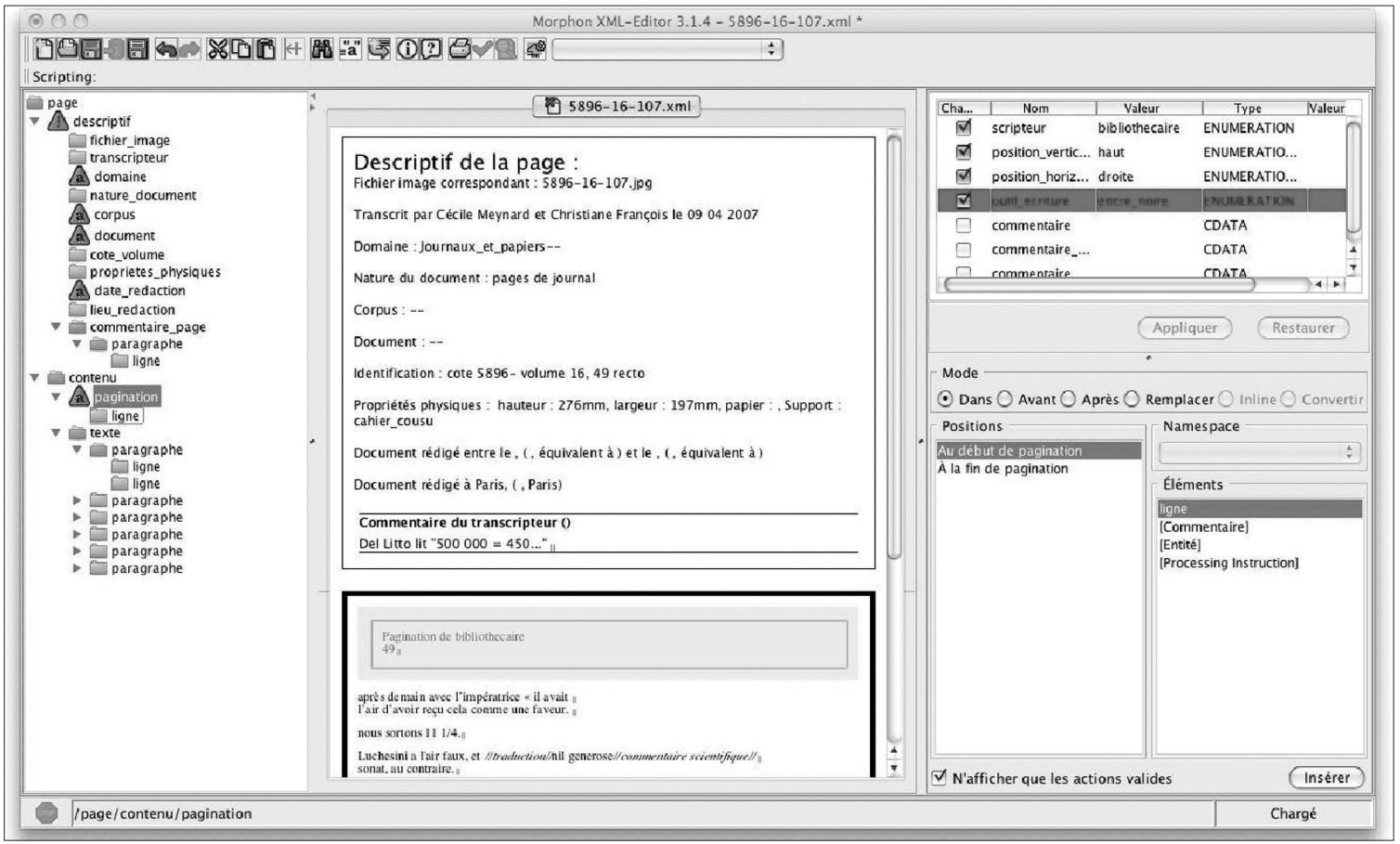
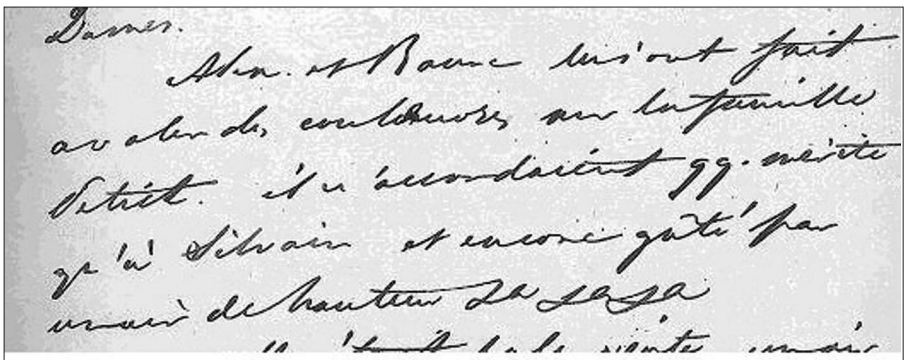


Fig. 5. Interface de transcription avec Morphon XML Editor.

– il est multilingue et surtout les menus et options sont en français, ce qui rend l'utilisation plus conviviale ;

– il permet une approche WYSIWYG[7] de la rédaction grâce à la feuille de style qu'il prend en compte ; ceci permet au transcripteur de vérifier la pertinence de son encodage au fur et à mesure de la frappe (voir en figure 6 un exemple de texte simple à décrire, son codage XML et l'apparence de la transcription telle que la voit l'utilisateur avec Morphon ;

– il tient compte de la grammaire et empêche l'utilisateur de mal structurer son document tout en lui proposant les différentes possibilités.



```
49:       <paragraphe>
50:         <ligne><code equivalent="Alexandrine [Daru]" estime="atteste"
50: estime_par="contenu">Alex.</code> et Baure lui ont fait </ligne>
51:         <ligne>avaler des couleuvres sur la famille </ligne>
52:         <ligne>Petiet. il n'accordaient qq. mérite </ligne>
53:         <ligne>qu'à Silvain et encore gâté par </ligne>
54:         <ligne>un air de hauteur &amp;a&amp;a&amp;a </ligne>
55:       </paragraphe>
```

Alex.|Alexandrine [Daru]| et Baure lui ont fait ||
avaler des couleuvres sur la famille ||
Petiet. il n'accordaient qq. mérite ||
qu'à Silvain et encore gâté par ||
un air de hauteur &a&a&a ||

Fig. 6. Exemple d'encodage XML d'un paragraphe de manuscrit simple (5896, vol. 16, f° 48 recto), du code XML correspondant à sa transcription et de l'affichage pseudo-diplomatique.

L'utilisateur (le transcripteur) a donc la possibilité avec l'outil que nous lui avons proposé de rédiger ses transcriptions dans un format facilement manipulable par les outils informatiques, tout en ayant des modalités de

7. WYSIWYG : « *What You See Is What You Get* » (ce que vous voyez est ce que vous obtiendrez).

description dans ses termes de spécialiste et avec un aperçu du résultat pseudo-diplomatique[8].

3. Parité base de données - base documentaire

Par ailleurs, la méthode consistant à encoder les transcriptions en XML selon une grammaire précise présente aussi un avantage non négligeable. L'information étant structurée et contrainte, il est possible de découper les fichiers de transcription en fonction des critères de la base de données. Nous disposons donc d'une parité parfaite entre la base de données et les fichiers XML de transcription. Bien que ceci génère des redondances de stockage de l'information, nous pouvons tirer profit de la double représentation des données :

– la base de données permet un accès et une recherche rapides dans les métadonnées ;

– le format XML assure un stockage pérenne de l'ensemble des informations et un affichage conforme aux intentions des transcripteurs ;

– la base de données permet une indexation fine, y compris des données de transcription (quels termes sont utilisés dans quels feuillets par quel scripteur) ;

– le format XML assure un affichage identique à celui voulu par le transcripteur sans aucun traitement informatique (allégeant ainsi la tâche du développeur informatique tout en réduisant la charge du serveur en ligne).

Une interface en ligne permet aux transcripteurs de déposer les transcriptions en les vérifiant une à une, l'expérience ayant montré que seule la mise en ligne montre clairement les erreurs d'encodage de l'information de transcription. Lors du téléversement (*file upload*), le système décortique le fichier XML afin d'enregistrer l'ensemble des données dans la base. Inversement, de manière à pérenniser les fiches existantes de l'ancienne base, celles-ci ont été converties au format XML et mises à disposition des transcripteurs afin qu'ils les améliorent avant de les réintégrer dans la plateforme.

Une interface à fonctionnalités multiples

La plateforme que nous avons conçue a pour objectifs principaux d'être une plateforme collaborative pour l'équipe de transcription mais aussi la boîte à outils de tous les utilisateurs potentiels de CLELIA. Développée dans le

8. Pour des raisons de limites des capacités de représentation graphique évoquées ci-dessus (note 6), et également parce que l'objectif restait de donner essentiellement une aide à la lecture, nous avons en effet renoncé à donner de chaque page une transcription diplomatique, qui aurait alourdi inutilement la tâche de saisie, l'utilisateur pouvant de toute façon se reporter à l'image numérisée s'il souhaite obtenir des informations plus précises sur la mise en page stendhalienne.

langage PHP, pour sa simplicité et son interopérabilité, elle est hébergée sur un serveur spécialisé par la Maison des sciences de l'Homme-Alpes[9].

Différentes catégories d'utilisateurs ont été définies suivant les usages envisagés. De nombreuses rubriques sont communes, mais peuvent diverger dans leur interprétation logicielle suivant la catégorie de l'utilisateur.

1. Consultation

Deux modes de consultation sont prévus : consultation grand public et consultation spécialisée. La base documentaire ayant comme objectif principal, dans le cadre du partenariat avec la municipalité de Grenoble et la Bibliothèque municipale, la valorisation du fonds des manuscrits de Stendhal, l'interface initiale est celle destinée aux utilisateurs « grand public ». Pour ceux-ci, les modalités de recherche d'information doivent être aussi simples que possible et la visualisation la plus légère possible. Par conséquent, après une page d'accueil décrivant succinctement les principes du projet des manuscrits et la façon de trouver ce qu'ils cherchent sur le site, ils ont le choix entre les options suivantes :

– Feuilleter les manuscrits : par groupes de dix pages miniaturisées, l'utilisateur parcourt l'ensemble de la base documentaire. Un simple clic sur une miniature lui permet d'accéder à une présentation en vis-à-vis de l'image de la page et de sa transcription linéaire. Cette présentation permet aussi de feuilleter les pages les unes après les autres. Il existe évidemment une fonction de zoom sur l'image.

– Rechercher des mots-clés au sein des manuscrits : les mots-clés sont recherchés uniquement dans le texte de transcription (et non dans les commentaires des transcripteurs) et un catalogue des pages est affiché de la même manière que pour l'option « feuilleter ». De même, en cliquant sur une miniature, l'utilisateur accède à la présentation en vis-à-vis de la page et de sa transcription linéaire.

– Accéder à des catalogues prédéfinis : une liste de catalogues, conçus par l'équipe de transcription, permet à un utilisateur de feuilleter les registres dans leur ordonnancement matériel et des éditions virtuelles, par exemple les dossiers de *L'Histoire de la peinture en Italie* ou de *L'Italie en 1818*, qui, par le hasard des classements successifs, se sont trouvés disséminés dans plusieurs registres.

– Concevoir son propre catalogue : au fur et à mesure de son parcours, l'utilisateur peut enregistrer une ou plusieurs pages dans un catalogue personnel. Cette option a été envisagée particulièrement pour les enseignants qui

9. http://stendhal.msh-alpes.fr/CLELIA.

souhaiteraient concevoir des supports pédagogiques. On peut ainsi, pour ne prendre qu'un exemple, imaginer un dossier « Madame Daru », qui regrouperait toutes les pages de documents et corpus divers (correspondance et journaux essentiellement) où est évoquée cette femme aimée par Stendhal.

L'ensemble de ces accès se fait sans identification de l'utilisateur. L'un des avantages de cet accès non contrôlé est qu'il n'est pas rébarbatif pour l'utilisateur peu chevronné ou simple curieux. Par ailleurs, l'absence d'identification permettra aussi l'indexation de l'ensemble des pages par les moteurs de recherche et permettra aux utilisateurs de s'échanger des liens vers des pages spécifiques.

L'utilisateur spécialisé dispose des mêmes accès que l'utilisateur grand public, et d'options supplémentaires :

– Il peut choisir d'accéder à un moteur de recherche avancée qui lui permettra d'effectuer des requêtes complexes (selon les particularités des pages : scripteur, présence de graphismes, recherches complexes de mots avec opérateurs booléens, par exemple « jésuite » et « politique » pour n'afficher que les pages comportant ces deux mots …) et de préciser si la requête doit être étendue aux commentaires intégrés dans les données de transcription.

– Le fait d'avoir un accès identifié lui permet d'enregistrer des catalogues personnels et, s'il le souhaite, de les partager avec les autres utilisateurs (quels que soient leurs droits d'accès), c'est-à-dire que son catalogue sera accessible, comme tout autre catalogue, dans l'interface des catalogues prédéfinis mentionnée ci-dessus.

La notion d'utilisateur spécialisé ne limite en rien le statut de l'utilisateur. Il peut être chercheur ou simple passionné. Aucune validation manuelle n'est effectuée lors de l'inscription, seule l'adresse courriel est vérifiée par un « lien magique » (*magic link*)[10].

2. Intégration des données

L'intérêt d'une plateforme logicielle est de permettre différentes formes d'accès à des données centralisées. Ainsi, par un accès contrôlé par identifiant et mot de passe, les transcripteurs et le comité de validation des transcriptions accèdent sur le même site et de manière sécurisée à des interfaces spécialisées.

Les transcripteurs peuvent travailler hors ligne avec l'outil Morphon XML Editor et ainsi revenir sur une transcription inachevée sans contrainte. À tout moment, ils peuvent déposer leurs transcriptions sur la plateforme afin d'en

10. Le principe du *magic link* permet de valider une inscription en ligne sans intervention humaine. L'utilisateur donne une adresse courriel lors de son inscription ; un message lui est alors envoyé avec un lien vers une page Web, ce qui permet de s'assurer que l'utilisateur est bien propriétaire de l'adresse indiquée.

vérifier l'affichage en ligne. Dans cette interface, ils peuvent visualiser tous les modes d'affichage afin de s'assurer de la pertinence de leur transcription selon les publics visés.

S'ils constatent des erreurs, les transcripteurs ont la possibilité de corriger leur fichier XML hors ligne et de le mettre à jour sur le serveur.

Quand ils le souhaitent, ils peuvent soumettre leurs transcriptions au comité de validation. Celui-ci a pour fonction de s'assurer de la qualité des transcriptions avant leur mise à disposition du public. La constitution du comité, ses principes et son éthique de fonctionnement restent indépendants de l'interface logicielle constituée pour la validation des transcriptions.

Le comité de validation est donc un « utilisateur » en tant que personne morale (par opposition aux personnes physiques que sont les transcripteurs). Lorsque le comité se connecte, il a accès à la liste des transcriptions considérées par leurs transcripteurs comme terminées. Pour chacune d'elles, il peut vérifier l'ensemble des données de transcription et les différentes représentations visuelles correspondant à chaque type d'utilisateur.

Afin de ne contraindre en rien les modes de fonctionnement du comité, trois possibilités sont mises à disposition :

– contacter le transcripteur par courriel afin de lui demander des modifications de sa transcription ;

– télécharger le fichier XML afin d'y apporter des modifications (mineures et dans les limites de l'éthique scientifique) puis le mettre à jour sur le serveur ;

– valider la transcription, ce qui revient à la mettre immédiatement à disposition du public.

3. Un modèle de séparation des données

Afin de ne prendre aucun risque de perte de données, lorsqu'un fichier XML est remplacé sur le serveur, une copie de la version précédente est préservée en cas d'erreur de manipulation par un utilisateur, qu'il s'agisse du transcripteur ou du comité de validation. C'est d'ailleurs dans cette optique que la plateforme est conçue : protéger au mieux les données des erreurs de manipulation.

Par ailleurs, l'ensemble du système a été prévu de manière à bien séparer les données :

– les images, propriétés de la Bibliothèque municipale de Grenoble ;

– les transcriptions et données XML, créées par les membres de l'équipe « Manuscrits de Stendhal » ;

– la base de données – reprenant différemment les données de transcription – et l'interface logicielle – développées par le laboratoire LIDILEM.

Cette séparation logique des données permet d'envisager une séparation physique, c'est-à-dire d'héberger sur différents serveurs les différentes sources de données.

Conclusions : un projet à long terme

La plateforme logicielle que nous avons présentée ne représente toutefois que la partie émergée de l'iceberg : il s'agit du travail nécessaire et préalable afin de permettre aux transcripteurs de travailler dans de bonnes conditions.

Grâce à la modélisation de document (DTD) et à la série de feuilles de style (utilisation du logiciel libre Morphon pour la saisie), les chercheurs peuvent déjà saisir des fiches, qui seront centralisées et vérifiées par un comité de validation. La phase de test et de perfectionnement de l'outil pour qu'il réponde au mieux aux besoins de saisie des chercheurs et aux besoins de consultation des utilisateurs (grand public ou chercheurs) est à l'heure actuelle quasiment terminée (version 1.0 mise à disposition en novembre 2007 sur le Wiki collaboratif de l'équipe). Depuis mars 2008, les chercheurs de l'équipe peuvent déposer directement leurs transcriptions (partielles ou définitives) sur le serveur dont le fonds s'enrichira de manière incrémentale. Par ailleurs, ce serveur servira de support de communication collaborative aux différents transcripteurs : ils pourront échanger points de vue, difficultés et solutions pour continuer la saisie des pages des registres qui n'ont pas encore été analysées.

1. Restructuration et édition

Le principe est celui d'une restructuration dynamique. Nous entendons par là l'extraction et l'ordonnancement de sous-arbres XML afin de constituer un nouveau texte. L'annotation des transcriptions, notamment la datation des unités de texte ainsi que l'attribution de ces unités de texte aux divers scripteurs, permet une restructuration du corpus en fonction de critères définis par l'utilisateur. Ainsi, une requête relativement simple permettra d'extraire l'ensemble des marginales d'un ou plusieurs registres des manuscrits et de les ordonner en fonction de leur date de rédaction.

Il est à noter toutefois que cette restructuration est automatisée et n'est donc pas parfaite. L'utilisateur a l'obligation de vérifier et éventuellement modifier le corpus restructuré. Cette restructuration se combine donc avec le module de catalogue personnalisé, la restructuration dynamique n'étant qu'une aide, qu'un premier pas vers la constitution de corpus littéraires ou linguistiques.

CLELIA est conçu pour permettre des éditions numériques et papier à la demande, et comme un outil au service de la génétique textuelle. En effet, le

module de catalogue personnalisé ainsi que celui de restructuration dynamique permettent de concevoir des ensembles ordonnés de pages ou d'extraits de pages conçus selon des critères émanant de l'enrichissement apporté par les transcripteurs…

Les catalogues personnels constituent en soi des éditions numériques que les utilisateurs inscrits peuvent mettre à la disposition des autres utilisateurs. Les données étant stockées en format XML, les affichages en modes linéaire et pseudo-diplomatique seront accessibles. À terme, il sera possible d'exporter ces catalogues afin qu'ils soient affichés de manière autonome (indépendamment de la plateforme) et par conséquent qu'y soit associée une feuille de style particulière afin de correspondre à un format d'affichage particulier.

Les feuilles de style permettent non seulement de définir un format d'affichage mais aussi de concevoir d'autres formats et notamment ceux d'impression et donc d'édition papier. En effet, la structuration XML des informations permet de produire des fichiers dans des formats réutilisables par les logiciels de traitement de texte (OpenDocument par exemple) ou par les éditeurs (PDF notamment).

Une application immédiate des potentialités de CLELIA est la préparation d'une édition des « Journaux et papiers » de Stendhal, projet actuellement en cours de réalisation en partenariat avec les ELLUG (Presses universitaires de l'Université Stendhal - Grenoble 3) qui récupéreront directement les fichiers XML, ce qui facilitera leur travail éditorial. La parution du premier tome de cette édition (correspondant à la période 1801-1804) est prévue fin 2010 ; et la publication devrait s'échelonner jusqu'en 2013 environ. Cette édition remettra en question les choix éditoriaux antérieurs, à commencer par celui, fait par Victor Del Litto, de démembrer les documents stendhaliens en obéissant à une double logique : le respect absolu de la chronologie, et la séparation entre textes jugés « intimes » et textes jugés « littéraires » (il s'agit de la célèbre et contestée distinction entre « Journal » et « Journal littéraire »[11]).

L'alimentation de la base à partir du travail sur les manuscrits permet en effet de développer toute une réflexion génétique sur les unités matérielles voulues par Stendhal (cahiers, liasses, etc., quand il est encore possible de les identifier) et sur les notions – souvent complexes chez Stendhal – de « documents » et de « corpus », qui remettent complètement en cause certains « textes » fournis par les éditeurs antérieurs – entre autres, *Pensées, Filosofia nova,* et bien sûr, *Journal littéraire* – et qui n'ont de réalité que comme créations éditoriales.

11. En réalité, l'examen des manuscrits révèle qu'un même paragraphe (voire une même phrase) peut contenir aussi bien des réflexions de caractère diariste que des considérations théoriques et esthétiques, et qu'il est par conséquent extrêmement difficile de les séparer en ensembles distincts.

De plus, la possibilité de reclassement virtuel à partir des données de transcription (analyse des papiers, identification des scripteurs, observations sur les pages – trous de couture révélant que telle feuille volante appartenait autrefois à une liasse, qu'il sera peut-être possible de reconstituer…–) permet également de rêver à une reconstitution au moins partielle d'un ordre initial perturbé peut-être déjà par Romain Colomb, le cousin et exécuteur testamentaire de Stendhal, mais surtout, à la fin du XIX^e siècle, par des bibliothécaires bien intentionnés mais maladroits. C'est donc une série de réflexions génétiques très riches que permet la constitution de cette base.

2. Un système modulaire

La conception de la plateforme est modulaire : les données sont séparées des traitements et nombre de traitements sont indépendants les uns des autres. Par conséquent, il est possible d'intégrer des greffons (*pluggins*) au fur et à mesure des besoins et des avancées du projet sans pour autant requérir une reconstitution complète de la suite logicielle.

Pour ce faire, un moteur de recherche avancé a été mis en place. L'un des éléments fondamentaux de la plateforme est le système de recherche. La structure XML que nous avons définie permet d'effectuer des recherches dans le texte, les commentaires scientifiques et en fonction des propriétés des feuillets (présence de biffes ou dimensions des feuillets).

Toutefois, comme tout un chacun a pu le constater, les moteurs de recherche, tout aussi perfectionnés soient-ils, exigent de l'utilisateur de trier les résultats parfois incongrus qu'il obtient. Cette déficience est notamment due à ce que les documents sont indexés par les graphies (c'est-à-dire les suites de caractères) et en fonction de propriétés statistiques simples, comme le fait par exemple Google[12].

Dans notre cas, nous ne sommes pas contraints par un volume de données très important (32 000 feuillets contre plusieurs dizaines de milliards de documents indexés par les grands moteurs de recherche) et nous pouvons donc utiliser des outils linguistiques performants pour améliorer la qualité des recherches sans que les processus de calcul soient perceptibles pour les utilisateurs.

Ainsi, l'intégration d'un système d'analyse syntaxique permettra d'enrichir les données transcrites avec les catégories syntaxiques des mots calculées en contexte ainsi que les lemmes correspondant à chacun des mots. À titre d'exemple pédagogique, une requête sur le terme « général » retournera toutes les occurrences du mot, fût-il substantif, adjectif ou adverbe, et l'utilisateur

12. L. Dekang, Google Inc., *Knowledge Acquisition from Text,* actes de la conférence TALN 2007, IRIT Press.

cherchant l'usage du terme militaire (et donc substantif) doit trier les résultats. Une requête précisant que le terme recherché est un substantif dans le système que nous implantons ne retournera que les occurrences voulues.

Il faut toutefois garder à l'esprit que les systèmes d'analyse linguistique automatique ont des taux de rappel et de décision[13] approchant les 95 %[14] et que par conséquent des erreurs apparaîtront.

Ce phénomène d'erreurs sera d'autant plus amplifié que les textes sont retranscrits tels quels et contiennent des fautes d'orthographe, de grammaire, des mots inconnus, des énoncés incomplets et éventuellement des phrases en langue étrangère (anglais et italien en particulier). Nous devons donc concevoir un outil d'analyse capable de s'adapter à ces variations. Nous nous inspirerons des travaux d'Agnès Souque[15] en matière d'analyse syntaxique robuste gauche-droite[16].

Par ailleurs, des travaux sont en cours afin de tenter d'apporter des solutions innovantes d'accès au contenu par cartographies du corpus. L'approche cartographique que nous avons déjà mise au point dans le domaine de la typologie de textes pour le projet CAPET[17] et dans le domaine de l'aide à la décision juridique pour le projet CATMInE[18] permet de visualiser un ensemble de documents en fonction de similarités définies par l'utilisateur. Par exemple, l'utilisateur pourra demander à visualiser l'ensemble des pages des manuscrits en fonction de la présence ou non d'un champ lexical qu'il définira. Grâce à la projection de Sammon[19], les pages seront affichées dans un nuage de points,

13. G. Adda, J. Mariani, P. Paroubek, M. Rajman, & J. Lecomte, «L'action GRACE d'évaluation de l'assignation des parties du discours pour le Français», *Langues*, 2 (2), 1999, p. 119-129.

14. Voir le site GRACE : Grammaires et Ressources pour les Analyseurs de Corpus et leur Évaluation, site institutionnel du consortium, 1995 : http://www.limsi.fr/TLP/grace.

15. A. Souque, «Conception et développement d'un formalisme de correction grammaticale automatique : application au français», mémoire de master Recherche Industries de la langue, Université Grenoble 3 - Stendhal, 2007, http://agnes.souque.free.fr/docs/Memoire.pdf.

16. Par analyse syntaxique robuste gauche-droite, nous entendons une analyse syntaxique résistant aux imprévus tels les fautes d'orthographe ou de grammaire et les énoncés tronqués (robustesse) et prenant en compte le fait qu'un écrit se rédige par unités linguistiques successives de gauche à droite (contrairement à la plupart des analyses syntaxiques qui considèrent la phrase comme une entité sécable en unités linguistique).

17. Th. Lebarbé, *Catégorisation A-syntaxique et Protocole d'Échange de Textes (CAPET)*, actes du séminaire Island, Université de Caen, 2002, http://users.info.unicaen.fr/~psluquet/Island/documents/Thomas-CAPET-ISLAND.ppt.

18. Th. Lebarbé, P. Breese, *Computer Assisted TradeMark Infringement Evaluation (*CATMInE*)*, 3rd American-French Conference on Technology and Legal Practice, Syracuse (USA), 2001. Voir le résumé de ce projet à cette adresse : http://www.greyc.unicaen.fr/anciens-evenements/seminaires-i3-2001/Lebarbe-resume.html.

19. J. W. Sammon, *A nonlinean mapping for data structure analysis*, IEEE Transactions on Computers, 18, 1969, p. 401-409.

celles qui auront les mêmes propriétés lexicales seront représentées par des points proches les uns des autres ; la distance sur la carte entre deux points (pages) sera représentative de leur similarité. Ces travaux seront mis à disposition du public sur CLELIA pour la rentrée universitaire 2008.

La réalisation de cette base documentaire est un projet extrêmement stimulant par sa transversalité, puisque, par cette réflexion sur un corpus double, à la fois littéraire et linguistique, il associe de manière originale des littéraires et des linguistes avec des perspectives de recherches à la fois complémentaires et intrinsèquement liées : de fait les transcriptions fournies par les littéraires constituent la base des recherches des linguistes, qui fourniront en retour des outils d'analyse et d'interprétation permettant aux littéraires de fonder leurs hypothèses sur la genèse de textes de Stendhal. Ce système de va-et-vient est au centre même du projet, fondé sur des intérêts réciproques.

La base documentaire CLELIA étant conçue comme une plateforme évolutive et collaborative, elle permet donc d'imaginer l'ajout ultérieur de fonctionnalités, selon les intérêts des chercheurs, mais aussi des adaptations à d'autres fonds de manuscrits.

Il existe un certain nombre de normes de description, à la fois dans le domaine du traitement automatique des langues et des textes – ainsi la TEI[20] – et dans le domaine de la gestion d'archives – tel le protocole OAI (Open Archives Initiative's Protocol for Metadata Harvesting[21]) – et de la gestion documentaire des bibliothèques – telle la norme Encoded Archival Description[22]. Toutefois, ces normes ont pour objectif d'homogénéiser des stockages d'informations génériques pour des objectifs spécifiques.

Dans le cadre du projet CLELIA, il aurait été possible de se conformer à ces modèles de représentation et de stockage. Nous avons ailleurs[23] critiqué cette approche qui consiste à contraindre un modèle que l'on conçoit (ici, le modèle de la page de manuscrit de Stendhal) par un modèle préexistant.

20. N. Ide, J. Véronis *The Text Encoding Initiative : Background and context*, Dordrecht, Kluwer Academic Publishers, 1995, 242 p.

21. M. Sévigné et F. Clavaud, « Vers des portails collaboratifs : le protocole OAI-PMH et les archives », *Culture et Recherche*, n° 103, octobre-novembre-décembre 2004, p. 20-21. http://www.culture.gouv.fr/culture/editions/rcr/cr103.pdf.

22. D. Pitti, *Encoded Archival Description, An Introduction and Overview*, *D-Lib Magazine*, vol. 5, n° 11, 1999.

23. Th. Lebarbé, « Hiérarchie inclusive des unités linguistiques en analyse syntaxique coopérative », thèse de Doctorat de l'Université de Caen, 2002. http://www.u-grenoble3.fr/lebarbe/uploads/Main/these-thomas-lebarbe-230502.pdf.

Notre objectif principal est de permettre à la fois aux transcripteurs de transcrire et aux utilisateurs d'accéder au contenu de transcription. Notre modélisation s'est donc orientée vers l'utilisateur. Ainsi, la structuration de l'information et la terminologie choisie répondent aux besoins de simplicité et de compréhensibilité exprimés par les différentes parties du projet (chercheurs littéraires, linguistes et informaticiens). Travailler ainsi nous a permis d'optimiser la collaboration interdisciplinaire, en développant un outillage informatique se conformant aux documents, plutôt qu'en contraignant le document et ses utilisateurs à un modèle inadéquat.

Cette approche n'empêche pas la constitution ultérieure d'outils d'extraction et de conversion afin que les données se conforment à telle ou telle norme si le besoin s'en fait sentir.