



Rapport de stage

Édition numérique de manuscrits modernes et
gestion du projet e-Man de l'ITEM



Période de stage : du 2 mars au 1^{er} juin 2015



Tuteur professionnel
RICHARD WALTER

Tutrice pédagogique
CÉCILE BOULAIRE

Jury de soutenance
CÉCILE BOULAIRE & CHRISTINE BÉNÉVENT

Résumé

Ce rapport de stage présente le projet e-Man de l'Institut des textes et manuscrits modernes au sein duquel j'ai effectué un stage de trois mois sous la responsabilité de Richard Walter. J'y expose les missions qui m'ont été confiées : la rédaction de documents pour présenter le projet, guider les contributeurs, ou donner des recommandations concernant la création d'une barre d'outils encodant en XML-TEI, ainsi que des recommandations en vue d'une prestation informatique.

Remerciements

Je suis très reconnaissante envers M. Marc Mézard, directeur de l'École normale supérieure et M. Richard Walter, mon tuteur professionnel et responsable du service de support aux projets numériques de l'ITEM, de m'avoir permis d'effectuer mon stage de fin d'études au sein de l'Institut des textes et manuscrits modernes.

Je tiens également à remercier une nouvelle fois Richard Walter, Claire Riffard, responsable de l'équipe « Manuscrits francophones », et Fatiha Idmhand, responsable de l'équipe « Manuscrits hispaniques », pour leur accueil chaleureux, nos échanges fructueux et leurs relectures très fines.

J'ai beaucoup appris lors de ce stage, où tous ceux que j'ai eu la chance de rencontrer et de côtoyer durant ces trois mois se sont montrés accueillants, disponibles et de bon conseil. Je tiens à citer M. Paolo D'Iorio, directeur de l'ITEM, M. Pierre-Marc De Biasi, précédent directeur de l'ITEM, spécialiste de Flaubert, de génétique artistique et de génétique des textes, Élise Anastase, responsable administrative de l'ITEM, Jean-Sébastien Macke de l'équipe Zola, Jérémy Pédrizzi, informaticien réseau de l'ITEM, Monique Sicard de l'équipe « Genèse de la photographie », Arghyro Paouri du pôle audiovisuel du CNRS, Sylvie Giraud de l'équipe Flaubert, Anne Garcia-Fernandez du Laboratoire d'anthropologie sociale, Agnès Tricoche du Labex TransferS, Annabelle Milleville du Labex TransferS, sans mentionner tant d'autres personnes qui m'ont éclairée, enrichie, conseillée durant ces quelques mois.

Enfin, je remercie l'équipe pédagogique du Master Patrimoine écrit et édition numérique du CESR de m'avoir offert une formation professionnalisante de grande qualité et de m'avoir offert la possibilité d'effectuer deux stages dans deux domaines passionnants, dont celui-ci dans le domaine des bibliothèques numériques.

Je n'oublie pas non plus les enseignants et responsables de mémoire qui m'ont encadrée et guidée pendant mes études d'angliciste et mes responsables de stage au Centre de recherche bretonne et celtique de Brest qui m'ont fait découvrir le monde des bibliothèques universitaires, des laboratoires de recherche, de l'édition et des bibliothèques numériques qui continuent de m'enthousiasmer.

Résumé	3
Remerciements	5
Introduction	9
Partie 1 : La structure d'accueil	11
1. L'Institut des textes et manuscrits modernes	13
a. <i>Un laboratoire CNRS et ENS</i>	13
b. <i>Un laboratoire de critique génétique</i>	13
c. <i>L'organigramme (non exhaustif) de la structure</i>	14
2. La plate-forme e-Man	15
a. <i>Une plate-forme Omeka</i>	15
b. <i>Deux équipes de recherche</i>	16
c. <i>Un service de support aux projets numériques</i>	16
3. Les autres projets de la plate-forme e-Man	16
a. <i>D'autres plates-formes dans e-Man</i>	16
b. <i>D'autres équipes, d'autres projets</i>	17
Partie 2 : L'édition numérique	19
1. Homogénéisation de la saisie dans e-Man	21
a. <i>Un nouveau guide de saisie</i>	21
b. <i>Faire connaissance avec la plate-forme</i>	21
c. <i>Se confronter à la gestion de projet numérique</i>	22
2. Élaboration d'une barre d'outils pour l'encodage des transcriptions	22
a. <i>D'une liste de boutons à un manuel d'encodage</i>	22
b. <i>Effectuer un état de l'art</i>	23
c. <i>Faire la synthèse des besoins et des points de vue</i>	23
3. Constitution du dossier génétique et définition des relations dans e-Man	24
a. <i>Comprendre les enjeux du dossier génétique et des relations</i>	24
b. <i>Effectuer une campagne de tests de plug-ins</i>	24
c. <i>Proposer des solutions exploitables informatiquement</i>	24

Partie 3 : La gestion de projet numérique	25
1. Travailler avec les acteurs de la critique génétique	27
a. <i>Le débat de la transcription diplomatique</i>	27
b. <i>Le débat de l'édition électronique ou de l'édition papier</i>	27
2. Travailler avec les acteurs des humanités numériques	28
a. <i>Former et se former</i>	28
b. <i>Former en travaillant par ateliers</i>	28
c. <i>Se former en travaillant par ateliers</i>	29
3. Être une interface entre le monde de la recherche et le monde du numérique	29
a. <i>Coordonner les différents acteurs d'un projet</i>	29
b. <i>Gérer plusieurs projets</i>	30
4. Communiquer	30
a. <i>Communiquer entre acteurs du projet</i>	31
b. <i>Donner de la visibilité au projet</i>	31
Conclusion	33
Annexes	35
<i>Annexe 1 : Présentation du projet e-Man</i>	35
<i>Annexe 2 : Guide de saisie des notices</i>	39
<i>Annexe 3 : Manuel d'encodage des transcriptions</i>	53
<i>Annexe 4 : Recommandations pour le traitement génétique des documents</i>	91
Liste des illustrations utilisées	99

APRÈS une licence d'anglais et un master de recherche en littérature anglophone à l'Université de Bretagne Occidentale, j'ai préparé l'Agrégation d'anglais et obtenu le CAPES d'anglais, grâce auquel j'ai pu faire l'expérience de l'enseignement auprès de collégiens. Cette expérience m'a ouvert les yeux sur la réalité de l'enseignement dans le secondaire et m'a donné envie de me reconverter dans un domaine en étroite relation avec les chercheurs ou enseignants-chercheurs. En effet, je souhaitais travailler au service de la recherche sans être moi-même enseignant-chercheur. J'ai alors demandé à effectuer un stage de découverte de deux semaines au sein du Centre de recherche bretonne et celtique à Brest. Durant ce court stage, j'ai eu un aperçu de tous les métiers qu'un tel laboratoire CNRS pouvait recouvrir : une bibliothèque universitaire possédant des fonds d'archives et ayant des initiatives de numérisation de ses fonds ; une cellule éditoriale préparant les ouvrages d'auteurs pour l'impression et un travail avec les chercheurs ayant un projet de création de base de données. Ce premier stage m'a confirmée dans le souhait de me former à ces métiers de valorisation de la recherche et m'a amenée à postuler au Master Patrimoine écrit et édition numérique du CESR.

Deux domaines m'ont alors attirée plus que les autres : le travail éditorial en presses universitaires et les humanités numériques, qui visent tous deux à produire une édition de textes (un livre, papier ou numérique, dans le premier cas ; une base de données ou une bibliothèque d'archives numérique dans l'autre). C'est assez logiquement que j'ai choisi d'effectuer un stage dans chacun de ces domaines : le premier à l'Institut des textes et manuscrits modernes, unité mixte de recherche du CNRS et de l'ENS à Paris, et le second aux Presses universitaires François-Rabelais de Tours.

Ce rapport présente le stage que j'ai effectué à l'Institut des textes et manuscrits modernes sous la responsabilité de Richard Walter et sur le budget du Labex TransferS de l'E.N.S. La mission qui m'a été confiée concernait la « création de fonctionnalités et la gestion de projets de la plate-forme e-Man en ligne (bibliothèque de manuscrits francophones africains et hispaniques) ».

La structure d'accueil

1. L'Institut des textes et manuscrits modernes

a. *Un laboratoire CNRS et ENS*

L'Institut des textes et manuscrits modernes est une unité mixte de recherche du Centre national de recherche scientifique et de l'École normale supérieure. L'ITEM est situé rue Pouchet dans le 17^e arrondissement de Paris. Il a déménagé de la rue d'Ulm, dans 5^e arrondissement de Paris, où il se trouvait originellement. Il a néanmoins conservé une antenne à l'ENS, dans la rue d'Ulm, et possède une autre antenne à la Maison des Sciences de l'Homme de Poitiers avec laquelle l'ITEM a un partenariat.

En tant qu'unité mixte de recherche du CNRS, l'ITEM est composé d'une centaine d'ingénieurs de recherche, d'ingénieurs d'études, d'assistants-ingénieurs et de chercheurs. L'ITEM relève de l'ENS en tant que laboratoire d'excellence ou Labex TransferS. Le Labex TransferS finance de nombreuses initiatives de l'ITEM. Par exemple, mon stage a été financé intégralement par le Labex TransferS.

b. *Un laboratoire de critique génétique*

Comme l'explique le texte de présentation de l'ITEM, la spécificité de cet institut est l'étude de la genèse artistique (dramaturgique, littéraire, des arts visuels, etc.) à partir d'archives et de brouillons d'écrivains, autrement dit :

l'étude du processus de création (artistique, littéraire, dramaturgique) à partir de la comparaison des différents témoins qui attestent des étapes de la genèse de l'œuvre (manuscrits ou tapuscrits, corrections, ébauches, brouillons, sources extérieures à l'œuvre – exogénétiques – notes, correspondances...). L'ensemble de ces éléments constitue le dossier génétique de l'œuvre.¹

Étudier le processus de création à partir d'archives d'écrivains est la vocation de l'ITEM depuis sa création : en effet, cet institut a été créé après le don au CNRS des archives du poète Louis Aragon. Il existe désormais une équipe de recherche qui étudie les manuscrits d'Aragon.

L'ITEM est actuellement dirigé par Paolo D'Iorio, spécialiste de Nietzsche. Il dirige depuis 2006 le projet HyperNietzsche de constitution d'une base de données de manuscrits de Nietzsche et de textes critiques sur l'auteur allemand.

Pierre-Marc de Biasi, le précédent directeur de l'ITEM, dirige l'équipe Flaubert qui étudie la genèse des œuvres de Flaubert. Il est également à la tête de l'équipe « Génétique des arts visuels », ou encore, en matière de projet numérique, du projet « Données internationales de génétique artistique » (DIGA). Il a écrit *Génétique des textes*, texte qui fait référence en matière de critique génétique au sein de l'ITEM et, internationalement, dans le domaine de la critique génétique.



1 Voir le texte de présentation d'e-Man en annexe 1, p. 1.

c. L'organigramme (non exhaustif) de la structure

Institut des textes et manuscrits modernes (ITEM)
Unité mixte de recherche (UMR) – UMR 8132 CNRS-ENS
59-61 rue Pouchet 75017 Paris

Directeur de l'ITEM
Paolo D'ITORIO

Directrice adjointe de l'ITEM
Aurèle CRASSON

Responsable administrative
Élise ANASTASE, ITA

Services numériques et
informatiques

Pôle « Humanités numériques »
Richard WALTER, ITA (IR)

Pôle « Humanités numériques »
Informaticien réseau
Jérémy PÉDRAZZI, ITA

Équipe « Manuscrits francophones »

Responsable de l'équipe
Claire RIFFARD

Équipe « Manuscrits hispaniques »

Responsable de l'équipe
Fatima IDMHAND, MCF HDR

Équipe Zola

Projet numérique « ArchiZ pour les
archives zoliennes

Jean-Sébastien MACKIE, docteur ès
lettres (associé)

Équipe Flaubert

Projet numérique « Flaubert et
le pouvoir des images » (FLIM)

Directeur de l'équipe
Pierre-Marc de BIASI, DR

Sylvie GIRAUD, post-doctorante

Équipe « Génétique des arts
visuels »

Projet numérique de « Données internationales de
génétique artistique » (DIGA)

Directeur de l'équipe
Pierre-Marc de BIASI, DR

Monique SICARD, CR

Travail éditorial

Revue *Genesis*
Presses de l'Université Paris-Sorbonne

Collection Planète Libre
Éditions CNRS

CNRS

DR : directeur de recherche

ITA : Ingénieurs, Techniciens, Assistants-ingénieurs

T : technicien / AI : assistant-ingénieur / IE : ingénieur d'études / IR : ingénieur de recherche

Chercheur CNRS : CR : chargé de recherche / DR : directeur de recherche

Université

MCF : maître de conférence

HDR : habilitation à diriger des
recherches

Un organigramme plus complet est disponible sur le site de l'ITEM à l'adresse suivante :
http://www.item.ens.fr/docannexe/file/578652/org_2012_aeres_def.pdf.

2. La plate-forme e-Man

a. Une plate-forme Omeka

E-Man est construit à partir du logiciel Omeka, développé par le Roy Rosenzweig Center for History and New Media de la George Mason University aux États-Unis. Ce logiciel est un CMS (*Content Management System*, ou Système de gestion de contenus) auquel on peut ajouter des modules d'extension (plug-ins). Il permet de gérer des bibliothèques numériques. C'est un logiciel *open source*, c'est-à-dire qu'il est gratuit et libre de droits et que son code est disponible et adaptable.

E-Man est-il une plate-forme ou un site internet ? Cette question, somme toute anodine, est intéressante en matière de gestion de projet. Le terme « plate-forme » est un terme technique qui fait référence au logiciel Omeka. Le terme « site », plus générique, peut faire référence à de nombreuses formes d'objets : des plus simples, comme les pages personnelles, aux plus complexes comme des bases de données. Le terme « site », bien que plus vague, est donc également plus communément employé. Par conséquent, « site » s'est avéré un terme plus engageant pour présenter e-Man aux participants au projet qui devront utiliser la plate-forme e-Man comme outil de recherche, par exemple auprès des chercheurs algériens, associés au projet « Manuscrits francophones », auxquels Claire Riffard a présenté la plate-forme en avril dernier. En revanche, le terme « plate-forme » reste le plus approprié pour mettre en évidence les spécificités d'e-Man liées à l'utilisation d'Omeka.

L'intérêt d'utiliser Omeka, dès l'origine du projet, résidait dans la facilité d'importer le nombre élevé de pages de manuscrits numérisées à disposition, et ce sans avoir à construire une base de données de toutes pièces, ce qui aurait nécessité l'intervention d'un développeur web². L'autre intérêt qu'offrait d'emblée Omeka était de contraindre les membres du projet à employer un vocabulaire Dublin Core pour renseigner les métadonnées, ce qui permettrait de les rendre moissonnables, par exemple, par les moteurs Isidore et Europeana, et donc de leur garantir une meilleure visibilité. En effet, le Dublin Core est un standard international de métadonnées défini en 1995 par un groupe de travail dans la ville de Dublin, aux États-Unis. Il est employé par de nombreuses institutions (comme l'Organisation Mondiale de la Santé).

La plate-forme e-Man est, en définitive, un site internet qui regroupe plusieurs sites internet. On peut donc considérer que c'est un portail de sites. En effet, quand on arrive sur la plate-forme e-Man, on découvre une page où ne figurent que les noms des différentes plates-formes Omeka auxquelles on peut accéder, et qui correspondent à des projets distincts.

Ci-contre, l'interface d'accueil d'e-Man³ au 16 juin 2015. L'accueil d'e-Man comporte désormais quatre projets distincts : les projets « Manuscrits francophones », « Manuscrits hispaniques », « Flaubert et le pouvoir des images » (FLIM) depuis mars 2015, ainsi que le projet « Cartographie des fonds d'archives de manuscrits francophones » (Cartomac) depuis juin 2015.



2 Un prestataire informatique a néanmoins modifié le thème d'e-Man.

3 Voir <http://eman-archives.org/>.

b. Deux équipes de recherche

16

J'ai rédigé, à partir d'un long texte de présentation, un texte synthétique de deux pages pour présenter le projet e-Man téléchargeable depuis la plate-forme dans un onglet « En savoir plus »⁴, visible dès l'interface d'accueil d'e-Man. La rédaction du texte de présentation m'a permis de découvrir l'histoire du projet, et plus généralement celle de l'ITEM et des équipes « Manuscrits francophones » et « Manuscrits hispaniques », respectivement dirigées par Claire Riffard et par Fatiha Idmhand.

De fait, le projet e-Man est l'aboutissement de ces deux projets. Le texte de présentation du projet e-Man fait l'historique⁵ de ses équipes fondatrices: l'équipe « Création d'outils pour l'exploitation de manuscrits hispaniques », ou Chispa (devenue depuis « Manuscrits hispaniques »), dirigée par Fatiha Idmhand, et l'équipe « Manuscrits francophones », dirigée par Claire Riffard.

c. Un service de support aux projets numériques

L'ITEM possède un service de support aux projets numériques. Richard Walter s'occupe de ce service de l'ITEM, et, plus particulièrement, des projets créés à partir du logiciel Omeka. Jérémie Pédrizzi, informaticien réseau à l'ITEM, s'occupe à la fois du serveur de l'ITEM et de la branche WordPress⁶ de l'aide aux projets numériques. Richard Walter et Jérémie Pédrizzi assurent des formations aux outils informatiques et aux humanités numériques. Par exemple, pendant mon stage à l'ITEM, Richard Walter a assuré deux formations Omeka financées par le CNRS et destinées aux personnels du CNRS. J'ai pu participer à l'une de ces formations au début de mon stage afin de découvrir les possibilités et les limites du logiciel Omeka.

3. Les autres projets de la plate-forme e-Man

a. D'autres plates-formes dans e-Man

La plate-forme e-Man a accueilli et doit accueillir de nombreux autres projets, tels que :

☞ « Flaubert et le pouvoir des images » (FLIM), dans e-Man depuis mars 2015 ;

☞ « L'Affaire Dreyfus à travers la collection Mitterand », à venir ;

☞ « Sources génétiques de l'opéra », à venir ;

☞ « Cartomac : réseaux d'écrivains et fonds d'archives », dans e-Man depuis juin 2015 ;

☞ « Aragon et Les Lettres françaises », à l'étude.

Chacune de ces plates-formes bénéficiera de la visibilité offerte par e-Man, mais sera gérée de manière distincte et autonome. En effet, le logiciel Omeka ne permet pas pour l'instant de gérer en même temps plusieurs plates-formes différentes de manière automatique.

4 Voir la capture d'écran de l'interface d'accueil d'e-Man supra. Le texte de présentation d'e-Man est disponible à l'adresse suivante : http://eman-archives.org/En_savoir_plus_Presentation_e-Man.pdf. Ce texte de présentation, à la rédaction duquel j'ai participé, est reproduit en annexe de ce rapport, en annexe 1.

5 Voir la partie « Historique du projet » du texte de présentation, en annexe 1, p. 1-2.

6 WordPress est un CMS permettant de construire des sites internet.

b. D'autres équipes, d'autres projets

L'équipe Flaubert est notamment composée de Sylvie Giraud. De passage dans les locaux de l'ITEM, celle-ci m'a présenté le projet « Optima » qui a permis de développer le logiciel Transcript. Celui-ci est malheureusement à l'arrêt depuis que le projet ANR, financé par l'Agence nationale pour la recherche, a pris fin en 2013. Comme je l'explique dans le Manuel d'encodage présent en annexe de ce rapport¹, l'encodage que devait permettre ce logiciel se faisait par zones de textes. Si nous n'avons pas gardé cette idée dans le projet d'encodage des manuscrits francophones et hispaniques, la liste établie par Sylvie Giraud, résumant les balises utiles pour l'équipe Flaubert, a servi de base de réflexion pour établir une première liste, qui s'est ensuite étoffée jusqu'à devenir un véritable manuel.

L'équipe Zola est, quant à elle, notamment composée de Jean-Sébastien Macke. Ce dernier a réalisé dans un premier temps, mais sans appui institutionnel, un site internet entièrement dédié à l'œuvre de Zola.² C'est à partir de l'expérience acquise grâce à ce site très apprécié des enseignants pour ses qualités didactiques, et à partir de l'iconothèque zolienne constituée dès les années 1980 par l'équipe Zola, que le projet ArchiZ a été élaboré³.

Enfin, l'équipe Cartomac, composée de Claire Riffard et de Rym Khene (doctorante), et bénéficiant de l'expertise d'Agnès Tricoche et du cartographe Julien Caverro, du Labex TransferS, vise à cartographier la présence des risques géopolitiques et naturels liés à la conservation des manuscrits du continent africain.

1 Voir la partie « Génétique des textes » du manuel d'encodage en annexe 3, p. 20.

2 Voir <http://www.archives-zoliennes.fr/>.

3 Voir <http://www.ezola.fr/>.

L'édition numérique

1. Homogénéisation de la saisie dans e-Man

a. *Un nouveau guide de saisie*

L'une des premières missions qui m'aient été confiées était l'homogénéisation de la saisie des notices de documents dans e-Man. Un premier guide avait été rédigé. Il m'a été demandé de l'améliorer. Ce guide était destiné aux contributeurs amenés à rédiger des notices de documents dans e-Man.

J'ai commencé par effectuer une relecture du document afin d'éliminer les dernières coquilles. Il m'avait été demandé, non seulement de mettre en page le document, mais de guider davantage le lecteur. J'ai donc rédigé une introduction afin que l'entrée en matière soit plus progressive. J'ai prolongé cette démarche afin de guider pas à pas le lecteur de la plate-forme générale à la plate-forme auteur, et de se connecter à l'interface administrateur afin de saisir des notices. Dans un but pédagogique, j'ai également illustré mes propos à l'aide de captures d'écran⁷.

J'ai inclus, dans le guide de saisie, des éléments qui n'y figuraient pas auparavant : des recommandations sur le choix des mots-clés à renseigner dans la notice Dublin Core, que j'ai accompagnés d'un exemple de nuage de mots-clés significatif⁸. J'ai également inclus le tableau comparatif des utilisateurs que j'avais construit après avoir fait des tests successifs, en m'identifiant tour à tour comme « super (administrateur) », « administrateur », « contributeur » puis « chercheur »⁹. Ce sont, en effet, les quatre statuts que propose Omeka.

Enfin, j'ai réorganisé le document afin de le rendre complet, d'en éliminer les redondances et d'y ajouter une table des matières pour plus de maniabilité. En outre, j'ai utilisé, pour la rédaction des quatre documents en annexes, un traitement de texte (OpenOffice) plutôt qu'un logiciel de Publication assistée par ordinateur (InDesign) afin que tous ces documents puissent être modifiés¹⁰ après mon départ.

b. *Faire connaissance avec la plate-forme*

Remanier le guide de saisie des notices m'a donné l'occasion de tester l'utilisation de la plate-forme du point de vue du rédacteur de notices. J'ai pu me rendre compte du travail que représentait leur rédaction : les quinze principaux champs du Dublin Core (qui composent le Dublin Core dit « simple » et qui seuls seront interrogés par les moissonneurs OAI-PMH) sont complétés par d'autres champs spécifiques au type de contenu auquel se rattache chaque document. Ces champs spécifiques composent le Dublin Core dit « étendu ». Or, les champs du Dublin Core simple et du Dublin Core étendu ne sont pas renseignés au même endroit, ce qui complique la tâche du rédacteur de la notice.

Les tests que j'ai effectués m'ont permis de remarquer d'autres limites à l'utilisation d'Omeka : ainsi, il est très facile de supprimer une notice définitivement par inadvertance. C'est la raison pour laquelle j'insiste particulièrement sur ce point dans le document. Par ailleurs, une autre contrainte, moins dangereuse, mais extrêmement fastidieuse est que, dans plusieurs situations, le rédacteur doit effectuer la même action à deux reprises : par exemple, dans le cas du choix des mots-clés, le rédacteur de la notice doit les indiquer dans le champ

7 Voir, par exemple, les pages 2 et 3 du guide de saisie, en annexe 2.

8 Voir la page 14 (dernière page) du guide de saisie, en annexe 2.

9 Voir la page 4 du guide de saisie, en annexe 2.

10 En effet, les membres de l'ITEM ne sont pas initiés à l'utilisation d'un logiciel de PAO et utilisent majoritairement le logiciel Microsoft Word pour tous leurs travaux de mise en page (y compris de la revue *Genesis*).

« sujet » du Dublin Core, et, s'il souhaite permettre la création d'un nuage de mots-clés, il doit copier-coller chaque mot-clé dans le champ « tags ». En effet, la création d'un nuage de mots-clés est permise par l'ajout d'un plug-in d'Omeka, c'est-à-dire un module d'extension que l'on a téléchargé pour compléter les fonctions de base d'Omeka. Ce plug-in ajoute ainsi un onglet distinct de celui du Dublin Core simple auquel appartient le champ « sujet ». Les doubles manipulations sont une réelle entrave au travail du rédacteur de notices parce qu'elles rendent encore plus complexe une tâche déjà longue et minutieuse. Les plug-ins, censés faciliter la vie par leur simplicité d'installation, s'avèrent donc être parfois une réelle limite d'Omeka.

c. Se confronter à la gestion de projet numérique

La rédaction du guide de saisie m'a fait prendre conscience de ce qu'impliquaient les statuts d'utilisateurs dans e-Man. En effet, le statut de « super (administrateur) » permettant d'effacer toutes les notices de la plate-forme, on ne peut accorder ce statut qu'aux personnes en qui l'on a une entière confiance et qui ne risquent pas de faire une mauvaise manipulation informatique. Le statut d'administrateur simple est très proche du statut de super administrateur, en ce qu'il permet également de modifier ou supprimer tout contenu. Mais il ne permet pas d'installer des plug-ins ou de modifier les paramètres. L'administrateur simple n'a pas non plus le droit de créer des comptes. Or, puisque tout membre décideur du projet n'est pas censé avoir accès à toutes les fonctionnalités de gestion informatique de la plate-forme, il peut être handicapant, pour lui, de ne pas pouvoir créer un compte pour un nouveau participant.

Par ailleurs, la différence entre le statut de chercheur et le statut de contributeur réside dans le fait que le contributeur n'a pas accès aux documents en accès privé, tandis que le chercheur a accès à tous les contenus, qu'ils soient marqués comme « privés » ou « publics ». Ainsi, dès qu'un chercheur, participant au projet e-Man, souhaite accéder à un contenu dont l'accès est restreint, il doit posséder un statut de « chercheur ». Il peut ainsi accéder non seulement à son propre contenu mais également à tous les autres contenus marqués comme « privés ». La manière dont sont gérés les statuts dans Omeka pose donc un réel problème. En effet, il n'existe, pour l'instant, aucun plug-in qui soit capable de donner accès à un seul contenu privé.

2. Élaboration d'une barre d'outils pour l'encodage des transcriptions

a. D'une liste de boutons à un manuel d'encodage

Je savais, en commençant mon stage, que mon rôle serait de réfléchir à l'encodage des transcriptions des manuscrits dans e-Man. Néanmoins, ma priorité, pendant tout le premier mois de mon stage, était, d'une part, de comprendre comment fonctionnait Omeka et, d'autre part, de comprendre les enjeux de la transcription. La formation sur Omeka et la journée d'études sur Mohammed Dib auxquelles j'ai assisté les 12 et 13 mars m'ont fait beaucoup avancer. Une fois terminée la rédaction des deux premiers textes qu'il m'avait été demandé de retravailler (le texte de présentation d'e-Man et le guide de saisie des notices), j'ai pu me consacrer à la réflexion sur les « balises-boutons » pour encoder les transcriptions dans e-Man. Comme mentionné précédemment, la liste de phénomènes à encoder, établie par Sylvie Giraud de l'équipe Flaubert, m'a fourni une base de travail à adapter au projet des équipes « Manuscrits francophones » et « Manuscrits hispaniques ». Cependant, cette première liste nécessitait d'être enrichie afin de donner lieu à un texte de recommandations, plus complet :

il fallait donc non seulement y ajouter des éléments précisant la politique éditoriale, les contraintes liées aux outils, mais également faire de ce texte un manuel explicitant la méthodologie à suivre. De plus, afin de se distinguer des manuels déjà existants, nous devons apporter des réponses spécifiques à l'encodage génétique des textes.

b. Effectuer un état de l'art

Le manuel d'encodage en annexe 3 présente l'état de l'art que j'ai effectué, avec l'aide des responsables du projet, pour rédiger le manuel d'encodage XML-TEI des transcriptions dans e-Man. Nous avons commencé par explorer la possibilité d'utiliser un plug-in d'Omeka pour la maniabilité qu'offre un tel outil. Il existe un plug-in d'Omeka dédié à la transcription collaborative. Nous avons donc choisi de l'installer sur e-Man afin d'effectuer des tests. La procédure d'installation s'avérant finalement plus compliquée que pour les autres plug-ins déjà installés sur e-Man, nous n'avons pu effectuer les premiers tests que la veille de la réunion où nous souhaitions faire une démonstration à l'équipe.

Comme je le développe dans le manuel, le plug-in Scripto d'Omeka fonctionne à partir d'un wiki de transcription. Ainsi, l'une des difficultés rencontrées résidait en l'incapacité de l'outil à prendre en charge le langage XML-TEI; de plus, ce plug-in nous obligeait à passer par le système MediaWiki en plus de la plate-forme Omeka, ce qui entraînait une complexité de gestion importante. Néanmoins, la découverte d'une barre d'outils de transcription en XML-TEI développée par le projet Bentham de l'University College London, laquelle passe également par MediaWiki, nous a incités à conserver cette solution comme une possibilité envisageable.

Une nouvelle solution s'est offerte à nous: il était également possible de passer directement par Omeka en utilisant la barre d'outils « Tiny MCE » intégrée au logiciel. L'équipe de l'atelier Omeka est parvenue à intégrer deux boutons à la barre d'outils existante: le bouton « barrer », ainsi qu'une palette de couleurs. Cependant, il semblerait que le nombre de boutons intégrables soit limité. En définitive, le choix entre les deux solutions ne pourra revenir qu'à un développeur informatique, plus à même d'envisager et de comparer les bénéfices et les contraintes propres à chacune d'entre elles.

J'ai toutefois tenté de synthétiser ces avantages et inconvénients en un tableau comparatif figurant en annexe de ce rapport¹¹.

c. Faire la synthèse des besoins et des points de vue

Dès le début, Fatiha Idmhand, maître de conférences à l'université de Dunkerque, a pris une part active à la rédaction du manuel. Dans un premier temps, c'est elle surtout qui s'est chargée de relire le document en cours d'élaboration. Au terme de deux semaines de travail, j'ai donc présenté le texte à Richard Walter. La relecture effectuée par celui-ci l'a convaincu de l'intérêt de l'encodage en XML-TEI pour notre projet, encodage qui nécessiterait l'appel à un prestataire mais constituait un bénéfice certain. Dès lors, il m'a demandé de développer la liste des balises XML-TEI afin que notre encodage soit le plus exhaustif possible. Il importait en effet d'anticiper, dès cette étape, les besoins ultérieurs.

Après de nombreuses relectures effectuées par Fatiha Idmhand et Richard Walter, Claire Riffard m'a à son tour fait part de ses besoins spécifiques d'encodage concernant la dimension génétique du projet. Ces remarques

allaient figurer dans le manuel, certaines définitivement, d'autres temporairement, ces dernières donnant lieu à une partie des recommandations pour le traitement génétique des documents dans e-Man figurant en annexe 4 de ce rapport.

Ainsi, l'élaboration du manuel d'encodage a nécessité de passer par sept versions successives, au gré des différents apports de mes trois responsables.

3. Constitution du dossier génétique et définition des relations dans e-Man

a. *Comprendre les enjeux du dossier génétique et des relations*

Le dossier génétique d'une œuvre est l'ensemble des manuscrits rédigés par un auteur avant l'impression de la première édition définitive. Dans ce dossier génétique figurent également les sources extérieures à l'œuvre dont l'auteur s'est servi pour la rédaction de l'œuvre en question. Fatiha Idmhand et Claire Riffard ont rédigé le texte scientifique concernant la constitution du dossier génétique pour le projet e-Man.

b. *Effectuer une campagne de tests de plug-ins*

Les relations entre documents à travers des liens dynamiques sont un enrichissement spécifique que permet l'outil informatique. Si l'édition papier permet d'effectuer des liens intellectuels entre les documents, l'édition électronique permet quant à elle de montrer l'intégralité des manuscrits et d'établir des liens entre eux. Dans le cadre du projet e-Man, l'outil informatique qui doit prendre en charge le dossier génétique et les relations entre documents est le logiciel Omeka. Plus précisément, nous avons décidé d'utiliser deux plug-ins d'Omeke pour réaliser ces opérations : le plug-in « Collection Tree » pour organiser le dossier génétique, et le plug-in « Item Relations » pour établir des liens dynamiques entre les documents.

Pour définir comment nous allons utiliser ces deux plug-ins, j'ai effectué une campagne de tests sur la plate-forme Omeka créée par Richard Walter pour les formations qu'il organise auprès de bibliothécaires, documentalistes et archivistes sur l'utilisation de ce logiciel.

c. *Proposer des solutions exploitables informatiquement*

Les recommandations que j'ai rédigées pour l'encodage en XML-TEI des transcriptions à l'aide d'une barre d'outil¹² s'adressaient à tous les membres du projet, y compris un éventuel prestataire informatique. En revanche, les recommandations concernant le dossier génétique et les relations entre documents dans e-Man¹³ sont destinées essentiellement à l'informaticien auquel le responsable « humanités numériques » du projet, Richard Walter, fera appel à l'été 2015. Aussi importait-il que ces documents soient tout deux lisibles dans une perspective de développement informatique.

12 Voir le manuel d'encodage en annexe 3.

13 Voir les recommandations en annexe 4.

La gestion de projet numérique

1. Travailler avec les acteurs de la critique génétique

a. *Le débat de la transcription diplomatique*

La réunion du 24 mars dernier avait pour objectif de faire le point sur l'avancée du projet avec tous les acteurs concernés. Étaient notamment conviés le directeur de l'ITEM, Paolo D'Iorio, ainsi que Pierre-Marc de Biasi. J'y ai présenté une première liste de balises, première étape de la rédaction de l'actuel manuel d'encodage. Le titre du document annonçait une « liste de balises-boutons pour une transcription diplomatique dans e-Man. » Une transcription diplomatique se distingue d'une transcription littérale en ce qu'une transcription littérale ne tient pas compte de la place des mots sur la page, tandis qu'une transcription diplomatique prend en considération les spécificités physiques du texte. Ainsi, le titre est par exemple saisi à la même place que sur le manuscrit (on le présente donc centré s'il est centré sur la page du manuscrit) ; on effectue également un retour à la ligne à l'endroit où l'auteur est passé à la ligne sur le manuscrit, etc. Cette méthode est surtout intéressante quand on transcrit un manuscrit, mais elle peut aussi être utilisée quand on transcrit un document imprimé sur lequel il y a eu des corrections manuscrites : passer à la ligne là où le texte original va à la ligne permet de se repérer sur la page et de pouvoir placer les ajouts et le texte barré.

La définition de la transcription diplomatique ne faisait pas consensus parmi les participants à la réunion, certains considérant qu'on n'allait pas assez loin et qu'il fallait envisager une transcription « ultra-diplomatique ». Cette dernière a l'ambition de respecter scrupuleusement l'apparence de la page de manuscrit. Le projet Optima, que je décris rapidement dans le manuel d'encodage en annexe 3, a développé une méthode d'encodage imitant la page des manuscrits de Flaubert grâce à un encodage par zones de texte. Nous avons choisi de ne pas adopter cet encodage par zones qui est plus contraignant qu'un encodage classique. Nous avons donc préféré éviter de qualifier notre projet d'encodage de « diplomatique » afin de ne pas imposer à notre projet des contraintes liées à d'autres projets, préférant nous concentrer sur les besoins effectifs des deux projets à l'origine d'e-Man, à savoir les projets « Manuscrits francophones » et « Manuscrits hispaniques ».

b. *Le débat de l'édition électronique ou de l'édition papier*

La journée d'étude sur Mohammed Dib du 13 mars dernier avait notamment pour objet d'étudier des manuscrits de cet auteur : ceux-ci sont en cours de numérisation par la BnF avant leur édition critique. Les contributeurs du projet étaient réunis : Daniel Ferrer (chercheur à l'ITEM), chercheurs algériens et français, écrivains, doctorants, post-doctorants. La journée d'études était organisée par Claire Riffard. Cette journée s'est révélée riche en débats sur le format de publication et d'édition du prochain ouvrage collectif de critique génétique sur Mohammed Dib. En effet, les propositions de solutions ont abondé : certains suggéraient de choisir le livre papier pour y transcrire le texte de l'édition définitive en regard d'un état antérieur du texte et de proposer une critique génétique en notes. D'autres préconisaient plutôt de rediriger le lecteur vers le manuscrit numérisé en ligne une fois que celui-ci serait disponible sur le site Gallica.fr de la BnF, puis sur e-Man. Cette solution nécessiterait un encodage en XML-TEI des différents états du manuscrit avant d'édition du texte. L'édition papier ne serait alors là que comme édition critique de l'édition définitive du texte avec notes génétiques en bas de page. Cette solution reviendrait à conserver le format actuel de la collection Planète libre de CNRS éditions et de proposer une visualisation enrichie des manuscrits en ligne grâce à un encodage en XML-TEI. Le débat est toujours vif.

2. Travailler avec les acteurs des humanités numériques

a. *Former et se former*

Les responsables de projets numériques ont un réel rôle d'auto-formation et de formation : il leur faut se tenir informés de l'évolution des humanités numériques, et ils échangent avec profit avec les responsables d'autres projets numériques afin d'envisager des solutions à des problèmes complexes. Ils ont également pour mission de former leurs équipes aux outils employés et de les familiariser aux enjeux et aux limites de ceux-ci. En outre, leur rôle de formateur peut également s'étendre à un public bien plus large : ainsi, au cours de mon stage, Richard Walter a non seulement donné des formations Omeka pour les personnels CNRS en France, mais également présenté en Uruguay notre plate-forme Omeka, et plus généralement les humanités numériques, à un public de chercheurs extrêmement demandeur de l'université de Montevideo, partenaire du projet « Manuscrits hispaniques » d'e-Man.

b. *Former en travaillant par ateliers*

Des ateliers informatiques ont été mis en place à l'ITEM depuis deux ans : tous les premier et troisième mardis du mois, le personnel CNRS (mais également un public plus large) peut assister à un « café gourmand » informatique : autour d'un café, un intervenant présente rapidement (pendant une heure) le fonctionnement d'un outil utile à la recherche en Sciences humaines et sociales. Par exemple, les « cafés gourmands » auxquels j'ai assisté pendant mon stage ont eu successivement pour thèmes :

- ☞ « Des outils d'organisation : Agora et MediaWiki » ;
- ☞ « Les cartes heuristiques : des outils pour organiser ses idées » ;
- ☞ « La politique de sauvegarde des données : quelques outils pour ne pas perdre de fichiers » ;
- ☞ « Gérer et travailler sur des fichiers vidéo : des outils de montage vidéo » ;
- ☞ « Sondage et agenda : des outils pour faire des sondages ».

J'ai très concrètement pu apprendre à me servir d'outils fort utiles pour mon stage : j'ai découvert le fonctionnement d'un wiki, ce qui m'a permis de tester le fonctionnement du plug-in Scripto de transcription collaborative d'Omeka ; j'ai été initiée à l'utilisation du logiciel Agora pour créer un environnement numérique de travail ; je me suis servie de l'outil de sondage pour demander la disponibilité des participants de l'atelier Omeka que l'on souhaitait organiser et pour lequel nous n'avions pas encore pu fixer de date. Enfin, la réalisation de cartes heuristiques m'a permis de représenter visuellement des comparaisons entre outils sans passer en premier lieu par la rigidité d'un tableau¹⁴.

Par ailleurs, une fois par mois en moyenne, le « café gourmand » se prolonge par un séminaire technique d'une heure et demie où un intervenant fait un retour d'expérience sur un projet numérique. J'ai donc eu la chance,

14 Voir les cartes heuristiques que j'ai réalisées aux adresses suivantes : <https://framindmap.org/c/maps/46518/public> (« Comparaison entre deux outils d'édition : le wiki et le carnet de recherche ») et <https://framindmap.org/c/maps/47127/public> (« Comparaison entre deux outils de transcription collaborative : le projet Bentham (en TÊI) et la combinaison de la barre d'outils d'Omeka TinyMCE et du plug-in Scripto d'Omeka »).

pendant mes trois mois de stage, d'assister à trois séminaires techniques :

☞ une présentation le 17 mars 2015, par Françoise Rubellin de l'Université de Nantes, du projet « Parodier l'opéra » ayant abouti à la base Theaville¹⁵ ;

☞ une présentation, le 7 avril 2015, par Nathalie Fargier, du nouveau site internet de la revue Persee.org¹⁶ ;

☞ une présentation, le 19 mai 2015, par Jean-Sébastien Macke de l'ITEM, du projet ArchiZ dont l'objectif est de mettre en ligne les archives zoliennes de A à Z.

c. *Se former en travaillant par ateliers*

Les responsables des projets numériques de l'ITEM, du Laboratoire d'anthropologie sociale (LAS) et du Labex TransferS de l'ENS ont décidé de mutualiser leur expérience en créant des ateliers de réflexion pratique sur la plate-forme Omeka. L'idée de ces ateliers a émergé lors de mon stage. En effet, Agnès Tricoche du Labex TransferS de l'ENS et Anne Garcia-Fernandez du Laboratoire d'anthropologie sociale (LAS), qui devaient toutes deux utiliser Omeka pour un projet numérique (respectivement le projet Cartomac de l'ITEM et un projet de transcription de contes populaires), ont assisté à la formation Omeka du 12 mars délivrée par Richard Walter. À la suite de cette formation, Agnès Tricoche et Anne Garcia-Fernandez ont également assisté à des « cafés gourmands » informatiques de l'ITEM, notamment sur l'utilisation des wikis, ainsi qu'à la réunion sur le projet e-Man du 24 mars 2015. À la suite de ces différentes rencontres, A. Tricoche, A. Garcia-Fernandez et R. Walter se sont rendu compte qu'ils pouvaient s'entraider sur le fonctionnement en PHP d'Omeka ainsi que sur l'utilisation des plug-ins. Des séances d'atelier informelles (de plusieurs heures) se sont alors mises en place. Pendant ces ateliers, les trois responsables de projets numériques ont exposé leurs problèmes respectifs et ont réfléchi à des solutions concrètes. Souvent les solutions ont émergé pendant l'atelier. Lorsque ce n'était pas le cas, A. Tricoche, A. Garcia-Fernandez et R. Walter se sont servis du forum de l'espace Agora dédié au projet e-Man pour échanger fichiers et solutions testées entre chaque atelier.

3. Être une interface entre le monde de la recherche et le monde du numérique

a. *Coordonner les différents acteurs d'un projet*

Le rôle du responsable de projet numérique est avant tout d'en coordonner les différents acteurs : il est un chef de projet numérique, c'est-à-dire qu'il propose des solutions concernant les dimensions numérique et informatique du projet, mais n'a pas de pouvoir hiérarchique sur les autres membres responsables du projet. Dans le cas de notre projet, Richard Walter, Claire Riffard et Fatiha Idmhand ont chacun le même pouvoir décisionnel concernant la réflexion sur le travail en cours, et ils valident en concertation les solutions à apporter. Dans le cas précis du manuel d'encodage, les apports de chacun ont été discutés et validés par tous les membres. Cela a impliqué plusieurs passages par la rédactrice principale du document que j'étais. A posteriori, après avoir élaboré sept versions successives du document, nous avons conclu, Richard Walter et moi-même, que nous aurions dû organiser davantage de réunions de concertation pour réunir les trois relecteurs du texte. De plus,

15 Voir <http://www.theaville.org/kitesite/index.php>.

16 Voir <http://www.persee.fr/web/guest/home>.

nous aurions dû systématiser les corrections commentées du document et mettre systématiquement en copie les quatre participants à la rédaction. J'ai donc pu, à la manière d'un chef de projet, faire l'expérience de la coordination des acteurs d'un projet, à l'échelle réduite d'un document de référence.

b. Gérer plusieurs projets

Gérer plusieurs projets est la réalité à laquelle est confrontée le responsable de projets numériques : pour prendre l'exemple de mon tuteur de stage, celui-ci gère trois projets en même temps. Il en coordonne donc tous les acteurs, dont une stagiaire pour la plate-forme e-Man (au moment où j'effectuais mon stage), et l'équivalent de deux stagiaires à temps plein pour la plate-forme « Données internationales de génétique artistique » (DIGA).

Au niveau de la gestion des plates-formes en tant que telles, Richard Walter gère trois plates-formes Omeka à l'intérieur de la plate-forme e-Man : celles consacrées aux « Manuscrits francophones », aux « Manuscrits hispaniques », et au projet « Flaubert et le pouvoir des images » (FLIM). Par ailleurs, il gère également, en coordination avec un prestataire informatique régulier, la plate-forme DIGA qui a été construite intégralement par ce développeur web. Le projet DIGA consiste en la création de fiches sur des concepts de génétique artistique. Il ne s'agit donc pas de création de bibliothèque numérique mais d'une base de données. Si le logiciel Omeka a également une dimension de base de données, il permet avant tout de créer une bibliothèque numérique en intégrant des fichiers images sur la plate-forme. Dans le cas du projet e-Man, l'utilisation d'Omeka était tout indiquée puisqu'il s'agissait bien d'intégrer des fichiers images issus de la numérisation de manuscrits d'écrivains. Le rôle de Richard Walter diffère par conséquent dans ces deux cas : dans le cas de DIGA, il s'agit de coordonner chercheurs et développeur informatique, alors que dans le cas d'e-Man, il est à la fois coordonnateur et développeur (sauf lorsque les tâches nécessaires s'avèrent d'une complexité excessive et doivent être confiées à un développeur informatique).

4. Communiquer

La communication fait partie du travail en mode projet. Celle-ci comprend deux volets : la communication avec les acteurs du projet ainsi que la communication auprès du public.

a. Communiquer entre acteurs du projet

Nous avons expérimenté l'utilisation d'un wiki pour communiquer entre acteurs du projet avant de choisir l'utilisation d'un espace numérique de travail déjà en place et largement utilisé à l'ITEM : l'espace numérique Agora Project du CNRS. Par exemple, il existait déjà un espace Agora pour partager les présentations des cafés gourmands et des séminaires techniques. Cet espace Agora possède une centaine d'utilisateurs, qu'ils soient rattachés à l'ITEM, associés aux projets de l'ITEM, ou intervenants lors des séminaires techniques.

Le wiki et l'espace numérique de travail, permettent d'échanger des documents. Ainsi ce sont tous les deux des outils de travail collaboratif. Ils permettent de gérer les utilisateurs : on peut complètement ouvrir le wiki au public ou le restreindre par l'utilisation d'un mot de passe. On peut aussi choisir de fermer certains articles du wiki à l'écriture par les utilisateurs. On peut déposer toutes sortes de contenus sur le wiki. Mais, pour cela, il faut maîtriser un minimum le langage MediaWiki pour pouvoir créer un nouvel article, ajouter des médias ou mettre en page son article. Pour mettre un mot ou un passage en gras ou en italique, le wiki utilise un langage

spécifique à l'aide d'apostrophes : entourer un mot de deux apostrophes de part et d'autre permet de mettre ce mot en italique : par exemple : "mot en italique". Entourer un mot de trois apostrophes de part et d'autre permet de mettre ce mot en gras : "'mot en gras'". Le wiki utilise également un système de balises simplifiées : la balise pour barrer un mot en langage MediaWiki est, par exemple : <strike>texte barré</strike>. Les wikis sont dotés d'une barre d'outils pour encoder la plupart de ces phénomènes. Malgré tout, certaines actions nécessitent l'apprentissage d'un nouveau langage. De plus, l'interface d'un wiki est différente de celle d'un ENT. Comme le wiki n'apportait rien de plus qu'un espace Agora et que tous les membres du projet connaissaient déjà bien ce système, nous avons décidé de créer un espace Agora spécifique au projet e-Man, lequel comprend les documents de travail liés aux différents projets de la plate-forme e-Man.

L'espace Agora permet, en plus de partager des documents, de rédiger des actualités sur le projet.

On peut aussi choisir d'installer un calendrier dans l'espace Agora. Nous n'avons pas fait ce choix parce que chaque projet n'a pas le même calendrier et que les membres des projets fonctionnent déjà par échanges de mails.

En outre, l'espace Agora d'e-Man possède un forum de discussion. Sur ce forum, nous avons créé une rubrique concernant la documentation internet. J'ai également pu présenter les cartes heuristiques que j'ai réalisées. Enfin et surtout, la rubrique la plus active du forum est consacrée à l'atelier Omeka qui s'est organisé autour de Richard Walter, Agnès Tricoche (du Labex TransferS) et Anne Garcia-Fernandez (du Laboratoire d'anthropologie sociale). Il s'intitule le « Coin des développeurs et des développeuses ».

b. Donner de la visibilité au projet

Communiquer avec le public permet, avant tout, de donner de la visibilité au projet. Pour communiquer sur un projet numérique d'un laboratoire scientifique, le Centre pour l'édition électronique (Cléo) a mis en place un outil au service des chercheurs, doctorants et ingénieurs en Sciences humaines et sociales : la plate-forme Hypothèses qui propose l'ouverture d'un blog scientifique appelé « carnet de recherche ».

Nous avons décidé, au sein du projet, d'ouvrir un carnet Hypothèses au début du mois d'avril, suite à la réunion avec tous les acteurs du projet au sein de l'ITEM. En effet, il était apparu très nettement qu'il était nécessaire de définir notre ligne éditoriale. Ce carnet de recherche avait donc pour premier objectif de poser les lignes claires de notre projet d'édition électronique.

Le texte de présentation du carnet, rédigé par Fatiha Idmhand en collaboration avec Claire Riffard et Richard Walter, est visible sur le carnet de recherche du projet e-Man, à l'adresse suivante : <http://eman.hypotheses.org/>.

Le délai d'ouverture d'un carnet Hypothèses, après examen de la demande par le Cléo, est assez long. C'est la raison pour laquelle je n'ai pas eu l'occasion de rédiger d'article pour le carnet avant la fin de mon stage. Mais, ayant obtenu un accès personnel pour rédiger sur le carnet Hypothèses e-Man, je prépare un article sur la problématique de l'édition électronique des manuscrits d'e-Man et les choix d'encodage définis dans le manuel que j'ai rédigé sous la responsabilité de Fatiha Idmhand, Claire Riffard et Richard Walter. Cet article sera un retour d'expérience sur la principale mission qui m'a occupée durant les trois mois de mon stage à l'ITEM. Il sera présent sur le carnet Hypothèses d'e-Man dès qu'il aura été validé par les responsables du projet. Il sera alors visible à l'adresse suivante : <http://eman.hypotheses.org/>.

Le stage effectué au service des supports numériques de l'Institut des textes et manuscrits modernes du 2 mars au 1er juin 2015 m'a permis de produire des documents utiles aux équipes du projet e-Man : un manuel d'encodage XML-TEI comprenant l'ensemble des balises utiles pour l'encodage des manuscrits du projet et devant servir, à terme, au développement d'une barre d'outils d'encodage XML-TEI dans un wiki de transcription ou directement dans la plate-forme Omeka. Ce développement, s'il a lieu, sera effectué par un prestataire. Le deuxième document que j'ai produit servira, de manière concrète, de document de travail au prestataire informatique qui travaillera dès cet été à l'amélioration de la plate-forme Omeka tant sur le plan de l'établissement des relations entre documents que de la constitution du dossier génétique des œuvres des projets « Manuscrits francophones » et « Manuscrits hispaniques » du projet e-Man.

Ces trois mois de stage ont été une mise en situation et un approfondissement des compétences acquises lors de la formation du Master Patrimoine écrit et édition numérique : j'ai adapté les connaissances acquises auprès des Bibliothèques virtuelles humanistes en matière d'encodage XML-TEI et mis en pratique des méthodes de gestion de projet et de travail collaboratif apprises lors de ma formation au CESR. Ce stage m'a également donné l'occasion d'appliquer des connaissances d'édition internet, notamment en ce qui concerne l'utilisation des CMS (plus particulièrement Omeka à travers la plate-forme e-Man, mais aussi WordPress puisque la plate-forme de blogs scientifiques Hypothèses utilise ce CMS).

J'ai également pu observer la manipulation d'un serveur grâce à Richard Walter qui m'a ainsi montré quelques requêtes rédigées en langage PHP. Afin de me sentir vraiment à l'aise dans un métier d'ingénieur en humanités numériques, je souhaite développer la maîtrise des langages PHP et MySQL nécessaires à la prise en main des CMS de bibliothèques numériques et des bases de données.



Plate-forme d'édition de Manuscrits modernes

Présentation du projet

Le projet **e-Man** (Plate-forme d'édition de **Manuscrits modernes**) est développé au sein de l'Institut des textes et manuscrits modernes (ITEM) sous la responsabilité de Fatiha Idmhand, Claire Riffard et Richard Walter ; il vise à créer un outil de publication interopérable favorisant l'exploitation numérique de documents.

e-Man est le résultat de l'association de chercheurs et d'ingénieurs membres de l'ITEM, regroupés autour des équipes « Manuscrits Francophones » et du projet « CHispa : création d'outils pour l'exploitation de manuscrits hispaniques ». Il est soutenu par le laboratoire d'excellence (ou labex) TransferS, l'Agence Nationale de la Recherche (pour CHispa), le Campus Condorcet, le consortium Corpus d'auteurs pour les humanités, information, édition, recherche (consortium CAHIER), la Maison des sciences de l'homme de Lille et le Centre d'études en civilisations, langues et littératures étrangères (CECILLE, EA-4074, Lille), l'action de coopération franco-uruguayenne ECOS-Sud Uruguay et le réseau mixte algéro-français de recherche-formation et de recherche sur la langue française et les expressions francophones (réseau LaFEF).

La démarche de l'ITEM

La spécificité de l'ITEM¹ est l'étude du processus de création (artistique, littéraire, dramaturgique) à partir de la comparaison des différents témoins qui attestent des étapes de la genèse de l'œuvre (manuscrits ou tapuscrits, corrections, ébauches, brouillons, sources extérieures à l'œuvre – exogénétiques – notes, correspondances...). L'ensemble de ces éléments constitue le dossier génétique de l'œuvre.

Les manuscrits numérisés

Les documents qui ont servi à l'expérimentation sont des manuscrits et archives dont la genèse s'étend entre la fin du XIX^e siècle et le début du XXI^e siècle. Nous avons travaillé sur des archives et manuscrits issus des fonds francophones (Jean-Joseph Rabearivelo et Mouloud Feraoun) et des fonds hispaniques (José Mora Guarnido, Carlos Denis Molina, Carlos Liscano). Ces fonds ont été numérisés ou sont en cours de numérisation. En ce qui concerne l'hébergement des fonds originaux, ils sont actuellement conservés chez leurs propriétaires ou leurs ayants droits (en Amérique latine, en Algérie, au SCD Lille 3, à la Biblioteca Nacional de l'Uruguay et à l'Institut français d'Antananarivo de Madagascar).

Historique du projet

Deux actions préalables ont été menées par les équipes CHispa et « Manuscrits francophones » :

☞ Du côté de CHispa, l'équipe a organisé le classement et le traitement physique d'une partie des fonds (près de 6000 documents) ainsi que leur numérisation (2500 fichiers numériques pour l'instant). En 2009-2010, un site a également été développé à partir d'un CMS appelé Spip pour feuilleter les archives numériques

¹ L'Institut des textes et manuscrits modernes, actuellement dirigé par Paolo D'Iorio, est une Unité mixte de recherche du Centre national de recherche scientifique (CNRS) et de l'École normale supérieure (ENS). Voir le site de l'ITEM : <http://www.item.ens.fr/> et <http://www.item.ens.fr/index.php?identifier=l-item>.

et zoomer sur la page.² Ce site, bien qu'il ait permis la publication des manuscrits au cœur du projet CHispa, n'intégrait malheureusement pas les fonctionnalités du web 2.0 et ne comportait aucun travail de structuration des métadonnées (archives, contenus). Néanmoins, malgré ses limites, et près de cinq ans après son ouverture, cette première version a permis d'atteindre deux objectifs principaux : celui de faire connaître ces archives, d'abord, et de susciter l'intérêt sur ces fonds (en Espagne et dans les Amériques) et sur l'approche analytique et éditoriale de l'équipe CHispa.

☞ Du côté des manuscrits francophones, des banques d'images ont été constituées depuis la création de l'équipe au sein de l'ITEM. La banque d'images Rabearivelo contient plus de 5000 fichiers numériques issus d'une campagne de numérisation exhaustive du fonds par un partenaire scientifique malgache (2008). Ces images ont servi à une édition critique à orientation génétique de l'œuvre complète de Rabearivelo (2010, 2012, CNRS Editions). Mais elles n'étaient pas accessibles à la communauté des chercheurs. La même situation se retrouve dans le cas des archives Mouloud Feraoun. L'équipe a été invitée par l'Ambassade de France à expertiser le fonds, à accompagner son inventaire, son conditionnement et sa numérisation (2014). Mais il fallait penser une accessibilité de la banque d'images ainsi constituée.

Deux équipes, une convergence de besoins

La convergence des besoins en termes d'exploitation numérique et de travaux génétiques a incité les deux équipes à s'associer et à poursuivre les recherches afin de créer un outil générique et modulaire qui répondrait à trois ambitions principales :

☞ la première, éditer les manuscrits numérisés par les équipes³ avec des métadonnées⁴ qui respectent les normes et standards internationaux du W3C⁵ ;

☞ la deuxième, pouvoir associer au sein de la plate-forme les documents entre eux grâce à des liens logiques, temporels et génétiques ;

☞ enfin, avoir la possibilité d'associer le manuscrit numérisé à sa transcription enrichie mais aussi à d'autres manuscrits pour constituer un dossier génétique.

Droits et autorisations de reproduction et de diffusion

Les droits et autorisations de reproduction et de diffusion ont été négociés avec les propriétaires et ayants droits : ceux-ci ont consenti la mise en ligne de ces sources sous Licences Creative Commons de types CCBY-NC-ND ou CCBY-NC-SA, c'est-à-dire avec citation de la paternité, interdiction de modifier les documents et interdiction d'utilisation commerciale⁶ dans le cadre de projets de recherche. Ces deux Licences respectent la propriété intellectuelle tout en permettant l'échange de documents sur le web.

Pourquoi Omeka ?

La plate-forme e-Man a été développée avec le logiciel Omeka, un système de publication web spécialisé dans l'édition de collections muséales, de bibliothèques numériques et d'éditions savantes en ligne. Omeka est un logiciel open source

² Flipping Book PDF Publisher Basic, <http://flippingbook.com>.

³ L'équipe CHispa assume également la préservation physique des archives d'Alcides Giraldi et le conditionnement de toutes les pièces du fonds.

⁴ Les métadonnées organisent l'information et décrivent, précisément, la structure des objets décrits : métadonnées descriptives, administratives, de gestion – gestion des droits notamment –, de préservation.

⁵ Voir les guides de bonnes pratiques de Huma Num (ex. TGE Adonis) : <http://www.huma-num.fr/ressources/guides>

⁶ <http://creativecommons.org>

à la croisée du système de gestion de contenus (CMS), de la gestion de collections et de l'édition d'archives numériques. Il a été développé par le Roy Rosenzweig Center for History and New Media de l'Université George Mason (Virginie, Etats-Unis) en vue de permettre aux utilisateurs de « se concentrer sur le contenu et sur l'interprétation plutôt que sur la programmation ».⁷ En effet, Omeka permet de publier et d'éditer des contenus de façon simple et flexible : l'utilisation de modèles visuels (templates) permet, très facilement, d'avoir un rendu satisfaisant. Ceux-ci proposent une architecture simple que l'on peut adapter par la suite aux besoins spécifiques d'un projet par l'ajout de modules. L'autre intérêt d'Omeka est qu'il oblige à structurer les métadonnées des documents assurant ainsi l'interopérabilité des données, pour leur meilleure diffusion, leur archivage numérique et leur « moissonnage » par les grands portails de la recherche : Isidore, Gallica, Europeana, Biblioteca Virtual Cervantes... Enfin, Omeka étant un logiciel libre, il dispose d'une forte communauté qui assure sa maintenance.

Une typologie applicable à l'étude de nombreux manuscrits

Les manuscrits numérisés sont décrits à l'aide des quinze champs du Dublin Core, mais ils comportent également des éléments de description (des « champs ») qui correspondent à la fois aux caractéristiques de nos fonds (recueils, théâtre, référent spatiaux...) et aux intérêts scientifiques repérés et décrits par les chercheurs. L'élaboration de ces métadonnées supplémentaires au sein du logiciel, dites « de type contenus », est inspirée du modèle de description archivistique encodée (EAD). Ce travail de sélection a été réalisé en collaboration avec les membres de l'équipe CHispa (Cécile Brailon, Laurent Passion) et avec l'aide d'une archiviste (Cécile Martini, archiviste, Lille).

Le plan de description des manuscrits s'est appuyé, lui, sur le document qui a été publié en 2010 et rédigé sous la direction de l'Agence française de normalisation (AFNOR) et de la Direction du livre et de la lecture, en partenariat avec des chercheurs de l'Institut des Textes et Manuscrits Modernes (ITEM), le plan DeMArch. Celui-ci intègre la notion de dossier génétique. Nous avons pu établir une typologie des documents qui respecte les recommandations en vigueur, qui s'applique aux manuscrits francophones et hispaniques d'e-Man mais qui peut potentiellement s'appliquer à toute autre étude de manuscrits modernes et contemporains.⁸

Une approche par dossier génétique

La plate-forme e-Man, grâce à Omeka, permet la mise en relation des documents entre eux grâce à des liens actifs et spécifiés ; ceci permet de mettre en exergue le travail scientifique de constitution de dossiers génétiques que le chercheur réalise ainsi que le processus de création d'une œuvre. Le manuscrit – le document – publié est ainsi associé à son dossier génétique (avant-textes, ébauches, documents préparatoires, sources exogénétiques, correspondances...).

Une plate-forme ouverte à d'autres projets d'édition de manuscrits

La plate-forme e-Man est ouverte à d'autres projets d'édition de manuscrits. Elle vient notamment d'intégrer le projet Flaubert et le pouvoir des images (FLIM).

Cécile Brémon (Humanités numériques), Fatiha Idmhand (Manuscrits hispaniques), Claire Riffard (Manuscrits francophones), Richard Walter (Humanités numériques)

⁷ <http://omeka.org/about>

⁸ Le groupe de travail s'est inspiré de la norme Describing Archives : A Content Standard (DACS) qu'il a traduite et adaptée aux usages des établissements français. Il s'est également référé ponctuellement à d'autres normes existantes pour la description des ressources archivistiques (Règles pour la description des documents d'archives (RDDA) par ex.) et, le cas échéant, aux normes existantes pour des types de documents précis (archives de films par ex.). Voir Association française de normalisation. Commission de normalisation Modélisation, production et accès aux documents, *Description des manuscrits et fonds d'archives modernes et contemporains en bibliothèque : DeMArch* [en ligne], 2010, <http://www.bivi.fonctions-documentaires.afnor.org/livres-blancs/recommandation-de-description-des-manuscrits-et-fonds-d-archives>. Voir également, sur le site de la BnF, le projet en ligne : http://www.bnf.fr/fr/professionnels/normes_catalogage/a.ead_demarch.html.



PLATE-FORME D'ÉDITION DE MANUSCRITS MODERNES

Guide de saisie des notices, 25/03/2015

La plate-forme d'édition de manuscrits modernes e-Man utilise le logiciel Omeka qui fonctionne comme un CMS ou système de gestion de contenu. Ce logiciel permet de créer une bibliothèque numérique en utilisant un modèle (*template*) que l'on peut compléter de modules d'extension (*plugins*).

Le lien suivant vous envoie sur la plate-forme e-Man : <http://www.eman-archives.org/>.

Vous avez alors accès aux deux plate-formes :

« Manuscrits francophones » (<http://www.eman-archives.org/francophone/>) et

« Manuscrits hispaniques » (<http://www.eman-archives.org/hispanique/>)

D'autres plate-formes rejoignent progressivement e-Man, par exemple « Flaubert et le pouvoir des images - FLIM » (<http://www.eman-archives.org/FLIM/index.php>).

≈ SOMMAIRE ≈

Entrer dans e-Man	p. 2-3
Qu'est-ce qu'une collection ?	p. 3
Droits de saisie	p. 3
Autres remarques	p. 3-4
<i>Tableau comparatif des utilisateurs</i>	p. 4
CRÉATION DE NOTICES	p. 5
Généralités	p. 5
Onglets à remplir	p. 6-10
LISTE COMMENTÉE DES CHAMPS DUBLIN CORE	p. 8-10
LISTE COMMENTÉE DES MÉTADONNÉES DU TYPE DE CONTENU	p. 11-13
<i>Recommandations notices – Mots-clés</i>	p. 14
<i>Exemple de nuage de mots-clés</i>	p. 14

≈ PLATE-FORME D'ÉDITION DE MANUSCRITS MODERNES ≈

≈ Entrer dans e-Man

- L'interface publique



<http://eman-archives.org/>

Si vous cliquez sur « Manuscrits francophones », par exemple, vous accédez à l'interface publique de la plate-forme « Manuscrits francophones ».

(Voir la copie d'écran ci-après, qui correspond à l'adresse suivante : <http://eman-archives.org/francophone/>)

- Accès à l'interface administrateur depuis l'interface publique

En que contributeur au projet, vous avez accès à l'interface administrateur d'e-Man.

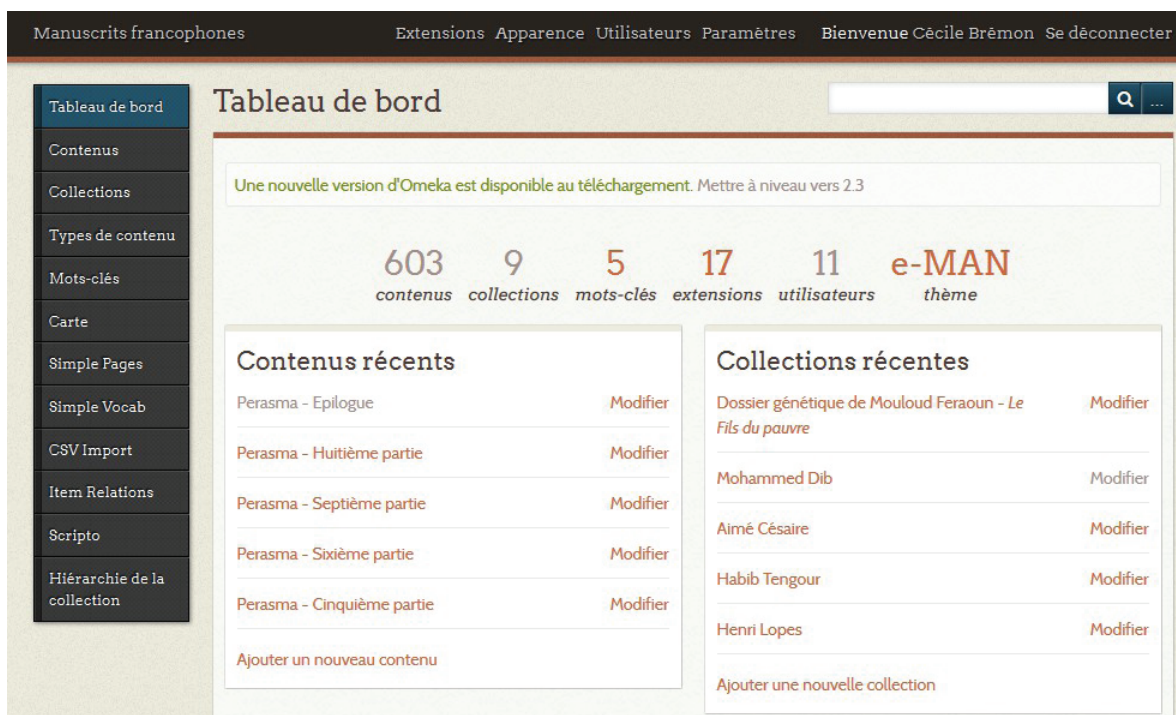
Pour accéder à l'interface administrateur, cliquez en bas de la page sur « Connexion »



Vous pourrez alors entrer l'identifiant et le mot de passe que vous aurez reçus.

- Menu de gauche du tableau de bord

Vous accédez alors au tableau de bord de l'interface administrateur.



Sur la gauche, vous avez le menu suivant : « Tableau de bord », « Contenus », « Collections », « Types de contenu », « Mots-clés » et « Carte » notamment.

Dans ce menu de gauche, vous n'avez besoin d'utiliser que « Contenus » et « Collections ».

- **Retour à l'interface publique**

En cliquant sur le nom du site en haut à gauche (« Manuscrits francophones » ou « Manuscrits hispaniques »), vous pouvez consulter le site tel que les internautes peuvent le voir.

☞ Qu'est-ce qu'une collection ?

Une collection correspond à l'ensemble des archives d'un auteur donné : il existe, par exemple, une collection « Mouloud Feraoun » sur la plate-forme « Manuscrits francophones » et une collection « Carlos Denis Molina » sur la plate-forme « Manuscrits hispaniques », etc.

Pour créer une collection, merci d'envoyer le nom de la collection et sa description à Richard Walter (richard.walter@ens.fr) ou au responsable de la plate-forme.

☞ Droits de saisie

Omeka propose 4 niveaux d'utilisateurs :

Super administrateur/ Administrateur/ Contributeur/ Chercheur

Pour plus d'explications, voir le tableau comparatif des utilisateurs en page 4

et la documentation d'Omeka en anglais : http://omeka.org/codex/User_Roles.

En tant que chercheur, vous ne pouvez que voir les notices non publiées.

En tant que contributeur, vous pouvez gérer des notices mais vous ne pouvez pas modifier les notices créées par une autre personne.

En tant qu'administrateur, vous pouvez corriger des notices faites par d'autres utilisateurs.

☞ Autres remarques

- **Barre de recherche et contenu non public**

La barre de recherche (ou de recherche avancée) permet aux visiteurs d'accéder au contenu public.

Remarque : quand vous êtes connectés en tant que contributeur, chercheur ou administrateur, vous accédez à *tous* les contenus (publics ou privés), même via l'interface publique. Soyez rassurés, cela ne signifie pas que ce contenu soit devenu public pour autant.

- **Suppressions définitives**

L'interface est intuitive : n'hésitez pas à tester, vous ne pourrez pas casser grand-chose (mais les suppressions de contenus sont définitives).

Attention : appuyer sur le bouton

Supprimer

supprime la notice définitivement !

(C'est donc un bouton à utiliser avec précaution. Si jamais, vous effaciez une notice par erreur, prévenez tout le suite Richard Walter (richard.walter@ens.fr) pour qu'une restauration des données de la veille puisse être effectuée.)

UTILISATEURS E-MAN / OMEKA

Super (administrateurs) Gestion de l'administration :		Administrateurs Gestion du contenu		Contributeurs Peuvent :		Chercheurs Peuvent :	
• création des comptes • installation des plug-ins		Menu de gauche	Menu horizontal	• voir les contenus ou collections non publiés		• voir les contenus ou collections non publiés	
Menu de gauche	Menu horizontal	Tableau de bord	Extensions	Ne peuvent pas :		Menu de gauche	Menu horizontal
Tableau de bord	Extensions	Contenus Possibilité de modifier ou supprimer tous contenus	Apparences	• publier leur contenu sans l'aide d'un super administrateur • voir les contenus non publiés		Tableau de bord	Extensions
Contenus Possibilité de modifier ou supprimer tous contenus	Apparences	Collections	Utilisateurs	Menu de gauche	Menu horizontal	Contenus Possibilité de modifier seulement leur propre contenu	Apparences
Collections	Utilisateurs	Types de contenu	Paramètres	Tableau de bord	Extensions	Collections	Utilisateurs
Types de contenu	Paramètres	Mots-clés		Contenus - Possibilité d'ajouter du contenu - Possibilité de modifier seulement leur propre contenu ou de le supprimer	Apparences	Types de contenu	Paramètres
Mots-clés		Carte			Utilisateurs	Mots-clés	
Carte		Simple Pages			Paramètres	Carte	
Simple Pages		Simple Vocab				Simple Pages	
Simple Vocab		CSV Import				Simple Vocab	
CSV Import		Item Relations				CSV Import	
Item Relations						Item Relations	
				Mots-clés Peuvent ajouter des mots-clés de type « tags »			
				Carte			
				Simple Pages			
				Simple Vocab			
				CSV Import			
				Item Relations			

≈ CRÉATION DE NOTICES ≈

≈ Généralités

Pour ajouter un contenu, ne pas oublier de choisir une collection.

Quand vous entrez dans un contenu, cliquez sur « modifier la page ». On vous propose alors plusieurs onglets : « Dublin Core », « Métadonnées de type contenu », « Scripto », « Fichiers », « Mots-clés », « Carte » et « Item Relations ».

(Si vous entrez dans un fichier, on ne vous proposera que les onglets « Dublin Core » et « Scripto ».)

- **Rendre public**

Ne pas oublier de cocher la case « Public » si vous voulez que votre notice soit publiée et accessible de l'extérieur (sinon elle est indiquée dans l'interface de saisie comme « réservée »)

- **Enregistrer les changements**

Ne pas non plus oublier de sauvegarder les changements. Sinon les modifications ne sont pas enregistrées.

Attention : sauvegardez les changements avant de changer d'onglet sinon les modifications effectuées sur l'onglet précédent ne sont pas enregistrées.

Si vous avez plusieurs contenus différents à saisir pour un champ, vous devez saisir chacun d'entre eux dans une entrée spécifique (cliquez sur « Ajouter une entrée »)

Cliquez sur « utiliser le HTML » pour faire apparaître les « boutons » gras, italique, souligné...

- **Champs de métadonnées**

Les champs de métadonnées sont expliqués avec des commentaires génériques, qui ont été personnalisés pour le projet (DC et métadonnées spécifiques).

≈ Onglets à remplir



- **Dublin Core**

Le Dublin Core est un standard de description des métadonnées.

Voir le descriptif des champs du Dublin Core en pages 8 à 10.

Les champs du Dublin Core sont à remplir autant que possible. Ce sont eux qui constituent la notice.

Les transcriptions sont, pour l'instant, placées dans le champ « Description » du Dublin Core. À terme, nous utiliserons la fenêtre Scripto qui s'ouvrira à partir d'un lien qui sera situé sous le fichier à transcrire.

- **Métadonnées du type de contenu**

Le type de contenus est à choisir dans le menu déroulant :

Création, Correspondance, Dossier biographique, Iconographie, ou Presse.

Pour de plus amples explications voir la fiche « Métadonnées du type de contenu » en pages 11 à 13.

- **Fichiers**

Vous pouvez en rajouter plusieurs par notice.

La taille maximum du fichier est de 100 MB.

Ne pas oublier de sauvegarder les changements si vous voulez que votre fichier soit déposé sur le serveur. Chaque fichier a son propre descriptif Dublin Core.

- **Mots-clés**

Cette fonction est spécifique à Omeka : à partir de ce champ, on peut faire un nuage de tags ou une recherche par mots-clés. Elle ne se sert pas du champ « Sujet » du Dublin Core.

Vous devez donc ici copier/coller les mots-clés que vous aurez saisis dans le Dublin Core.

Pour de plus amples explications, voir les fiches « Recommandations notices » et « Nuage de mots-clés » page 14.

- **Carte**

Vous pouvez géolocaliser l'endroit où le manuscrit est localisé, grâce à GoogleMaps.

À terme, il faudrait pouvoir récupérer ces données pour les intégrer dans le champ DC « coverage ».

- **Item Relations**

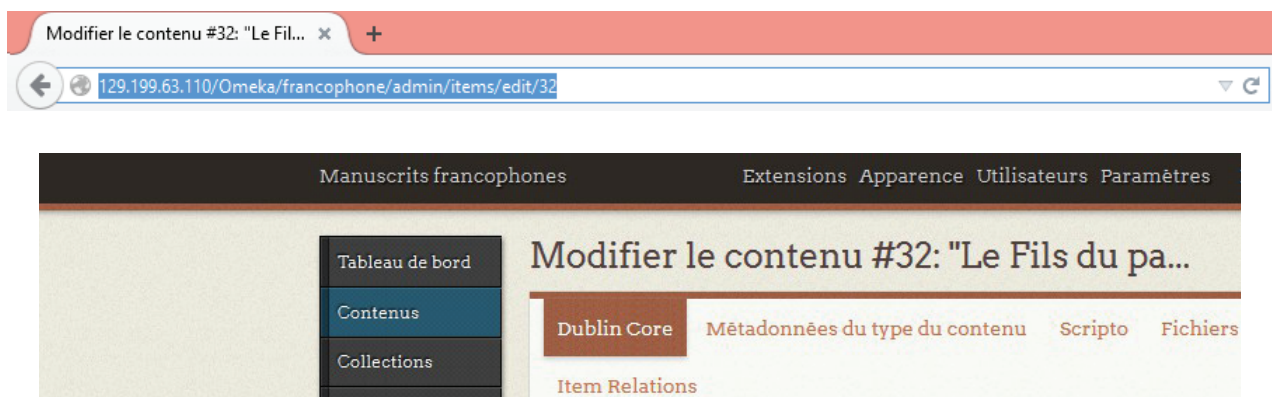
Cette fonction est spécifique à Omeka : à partir de ce champ, on peut établir des relations entre deux notices. On utilise cette fonctionnalité pour établir des relations entre les documents au sein d'un « dossier génétique ».

Malheureusement cette fonction n'est pas liée au champ « Relation / Dossier génétique » du Dublin Core. On a toujours besoin de remplir le champ DC « relations » pour pouvoir éditer en PDF ou autre format, une notice DC complète.

Dans Item-Relations, il faut donc ressaisir ici les indications de ce champ. Mais il ne faut pas indiquer la cote mais l'identifiant de la notice (ID). L'ID est un nombre de 2 à 3 chiffres : dans la barre de navigation, c'est le dernier chiffre de l'adresse (url) de la notice.

Exemple : <http://129.199.63.110/Omeka/francophone/admin/items/edit/32>

=> L'ID est 32



Le mieux serait de toujours avoir un onglet ouvert dans lequel vous pouvez consulter les autres notices saisies et récupérer l'ID. N'hésitez pas à tester après saisie.

Quand vous saisissez une relation dans une notice, elle se met automatiquement dans l'autre notice citée. Vous n'avez donc pas besoin de la saisir de nouveau (ouf !).

Nous avons sélectionné trois types de « relations » entre 2 notices :

- Avant texte de
- Fait partie de
- Est en relation avec

Indiquer la nature d'une relation est donc une opération intellectuelle/scientifique délicate : privilégiez, dans le doute, l'option « est en relation avec », car les 2 autres demandent l'expertise du chercheur spécialiste du corpus.

Vous pouvez toujours, dans le champ « Description », commenter ou spécifier avec plus de détails les différentes relations.

- **Les champs des notices :**

Pour faciliter la saisie, **voici la liste commentée des champs Dublin Core et celle des métadonnées du type de contenu**. Cette liste est à connaître par cœur pour saisir au bon endroit les informations recueillies.

Le jeu de champs de métadonnées Dublin Core est commun à tous les contenus sous Omeka (notices, fichiers et les collections). En fin de guide, nous proposons un tableau de correspondances entre les champs prévus pour le projet et le Dublin Core.

À l’affichage, les noms des champs DC seront modifiés pour les besoins du projet mais il est important de garder la dénomination DC quand on exporte la notice dans un autre format (PDF, XML, etc.).

≈ DUBLIN CORE ≈

En italique, est, à chaque fois, indiquée la définition « officielle » du Dublin Core.¹

1. Titre (*Title*)

C'est le nom que porte la ressource ou celui qui lui est donné.

Le titre sera celui de l'œuvre (*Les Misérables*), ou un « titre forgé », ou celui qui a été assigné à la ressource (carnet n°24 ; lettre de X).

- En cas d'absence de titre, indiquer la première phrase ou les premiers mots, entre crochets mais sans ponctuation finale (titre forgé) ; ex : [Longtemps, je me suis couché de bonne heure].
- Pas de guillemets aux titres.
- Titre le plus unique possible.

Quand des notices portent le même titre (ex. typique des variantes d'un même texte), indiquer un numéro ou un autre après le titre. Sinon on se retrouve avec une liste de notices portant toutes le même nom. (Si le ou la spécialiste peut différencier avec la cote, il faut ménager le lecteur...)

2. Auteur / Créateur (*Creator*)

L'entité principalement responsable de la création de la ressource.

Indiquer le nom de l'auteur du document, de l'œuvre, en utilisant la **forme « Nom, Prénom »**. La virgule permet de séparer le nom et le prénom, elle est très importante pour identifier l'un et l'autre).

- Format : Nom, Prénom
- Hormis la majuscule à la première lettre, ne pas utiliser la majuscule. Ex. Hugo, Victor et non HUGO, Victor.
- Quand l'auteur n'est pas le sujet de la collection, vous pouvez mettre les dates de naissance et de mort, entre parenthèses après le nom de l'auteur. Ex. Hugo, Victor (1802-1885).
- Si l'œuvre comporte plusieurs auteurs : mettre un nom par entrée.

3. Sujet (*Subject*)

Le sujet de la ressource.

Privilégier des mots-clés significatifs : en vue du moissonnage, on tentera de rendre compte de ce dont parle la ressource, en peu de mots (ex : guerre d'Algérie, franquisme, exil...). Il convient donc de choisir des mots-clés thématiques.

- Créer une entrée par mot-clé.
- Ne pas utiliser les guillemets.

Ne pas oublier d'ajouter ces mots-clés dans l'onglet « Mots-clés » pour la recherche par tags d'Omeka.

4. Description (*Description*)

Une présentation du contenu de la ressource

Indiquer ici tout commentaire ou analyse utile, quelques lignes comme « *Ce récit fait partie d'une série de récits rassemblés dans un recueil nommé « xxxxxxxx ».* Il comporte des mentions manuscrites sur »

Un autre champ existe pour indiquer l'auteur de cette analyse.

Attention à ne pas confondre l'auteur de l'analyse (champ Description), et l'auteur de la description (toute la fiche) et l'auteur révision

5. Cote (DC : *Source*)

Une référence à une ressource à partir de laquelle la ressource décrite a été dérivée.

Indiquer la cote du document donnée par l'institution qui héberge le fonds.

6. Éditeur (*Editor*)

L'entité responsable de la diffusion de la ressource.

¹Pour établir ce guide, nous avons également travaillé avec les définitions de la BnF. Voir en ligne : http://www.bnf.fr/fr/professionnels/formats_catalogage/a.f_dublin_core.html

Indiquez la personne et l'institution qui sont responsables de l'édition, sous la forme « **Nom du projet, Prénom Nom, Institution** ».

- Francophone : Projet e-Man, Claire Riffard, ITEM, CNRS-ENS, France
- Hispanique : Projet e-Man, Fatiha Idmhand, ITEM, CNRS-ENS, France
- Une entrée par information si vous avez besoin de spécifier un autre partenaire de l'édition, rôle à préciser en fin de notice et entre parenthèses (financier, hébergeur, éditeur).

Les contributeurs (auteurs des notices, descriptions et analyses) ont leurs champs spécifiques : cette information est essentielle pour mettre en valeur les apports de tous les membres de l'équipe.

7. **Date (Date)**

Une date ou une période associée avec un événement dans le cycle de vie de la ressource.

Indiquer la date d'écriture du document : AAAA-MM-JJ.

- Quand vous ne connaissez pas le jour, garder AAAA-MM.
- Si le document ne peut être daté précisément, indiquer une période à l'aide de dates extrêmes. Ex. : 1880/1900 (soit AAAA/AAAA) plutôt que fin XIXe siècle.
- Si on a plus de précisions pour les dates extrêmes, garder le format AAAA-MM-JJ. Ex. : 1880-03/1881-11 plutôt que mars 1880/novembre 1881.

Vous pouvez donner dans le champ « description » toute précision utile sur la chronologie de l'œuvre.

8. **Ville ou lieu géographique (DC : Coverage)**

La portée spatiale ou temporelle de la ressource, ou la juridiction pour laquelle la ressource est pertinente.

Dans le Dublin Core, cette information dépend du choix du chercheur, soit il privilégie l'ajout d'informations temporelles ou sur la ville ou une hiérarchie géographique (ex : Alger ; Algérie ; Afrique).

Ici, nous avons choisi d'indiquer la ville où est situé le fonds possédant la ressource décrite dans la notice : il est inutile d'apporter plus de précisions sur la localisation vu l'existence de l'outil de géolocalisation à remplir.

Il ne faut pas confondre ce champ avec « Référent géographique » (métadonnées du type de contenu) qui sert à spécifier le référent spatial de l'œuvre elle-même.

- Une indication par entrée.
- Aucune indication temporelle.

9. **Dossier génétique (DC : Relation)**

Une ressource dérivée.

Indiquez si le document appartient à un dossier génétique, en ne mentionnant que les cotes des documents associés. Utilisez l'onglet « Item – Relations » pour spécifier le type de relations existant entre les documents associés. Le dossier génétique sera construit avec cet onglet et non avec les données de ce champ, qui sert uniquement à la constitution d'une fiche de métadonnées complètes.

- Une entrée par cote

10. **Collation (DC : Format)**

Le format de fichier, le type de support ou les dimensions de la ressource.

Collation : nombre de feuillets (f.) ; nombre de page (p.) ; dimension (L x l en mm)

11. **Langue (Language)**

Langue de la ressource, de l'œuvre décrite

Choix dans une liste déroulante : Anglais / Arabe / Berbère / Espagnol / Français / Haïtien / Italien / Kabyle / Latin / Malgache / Portugais.

Si autre langue, mail à Richard Walter (richard.walter@ens.fr) ou au responsable de la plate-forme.

- Une entrée par langue

12. **Genre (DC : Type)**

La nature ou le genre de la ressource.

Choix dans une liste déroulante.

Si le type de la ressource n'est pas dans le menu déroulant, merci de voir avec les responsables du projet.

13. **Identifiant (Identifier)**

Une référence non ambiguë à la ressource dans un contexte donné.

Ne pas remplir

14. Contributeur (*Contributor*)

Une personne(ou une entité) qui a contribué à la création du contenu de la ressource.

Indiquer toute personne ou toute institution ayant contribué à l'édition numérique du document.

Par exemple, indiquer la bibliothèque qui a permis ou fait la numérisation du document décrit dans la notice.

Format : Nom, Prénom (rôle)

- Chaque contributeur est signalé dans une nouvelle entrée.
- Son intervention est signalée facultativement entre parenthèses en fin de champ.

Les autres contributeurs (auteurs de la description, de l'analyse ou de la révision de la fiche) ont leur propre champ personnalisé, ne pas les répéter ici.

15. Gestion des droits (*Rights*)

Information sur les droits détenus dans et sur la ressource.

Toutes les mentions légales et le nom d'un contact si possible

- Une information par entrée

≈ MÉTADONNÉES DU TYPE DE CONTENU ≈

≈ Type de contenu

Sélectionnez dans le menu déroulant :

- Correspondance : une lettre ou toute correspondance ;
- Création : les œuvres de création (tout genre, tout style) ;
- Dossier biographique : tout ce qui concerne le dossier biographique (documents autres que des œuvres) ;
- Iconographie : tout document iconographique (autre que du texte) ;
- Presse : un article dans une revue, périodique ou tout support de presse ;
- Documents pédagogiques : cours, plan de cours et autres documents pédagogiques créés par l'auteur

Document iconographique : document sans texte. La carte postale fait partie de la correspondance.

Les types de contenus « Création », « Correspondance » et « Presse » ont des métadonnées spécifiques.

≈ Métadonnées communes

1. Sous-titre

Si titre forgé, ne pas oublier de mettre les crochets...

2. Référent géographique

➤ Lieu(x) de l'action décrite dans le contenu de l'œuvre.

A ne pas confondre avec le champ n°8 du DC (coverage).

3. Nature du document

- Manuscrit, tapuscrit, copie dactylographiée, etc.
- Lettre, carte postale, photo, etc. (correspondance)
- Photographie, diapositive, etc. (iconographie)
- Passeport, carte d'identité, répertoire, agenda, etc. (dossier biographique)
- Journal, magazine, revue, etc. (presse)

4. Support

- Papier, carnet, feuillet,
- Carte plastifiée (dossier biographique), négatif, etc. (iconographie),
- Autre support : enregistrement sonore, vidéo, etc.

5. État général

➤ Menu déroulant : Bon, moyen ou mauvais

Pour tout commentaire sur l'état du document (présence de ratures, etc.), il faut mieux l'indiquer dans la rubrique « état génétique ».

6. État génétique

➤ Toutes les informations sur l'élaboration du document ou la phase d'écriture (1er brouillon, 1ères épreuves corrigées, présence de ratures ou d'autres éléments, etc.).

7. Localisation

Localisation géographique du document.

➤ Indiquer où est hébergé le document et le nom du fonds, avec l'adresse précise si possible

8. Notes

➤ Tout ce que vous n'avez pas pu mettre ailleurs ou en attente de création de champ spécifique.

9. Auteur de l'analyse (champ Description)

Auteur de l'analyse (champ DC n° 4 "Description")

➤ Indiquer l'auteur suivi de la date de l'analyse (**format français & entre parenthèses**).

- Une entrée par auteur

10. Auteur de la description (toute la fiche)

Auteur de la rédaction de la fiche

- Indiquer l'auteur suivi de la date de la description (format français & entre parenthèses).
- Une entrée par auteur

11. Auteur révision

Auteur de la révision de la fiche

- Indiquer l'auteur suivi de la date de la révision (format français & entre parenthèses).
- Une entrée par auteur

12. Auteur transcription

Auteur de la transcription si elle existe.

- Indiquer l'auteur suivi de la date de transcription (format français & entre parenthèses).
- Une entrée par auteur

13. Auteur informatique

Responsable de l'édition électronique (Richard Walter), cf. champ DC Contributeur.

Métadonnées spécifiques « Création »

1. Recueil

- Le document décrit dans la notice appartient à un ensemble : indiquez lequel.
- Un ensemble par entrée

2. Publication

La publication renvoie à un facsimilé ou à un livre. Si le document a été publié, indiquer toutes les références bibliographiques connues.

Deux choix :

- Inédit.
- Nom, Prénom (de l'auteur), Titre de la publication, Lieu, Maison d'édition, Date.

Indiquer éventuellement le volume, tome ou collection si nécessaire....

Une entrée par édition

3. Représentation

Champ spécifique au théâtre

Deux choix :

- Inédit.
- Mise en scène avec ville, date, metteur et scène puis toute indication nécessaire...
- Une entrée par édition

Métadonnées spécifiques « Correspondance »

Si l'expéditeur n'est pas l'auteur concerné par la collection, il faut l'indiquer dans le champ DC « Auteur ».

1. Destinataire

- Format : **Nom, Prénom**
- Hormis la majuscule à la première lettre, ne pas utiliser la majuscule. Ex. Hugo, Victor et non HUGO, Victor.
- Vous pouvez mettre les dates de naissance et de mort, entre parenthèses après le nom de l'auteur.
- Si la lettre a plusieurs destinataires, mettre un nom par entrée.

Si l'identification d'un éventuel surnom ou pseudonyme n'a pas été possible, on l'indiquera dans le champ DC « Description »

2. Lieu d'expédition

Format : Ville (Pays), puis Adresse complète si possible.

3. Publication

La publication renvoie à un facsimilé ou à un livre. Si le document a été publié, indiquer toutes les références bibliographiques connues.

Deux choix :

- Inédit.
 - Nom, Prénom (de l'auteur), Titre de la publication, Lieu, Maison d'édition, Date.
Indiquer éventuellement le volume, tome ou collection si nécessaire....
- => Une entrée par édition

🌀 Métadonnées spécifiques « Presse »

4. Titre de la publication

- Nom du journal ou de la revue

5. Ville

- Lieu d'édition du journal ou de la revue
- Indiquez uniquement la ville

6. Numéro de la publication

- Indiquez la numérotation qui apparaît sur la publication
- En cas d'absence, mettre l'année.

7. Directeur de la publication

- Format : Nom, Prénom (fonction)
- Fonction : rédacteur en chef, directeur de publication, etc.

8. Type de la publication

- Journal ou revue ou magazine

Le critère est celui de l'information que comporte le document : s'il se dit « magazine » on sélectionne « magazine ».

9. Périodicité

- Quotidien, hebdomadaire, mensuel, semestriel, annuel.

Le critère est celui de l'information que comporte le document. Sélectionnez donc la périodicité indiquée dans le document, même s'il a changé par la suite.

10. Nombre d'exemplaires

- Tirage de la revue

Deux types de mots-clés

Le terme « mot-clé », dans Omeka, désigne deux choses distinctes :

- d'une part, les mots-clés de type « subject » qui servent à faire les notices ;
- d'autre part, les mots-clés de type « tags » qui servent à faire des nuages de tags.

En pratique, on entre les mêmes mots-clés dans les deux cas : il faut, en fait, faire un copier-coller des mots-clés de type « subject » dans les tags.

On entre les mots-clés de type « subject » dans le champ « sujet », à partir de l'onglet « Dublin Core » du menu du haut.

On entre les mots-clés de type « tags » à partir de l'onglet « mots-clés » du menu du haut.

Des mots-clés thématiques

Pour que le nuage de tags soit pertinent et afin d'éviter toute redite par rapport aux « Item relations » et aux autres éléments déjà présents par ailleurs dans la notice, il convient de choisir un nombre très limité de mots-clés. Par ailleurs, ces mots-clés sont en grande majorité des mots-clés thématiques.

Par exemple « amitié » si le thème principal du document est l'amitié.

Ne pas inscrire de nom d'auteur, de titre, ou encore le genre du document dans les mots-clés. Les « Item relations » (titres d'œuvres en relation avec le document, par exemple), ou le « Dublin Core » (genre du document – « DC type ») sont faits pour ça.

Exemple de nuage de tags

Pour avoir une idée de ce qu'est un nuage de tags, voir l'exemple ci-dessous.

NUAGE DE MOTS-CLÉS

18/03/2015

Nuage de mots-clés de type tags sur le site du *Roman des Morand* de l'ENS Lyon :

Nuage de mots-clefs

accouchement affection alimentation amitié **argent** armée napoléonienne attendrissement
 belle-mère bibliothèque bonheur bonnes oeuvres Brotteaux cadeaux cancer carnaval chagrin
 Charbonnières chirurgie citoyenneté **climat** colère confiance conflit conflit conjugal
conflit fils-mère conseils maternels consolation courage découragement **deuil** dîner de
 veuve distractions domestiques douleur économie agricole économie domestique éducation
 éducation des enfants éducation des filles **éducation des garçons** éducation maternelle
 éducation parentale **éducation paternelle** égoïsme élections émotions enfance
 enfants épidémie espérance études Europe exil fêtes familiales fidélité filles foi garçons
 garde-robe gestion des appartements **gestion du pont** goutte gouvernement
 grossesse handicap homme hypothèque identité de genre impôts infidélité intendance
 intérêts inventaire jalousie joie justice lecture lessives **lettre remarquable** lettres-
 non-transcrites logement lois loisirs maladies malheur manque mariage mariage d'Albine
 Mariage d'Eléonore Morand de Jouffrey mariage des enfants médecin **médecine** mémoire
 familiale mobilier mode enfantine mode féminine modestie mort mort d'Antoinette musique
 naissance non-transcrites obligations parentales **opinion politique** parenté peine de mort
 pension plainte pratique religieuse pratiques aristocratiques pudeur querelle recherche de
 paternité reconnaissance reconstruction de Lyon régime alimentaire rentrée des classes
 reproches République réputation Restauration de 1814 retraite révolution royalisme royauté

<http://morand.ens-lyon.fr/items/tags>

Pour info, le site du *Roman des Morand* utilise Omeka et le plug-in TEI Display

sacrifice maternel **santé** sensibilité sensibilité masculine sensualité
sentiment amoureux sentiment grand-maternel sentiment grand-paternel sentiment
 maternel sentiments sexualité **sociabilité** soins paternels solidarité succession suicide
 susceptibilité tarifs taux d'intérêt tendresse filiale tendresse grand-maternelle tendresse
 maternelle tendresse paternelle test théâtre théâtre enfantin timidité masculine transports
 vacances vieillesse vieillissement visite du pape à Grenoble



Manuel d'encodage des transcriptions dans la plate-forme d'édition de manuscrits modernes e-Man

Manuel rédigé par Cécile Brémon¹

Sous la responsabilité de Fatiha Idmhand, Claire Riffard, Richard Walter²



Version n°7 du 29 mai 2015

I. À la recherche d'un outil pour encoder les transcriptions des textes publiés dans e-Man.....	4
A) Introduction.....	4
1 Présentation d'e-Man et enjeux de l'encodage des transcriptions en XML-TEI.....	4
2 À la recherche d'un outil de transcription en ligne.....	5
B) Une barre d'outils pour encoder la transcription en XML-TEI.....	6
C) Étapes de l'encodage.....	8
1 De la transcription au fichier XML avec table des matières.....	8
2 Structure du fichier XML.....	8
3 Tout encoder ?.....	9
D) Génétique des textes.....	9
II. Liste des balises adaptées aux besoins des équipes d'e-Man.....	10
A) Structure physique.....	10
1 Caractères spéciaux.....	10
a) Esperluette (&).....	10
b) Tiret long (—).....	10
2 Indice et exposant mathématiques.....	11
3 Retour à la ligne.....	11
4 Coupure de mot en fin de ligne.....	11
5 Saut de page.....	11
6 En-tête et pied de page de documents imprimés <fw> (Forme Work).....	11
B) Structure logique.....	12
1 Division.....	12
2 Tête de division : <head>	12
3 Paragraphe, alinéa et espace.....	13
a) Paragraphe.....	13

¹ cecile.bremon@gmail.com

² fatihaidmhand@yahoo.es, claire.riffard@item.cnrs.fr, richard.walter@ens.fr

b) Alinéa.....	13
c) Espace <space>.....	13
4 Centrer, aligner, justifier.....	13
5 Texte théâtral.....	14
a) Acte, scène.....	14
b) Réplique et locuteur.....	14
c) Didascalies	14
6 Correspondance épistolaire.....	15
a) Support.....	15
b) Sceau <stamp>.....	15
c) Date et lieu de rédaction <dateline>.....	15
d) Formule d'appel ou formule finale <salute>.....	15
e) Signature.....	16
f) Début et fin de la lettre <opener> <closer>.....	16
7 Poésie.....	16
a) Poème et recueil.....	16
b) Strophe et vers.....	17
8 Texte en prose : roman, essai, presse, etc.....	17
9 Liste, colonne et tableau.....	17
a) Liste <list>.....	17
b) Colonne <cb/>.....	17
c) Tableau <table>.....	17
10 Abréviations.....	18
11 Erreur et correction <sic> <corr>.....	18
C)☞ Indexation.....	19
1 Noms de personnages ou de personnes.....	19
2 Noms de lieux.....	19
D)☞ Transcription.....	20
1 Soulignement, double soulignement, surlignement.....	20
a) Soulignement simple.....	20
b) Double soulignement.....	20
c) Surlignement.....	20
2 Ajout, suppression, substitution et surcharge.....	21
a) Texte barré ou suppression.....	21
b) Ajout supralinéaire.....	21
c) Ajout infralinéaire.....	21
d) Ajout en marge de droite.....	22
e) Ajout en marge de gauche.....	23
f) Ajout en marge de haut de page	23
g) Ajout en marge de bas de page.....	23
h) Placer très précisément l'ajout ?.....	24
i) Substitution.....	24
j) Surcharge.....	25
3 Passage difficile à lire.....	26
a) Undeleted illegible text : Texte illisible non barré / texte non lu non barré.....	26
b) Deleted illegible text : Texte illisible barré / texte non lu barré.....	26
c) Questionable reading : « Lecture à débattre » / lecture hypothétique / lecture conjecturale.....	27

4 Instrument d'écriture (crayon à papier, feutre, stylo, plume, etc.).....	27
5 Couleur d'écriture.....	28
6 Langue.....	28
7 Changement de scripteur.....	28
8 Campagnes d'écriture.....	28
E) Notes et commentaires.....	29
1 Commentaires du transcripteur (invisibles pour l'internaute).....	29
2 Notes critiques courtes (visibles au survol).....	29
III. Conclusion.....	31
Tableau récapitulatif des balises.....	32
Proposition de présentation de barre d'outils.....	36

I. À la recherche d'un outil pour encoder les transcriptions des textes publiés dans e-Man

A) Introduction

Dans le cadre de nos projets éditoriaux sur e-Man, nous recherchons un outil susceptible de faciliter l'encodage (XML-TEI) des transcriptions des textes³. Certains des manuscrits dont nous publions les images numérisées sur e-Man peuvent poser des difficultés d'exploitation au lecteur non familiarisé avec l'écriture de l'auteur, il est donc important que le spécialiste puisse offrir une transcription de ces écrits afin d'en faciliter la lecture et l'exploitation. Lorsque les transcriptions des textes sont prêtes, nous les publions et les accompagnons des informations scientifiques disponibles au moment où celles-ci sont éditées. Dans le même temps, nous nous sommes dit qu'il fallait, lorsque cette transcription n'était pas encore disponible, élaborer un outil permettant à des personnes autorisées (équipe, chercheurs associés) d'apporter leur soutien au projet et de réaliser eux-mêmes la transcription qui les intéresse ou qu'on leur demande.

1 Présentation d'e-Man et enjeux de l'encodage des transcriptions en XML-TEI

Le projet de Plate-forme d'édition de Manuscrits modernes, e-Man, est développé au sein de l'Institut des textes et manuscrits modernes (ITEM). Il vise à créer un outil de publication interopérable favorisant l'exploitation numérique de documents. La plate-forme, conçue grâce au logiciel de création de bibliothèques numériques Omeka, a très vite accueilli des manuscrits numérisés « francophones » et « hispaniques ». Les notices décrivant les manuscrits numérisés sont enrichies progressivement. E-Man remplit son rôle de plate-forme d'édition de texte et de bibliothèque numérique.

Les notices des documents étant encodées en Dublin Core, un standard de jeux de métadonnées, la plate-forme e-Man est pleinement interopérable. Nous prévoyons donc de créer un entrepôt OAI-PMH afin de rendre les métadonnées d'e-Man moissonnables par des moteurs tels qu'Isidore ou Europeana et d'élargir ainsi la diffusion d'e-Man à un spectre plus large : chercheurs qui méconnaissent ces corpus, public d'amateurs curieux, etc.

Le potentiel d'e-Man dépasse toutefois la mise à disposition des manuscrits numérisés et de leurs métadonnées. En effet, les intérêts de recherches des responsables du projet portent sur l'étude des processus de la création (artistique, littéraire, dramaturgique). Ils travaillent à partir de la comparaison des différents témoins qui attestent de la genèse de l'oeuvre (manuscrits ou tapuscrits, corrections, ébauches, brouillons, sources extérieures à l'oeuvre, notes, correspondances). L'ensemble de ces éléments, appelé dossier génétique de l'oeuvre, pourrait être constitué sur la plate-forme e-Man. Comme le logiciel Omeka permet d'organiser les collections et de créer des liens entre les documents, ses fonctions vont être exploitées pour permettre l'édition numérique du dossier génétique.

Notre projet d'édition voulant être complet, il propose d'éditer les transcriptions de certains manuscrits. L'enrichissement que permet l'édition numérique des textes réside en la possibilité d'encoder ces derniers dans un langage XML structurant, comme celui de la *Text Encoding Initiative* par exemple. Dans le cas de la TEI, le Consortium a développé des balises spécifiques à l'étude des textes et un langage, le XML-TEI, qui sert à structurer les métadonnées d'une part, et la transcription de l'autre. En renseignant des métadonnées au sein du formulaire Dublin Core dans Omeka, il sera possible, à terme, de transférer automatiquement celles-ci dans le fichier XML de chaque manuscrit transcrit. La transcription, structurée elle aussi, respectera autant que possible la structure du texte dans la page du manuscrit. L'encodage XML-TEI sera également utilisé pour encoder des opérations génétiques spécifiques aux manuscrits comme les modifications effectuées par certains auteurs avant les mises au net ou la publication des œuvres. Adapter la TEI aux spécificités de nos corpus est également un enjeu du projet.

³ Le langage XML-TEI est, en effet, le langage informatique le plus reconnu pour l'encodage des manuscrits et imprimés.

2 À la recherche d'un outil de transcription en ligne

Nos contraintes de départ sont multiples : d'abord, nous souhaitons éviter de recourir à un logiciel d'encodage à cause du coût d'un tel outil, de sa difficile prise en main et du fait que les transcrip-teurs ne sont pas, systématiquement, des encodeurs professionnels. Ensuite, nous souhaitons effectuer les deux actions (transcriptions et encodage) sur une même plate-forme⁴ afin d'éviter une double manipulation informatique. Cela suppose donc qu'il nous faut trouver un outil ou une méthode compatible avec Omeka.

L'idéal, enfin, serait de pouvoir travailler sur cet encodage depuis une barre d'outils dotée de boutons qui permettent d'encoder la transcription en XML-TEI

En réalisant un état de l'art des outils de transcription disponibles à l'heure actuelle, nous avons mesuré les avantages et les inconvénients du plug-in Scripto d'Omeka, du wiki de transcription du Transcribe Bentham Project et de la barre d'outils intégrée au logiciel Omeka TinyMCE. Ce manuel fait état du résultat de ces investigations et propose des pistes de travail.

En ce qui concerne le plug-in Scripto d'Omeka, s'il donne, à première vue, la possibilité de gérer ensemble les interfaces de transcription et de saisie des notices au sein de la plate-forme Omeka, l'un des problèmes posés par cet outil de transcription collaborative est lié à la nécessité de passer par MediaWiki. Cela oblige donc le chercheur ou contributeur chargé de réaliser la transcription enrichie, à utiliser un wiki de transcription et à sortir de la plate-forme Omeka. Or l'une des ambitions du projet e-Man est de permettre la gestion, sur une même plate-forme, de l'édition des images et des transcriptions.

En revanche, le projet Bentham mené par les chercheurs de l'*University College London* a apporté des réponses intéressantes à nos recherches. Ce projet vise à diffuser les manuscrits de Jeremy Bentham et leur transcription en sollicitant la collaboration d'internautes. Il s'agit d'un projet ouvert dont le code est accessible. S'il a attiré notre attention, c'est parce que cette équipe a développé une interface simple destinée à faciliter l'encodage en XML-TEI des transcriptions saisies. Nous l'avons donc testé et nous sommes proposés de l'adapter à nos propres intérêts de recherche en intégrant à l'outil du projet Bentham, les éléments d'analyse génétique liés au processus de création.

À l'issue de différents tests, nous nous sommes rendus compte que si nous retenions la solution d'un wiki de transcription, nous pourrions faire un lien d'e-Man vers le wiki en retravaillant la charte graphique du wiki de sorte que l'internaute/utilisateur n'ait pas l'impression de sortir de la plate-forme e-Man (logo, couleurs, etc.). Le choix d'avoir deux plate-formes pose la question de la gestion des comptes et des mots de passe : gérer deux plate-formes veut dire gérer deux mots de passe par utilisateur. Si une telle solution était retenue, il faudra trouver une solution pour rediriger automatiquement les utilisateurs d'e-Man vers leur compte de transcription une fois connectés à la plate-forme e-Man. Cette solution, quoique complexe, n'est pas abandonnée.

Enfin, en ce qui concerne TinyMCE, nous l'avons comparé au plug-in Scripto et au MediaWiki et avons synthétisé, dans le tableau suivant, le résultat de cet examen des outils d'aide à la transcription :

	Wiki de transcription Bentham	Plug-in Scripto d'Omeka	TinyMCE barre d'outils d'Omeka
Type d'outil	Interface de transcription intégrée à MediaWiki	Plug-in d'Omeka passant par MediaWiki	Barre d'outils d'Omeka
Installation	Code disponible ✓	Installation réussie dans e-Man ✓	Installation par défaut dans Omeka de la barre d'outils minimale
	Étapes d'installation peu évidentes ✗		
	Formulaire de contact pour poser des questions ✓		

⁴ L'encodage, grâce à ses balises d'ajout et de suppression notamment, permet de structurer la transcription dans un langage normé.

Encodage	Barre d'outils intégrée ✓	Pas de barre d'outils intégrée ✗	Possibilité d'enrichir la barre d'outils de nouveaux boutons en HTML
	Encodage en XML-TEI et HTML	Par défaut, ne reconnaît que le langage MediaWiki (HTML simple et signes spécifiques à MediaWiki)	Possibilité de créer autant de boutons que souhaités ou nombre de boutons limité ?
			Possibilité d'importer une barre d'outils existante ?
Accès	Accès par le wiki	Accès selon les modalités du wiki auquel le plug-in est rattaché	Accès à la barre d'outils TinyMCE dans tous les formulaires d'e-Man
	Possibilité de faire un lien de la plate-forme e-Man vers le wiki	Accès public (et saisie) dans la plate forme e-Man	
	Possibilité de restreindre l'accès à la saisie des transcriptions par mot de passe	Possibilité de restreindre l'accès à la saisie des transcriptions par mot de passe	
	Possibilité de ne pas restreindre l'accès par mot de passe	Possibilité de ne pas restreindre l'accès par mot de passe	
Contrôle	Possibilité de rejeter une proposition de transcription	Obligation de passer par le wiki pour rejeter des propositions de transcriptions	Contrôle par e-Man
Gestion des comptes (et des mots de passe)	Deux comptes : e-Man + wiki	Deux comptes (e-Man + wiki) à partir d'une même plate-forme : possibilité de se connecter automatiquement sur le wiki en se connectant à e-Man ?	Gestion des comptes e-Man

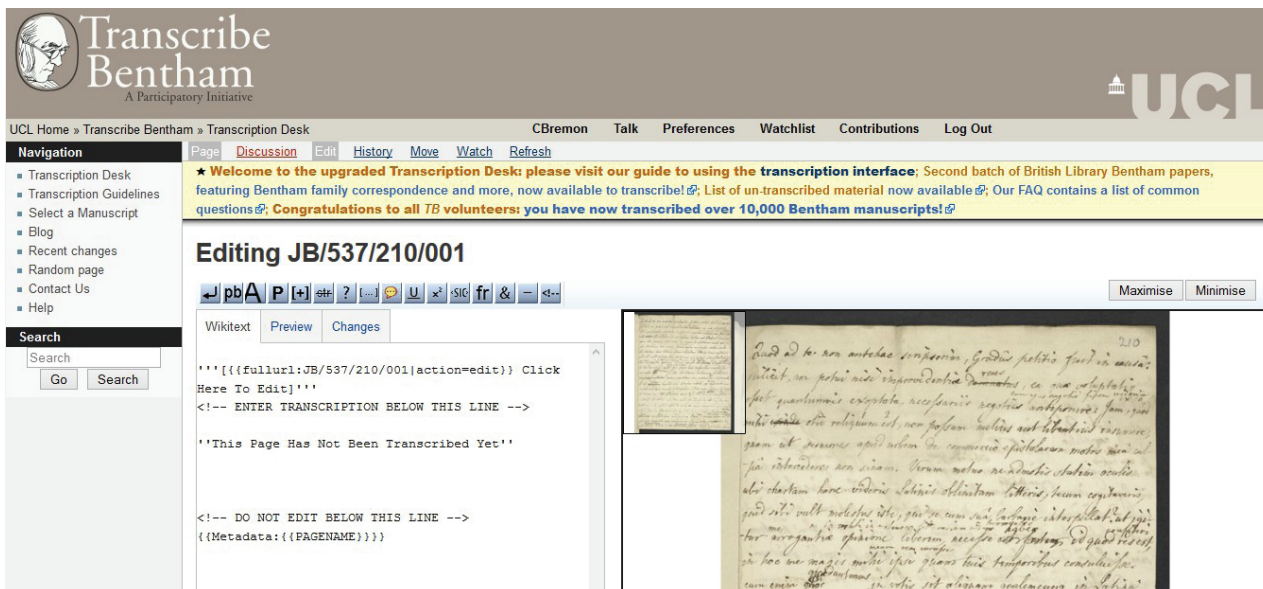
B) Une barre d'outils pour encoder la transcription en XML-TEI

Le projet Bentham a créé une barre d'outils qui permet d'encoder en XML-TEI. Le transcritteur n'a pas besoin de connaître par cœur les balises : comme avec un traitement de texte, il sélectionne le texte qu'il veut entourer de balises, clique sur l'icône correspondant à l'action. Lors de cette opération, une balise ouvrante et une balise fermante entourent le texte ou mot : cela simplifie la tâche du chercheur qui peut encoder un texte en XML-TEI et HTML sans être ralenti par l'apprentissage d'un langage informatique et sans risquer de faire une erreur d'inattention en tapant manuellement les balises.

Notre idée est d'intégrer à notre plate-forme une barre analogue à celle du projet Bentham et de l'enrichir de boutons additionnels pour encoder les phénomènes de génétique textuelle notamment. Pour cela, et selon la complexité de l'installation d'une telle barre d'outils dans Omeka, nous serions alors obligés de créer une interface nouvelle sous forme d'un wiki de transcription hors d'e-Man.

Dans le cas d'une barre d'outils intégrée à la plate-forme e-Man, seuls les membres du projet pourront transcrire dans l'interface administrateur de cette plate-forme.

Voici l'interface de saisie de transcription du projet Bentham. Elle comporte une barre d'outils (*toolbar*) dont les boutons permettent d'encoder le texte en XML-TEI et de visualiser le manuscrit sur le côté droit :



Pour visualiser les boutons de la barre d'outils et pour pouvoir saisir les transcriptions des manuscrits, le transcripateur du projet Bentham crée d'abord un compte dans le wiki de transcription. Grâce à ce compte, il peut proposer sa transcription. En cliquant sur « Select a Manuscript », celui-ci choisit un manuscrit à transcrire. Le transcripateur visualise le manuscrit sur la droite, peut zoomer sur l'image et la faire pivoter pour lire les ajouts en marge.

Voici la liste des seize boutons utilisés par le projet Bentham dans l'ordre où ils apparaissent dans leur interface de transcription. Nous les présentons avec leur description en anglais, leur traduction littérale et les balises XML-TEI ou HTML qu'ils utilisent pour l'encodage :

<i>Line breaks</i>	<i>Page breaks</i>	<i>Headings</i>	<i>Paragraphs</i>	<i>Additions</i>	<i>Deletions</i>	<i>Questionable reading</i>	<i>Illegible text</i>
Retour à la ligne	Saut de ligne	En-tête	Paragraphe	Ajout	Suppression	Lecture « à débattre »	Texte illisible
<lb/>	<pb/>	<head></head>		<add></add>		<unclear></unclear>	<gap/>
<i>Marginal notes</i>	<i>Underlined text</i>	<i>Superscript</i>	<i>Usual spellings</i>	<i>Non-English language</i>	<i>Ampersands</i>	<i>Dashes</i>	<i>User comments</i>
Notes en marge	Texte souligné	Exposant	Orthographe inhabituelle	Langue étrangère	Esperluette	Tiret long	Commentaire du transcripateur
<note></note>	<hi rend="underline"></hi>	<hi rend="superscript"></hi>	<sic></sic>	<foreign></foreign>	&	—	<!-- -->

A priori, tous les boutons proposés par le projet Bentham ne concernent pas à notre projet éditorial. Si nous ne pensons toutefois pas nous passer de la possibilité d'encoder tous les phénomènes, l'enjeu de notre travail consiste à créer les balises qui manquent ou à modifier les balises existantes. En effet, notre corpus est large et hétérogène et doit donc s'adapter à différents genres (théâtre, poésie, prose), à différentes dispositions du texte (pleine page, doubles colonnes, ajouts en marge (en haut, en bas, à gauche, à droite), etc.). Pour que la barre d'outils reste néanmoins parfaitement lisible, nous souhaitons organiser cette barre d'outils par « regroupements », à la manière d'un traitement de texte, mais en suivant l'organisation que nous avons établie dans la liste de la partie II de ce document (à partir de la page 10).



C) Étapes de l'encodage

1 De la transcription au fichier XML avec table des matières

L'encodage de la transcription se fait généralement dans un éditeur de texte XML tel que oXygen XML editor. Dans ce cas-là, la transcription peut être exportée dans un fichier au format .rtf ou .doc, par exemple.

Cette solution concerne surtout des textes imprimés facilement océrisables et peut difficilement s'envisager dans le cas d'une transcription de manuscrits d'écrivains couverts d'ajouts et de ratures.

La solution retenue pour l'encodage de manuscrits d'écrivains est donc souvent la transcription d'une page de manuscrit en vis-à-vis de l'image du manuscrit numérisé afin de pouvoir zoomer sur l'image et d'effectuer en même temps l'encodage des différentes opérations de transcription, notamment les ajouts et suppressions.

Nous souhaiterions combiner ces deux logiques : celle de la transcription d'une page de manuscrit après l'autre, et la constitution d'un fichier XML bien structuré et contenant l'œuvre entière.

La première étape consistera à transcrire et encoder les transcriptions au sein de chaque fichier correspondant à chaque page de manuscrit. Cette étape sera réalisée dans l'interface de transcription de la plate-forme e-Man.

Une seconde étape consistera à rassembler l'ensemble de ces fichiers en un seul comprenant toute la transcription du manuscrit. Cela permettra, à terme, de visualiser le texte intégralement mais également de constituer une table des matières et d'exploiter l'ensemble du manuscrit. En effet, seule la constitution d'un fichier contenant toute la transcription du manuscrit permet de naviguer dans le texte intégral.

Pour pouvoir générer une table des matières, nous devons, en outre, diviser la transcription en respectant les chapitres s'ils existent et en respectant les autres divisions inhérentes au manuscrit. Cela veut dire que si aucune division n'est manifeste, il ne faudra pas créer de divisions artificielles. En revanche, là où elles sont présentes (chapitres pour les romans, actes pour le théâtre, etc.), on les mettra en évidence grâce à des balises TEI appelées <div>.

2 Structure du fichier XML

Par ailleurs, on veillera à respecter la structure d'ensemble de l'œuvre qui se compose :

- du <teiHeader> au sein duquel on inscrit notamment la page de titre, certaines métadonnées et les noms des différents auteurs et réviseurs de la transcription et de l'encodage. Le teiHeader de chaque fichier XML de notre projet sera constitué à partir de la notice Dublin Core du manuscrit transcrit telle qu'elle a été remplie dans e-Man.
- et du <text> où est écrite la transcription.

Le <text> se compose lui-même :

- du <front>, qui est réservé aux pages préliminaires de l'œuvre (si elles existent). En effet, les pages préliminaires, telles que la préface, sont plutôt caractéristiques de l'œuvre imprimée et non du manuscrit d'écrivain⁵ ;
- du <body>, c'est-à-dire le corps de l'œuvre (où commence réellement le récit par exemple). Tout manuscrit comporte donc un <body> ;
- et, enfin, du <back>, qui est réservé aux pages post-liminaires de l'œuvre, si elles existent. En effet, comme les pages préliminaires, les pages post-liminaires, comme un index, un glossaire ou une bibliographie, sont plutôt caractéristiques de l'œuvre imprimée ou du manuscrit ancien que du manuscrit moderne.

3 Tout encoder ?

Enfin, nous rappelons que les boutons et balises que nous définissons ici sont au service du chercheur et non l'inverse. Chaque équipe définit ce qu'elle souhaite encoder pour chaque transcription. Nous proposons un maximum de balises afin que le choix ne soit pas restreint mais certaines balises peuvent être considérées comme facultatives dans certaines transcriptions. De même, si un chercheur souhaite ne transcrire qu'une partie d'un recueil, il peut le faire : le fichier XML restera valide.

D) Génétique des textes

La spécificité de l'Institut des textes et manuscrits modernes est l'analyse génétique des manuscrits d'écrivains d'époque contemporaine. Cela a plusieurs incidences sur l'encodage des transcriptions. Transcrire Proust ou Flaubert, par exemple, nécessite, en effet, un découpage de la page en différentes zones. Le projet Optima⁶ avait réalisé des tests pour établir un outil permettant de transcrire, par zones, les carnets de Flaubert ; d'autres expériences ont été menées sur les carnets de Proust et ont donné lieu à l'élaboration d'un prototype permettant de visualiser la page de manuscrit numérisée, découpée en zones de textes qui peuvent être affichées selon l'ordre chronologique de la rédaction de Marcel Proust. Pour élaborer cet outil et mettre en place cette visualisation du processus de l'écriture proustienne, le texte a d'abord été transcrit et encodé en XML-TEI avec des balises <zone>, ce qui permet de faire apparaître les transcriptions de chaque zone. L'encodage a également intégré la dimension chronologique afin d'afficher, l'une après l'autre, les étapes de l'écriture.⁷

Cette visualisation est très sophistiquée et n'est possible que lorsque l'auteur a écrit tous ses ajouts, corrections et notes éditoriales sur un même support, comme c'est souvent le cas avec les manuscrits d'écrivains comme Flaubert ou Proust qui, de plus, n'ont pas non plus réalisé d'innombrables versions de leurs écrits (ou n'en n'ont pas conservé tous les témoins intermédiaires).

Les auteurs du vingtième siècle, comme ceux des corpus francophones et hispaniques, ont utilisé, en plus des feuillets et carnets, des machines à écrire et multiplié ainsi les supports de l'écriture mais également les témoins génétiques. Ainsi, les dossiers génétiques de Mouloud Feraoun, Jean-Joseph Rabearivelo, Mora Guarnido ou Denis Molina offrent parfois jusqu'à huit états génétiques de leurs projets répartis sur plusieurs supports (carnets et feuillets libres, carnets et feuillets tapés à la machine, etc.). On peut représenter et encoder les corrections et changements intervenus dans le même état génétique. Par contre, nous devons trouver un moyen de représenter et d'encoder les corrections et les changements entre, par exemple, une page d'un carnet et les sept autres passages correspondants dans les autres manuscrits.

⁵ Cependant, le titre de l'œuvre figure sur les couvertures des cahiers des manuscrits francophones. Ces inscriptions feront partie du <front>.

⁶ Voir un article présente le projet Optima : <http://www.unicaen.fr/puc/images/preprint0062011.pdf>

⁷ Le projet est consultable en ligne "Prototype Proust": http://research.cch.kcl.ac.uk/proust_prototype/



II. Liste des balises adaptées aux besoins des équipes d'e-Man

Nous nous inspirons, pour construire cette liste, de manuels d'encodage et de recommandations existants, en particulier le manuel d'encodage des Bibliothèques Virtuelles Humanistes⁸ et les recommandations du projet Desgodets⁹.

Le Prototype Proust¹⁰ d'Elena Pierazzo (King's College, London) et Julie André (ITEM, Sciences Po Paris) nous a donné une base de réflexion sur laquelle nous appuyer pour l'encodage des campagnes d'écriture.

Enfin, lorsque ceux-ci existaient, nous avons utilisé les boutons développés par le projet Bentham pour leur barre d'outils d'encodage XML-TEI¹¹. Ensuite, nous avons décidé de créer ceux qui manquaient comme le montre la liste ci-dessous¹².

Un des défis serait de pouvoir sélectionner un type de support ou de genre et de pouvoir avoir uniquement les boutons correspondants. Par exemple, lorsque le chercheur sélectionne « théâtre » on lui proposera les boutons génériques et ceux spécifiques au théâtre et non à la poésie. Cela nous semble, à première vue, très complexe à mettre en œuvre.

A) Structure physique

Dans cette partie « Structure physique », nous faisons la liste des boutons servant à structurer la transcription en respectant la structure physique du document d'origine.

1 Caractères spéciaux

Afin qu'ils soient reconnus par les navigateurs, les caractères spéciaux suivants doivent être codés en HTML.

a) Esperluette (&)

Le code HTML qui correspond à l'esperluette est &

Le bouton , développé par le projet Bentham, nous permettra de coder ce caractère qui n'est pas reconnu par les navigateurs s'il est tapé par le clavier : « & ».

b) Tiret long (—)

De même, le tiret long doit être codé en HTML de la manière suivante : —

Le bouton qui permet de le faire est : 

⁸ <http://www.bvh.univ-tours.fr/XML-TEI/index.asp>

⁹ <http://www.desgodets.net/files/desgodetsDocumentation.pdf>

¹⁰ http://research.cch.kcl.ac.uk/proust_prototype/ (Référence du site du Prototype Proust déjà donnée précédemment)

¹¹ http://www.transcribe-bentham.da.ulcc.ac.uk/td/Help:Transcription_Guidelines

¹² La signification du bouton s'affichera lors du survol de ce dernier. Par exemple, en survolant le bouton du retour à la ligne, on verra s'afficher « retour à la ligne » ; en survolant le bouton de l'alinéa, on verra s'afficher « alinéa », et ainsi de suite.

Par ailleurs, nous ajouterons également un bouton qui permette d'importer le contenu de documents de traitements de texte (par exemple Word ou OpenOffice) sans les balises spécifiques à ces logiciels.

2 Indice et exposant mathématiques

Il existe un bouton dans l'interface de transcription du projet Bentham pour encoder les exposants : 

Ce bouton utilise les balises suivantes : `<hi rend="superscript"></hi>`

Sur le même modèle, nous créerons un bouton pour encoder les indices : 

Ce bouton utilisera les balises suivantes : `<hi rend="subscript"></hi>`

L'utilisation de ces boutons pour les projets d'e-Man sera probablement assez rare mais il est néanmoins important de les prévoir.

3 Retour à la ligne

Il est important d'encoder les retours à la ligne si l'on veut pouvoir lire la transcription en la comparant au manuscrit¹³ (quand la transcription et le manuscrit numérisé sont en regard dans l'interface de visualisation et de transcription).

Le bouton utilisé par le projet Bentham est le suivant : 

La balise XML-TEI qui se trouve derrière ce bouton est la suivante : `<lb/>`

Il s'agit d'une balise unique, c'est-à-dire qu'elle n'encadre pas la ligne à baliser mais elle se suffit à elle-même ; elle marque une rupture. Le projet Bentham préconise de cliquer sur le bouton correspondant directement après la fin de chaque ligne. Nous suivons cette recommandation qui nous paraît plus naturelle que de placer la balise `<lb/>` en début de chaque ligne. Ce bouton sera donc utilisé pour marquer la fin de chaque ligne de la transcription.

4 Coupure de mot en fin de ligne

L'encodage des césures se fait grâce à un attribut¹⁴ ajouté à la balise unique du retour à la ligne (`<lb/>`). Cet attribut est `@rend` et sa valeur est "hyphen" et ceux-ci permettent de placer automatiquement une marque de césure en fin de ligne afin que l'on puisse néanmoins faire des recherches plein texte sur le mot coupé. Dans le moteur de recherche, on pourra trouver le terme grâce à la balise « hyphen ».

Nous créerons donc un bouton qui permettra d'insérer la balise suivante : `<lb rend="hyphen"/>`

5 Saut de page

On utilisera un saut de page pour marquer la fin de chaque page de manuscrit transcrit : `<pb/>`

Une fois la page entièrement transcrite, le transcripateur pourra ainsi insérer le saut de page en cliquant sur 

6 En-tête et pied de page de documents imprimés <fw> (*Forme Work*)

Dans les imprimés anciens ou récents, on trouve des informations en en-tête et en pied de page : le titre courant en haut de page (`<header>`) ou en bas de page (`<footer>`), le numéro de page (`<pageNum>`), la signature ou marque de cahier (`<sig>`), ou la réclame (`<catch>`). Sont comptées comme inscriptions en en-têtes ou pieds de page les éléments qui reviennent quasiment systématiquement.

"header", "footer", "pageNum", "lineNum", "sig" et "catch" sont tous des valeurs de l'élément `<fw>` (pour *Forme*

¹³ Nous entendons par « manuscrit » tout document manuscrit ou tapuscrit figurant dans le dossier génétique de l'œuvre.

¹⁴ Les balises sont composées d'éléments (notés entre chevrons `<>`), qui peuvent avoir des attributs (notés `@`). Ces attributs ont eux-mêmes des valeurs (notées entre guillemets droits `"`).

Work¹⁵). Ce sont des valeurs de l'attribut @type. Celles-ci peuvent être complétées par des valeurs de l'attribut @place pour permettre un positionnement dans la page.

Ci-dessous un exemple¹⁶, extrait du manuel des Bibliothèques Virtuelles Humanistes, d'encodage du titre courant et du numéro de page :

```
<fw place="top-center" type="head">DE FOLIE ET D'AMOUR.</fw>  
<fw place="top-right" type="pageNum">19</fw>
```

L'encodage des titres courants, numéros de pages et autres informations des en-têtes et pieds de page de documents imprimés nous sera utile dans le cadre de l'encodage de documents tapuscrits ou si nous décidons d'encoder certaines éditions définitives.

De plus, il nous faut encoder toutes les inscriptions qui se trouveraient dans les en-têtes et pieds de page de cahiers manuscrits. En revanche, les annotations manuscrites seront encodées comme des ajouts de marge si elles ne se produisent qu'une fois ; si elles se reproduisent à chaque page ou régulièrement (par exemple, une pagination), elles seront alors encodées comme faisant partie de l'en-tête ou du pied de page¹⁷.

B) Structure logique

1 Division

Chaque page de manuscrit étant transcrite séparément, il est très rare qu'une division (chapitre, poème, scène, etc.) débute et finisse sur la même page de manuscrit, c'est pourquoi nous choisissons de créer un bouton pour ouvrir la division et un autre pour fermer celle-ci (et non un seul bouton pour entourer le texte sélectionné d'une <div> ouvrante et d'une </div> fermante).

Nous créerons un bouton <div> ouvrante :

```
<div>
```

Et un bouton </div> fermante :

```
</div>
```

Dans le champ Omeka « transcription », nous renseignons la division et le type quand ils sont présents. Quand il n'y a pas de division, il ne faut pas en créer.

Par exemple, pour encoder le premier chapitre d'une œuvre, nous utiliserons les balises suivantes :

```
<div n="1" type="chapitre">Texte</div>
```

La division, concernant plus spécifiquement le théâtre, la poésie et la correspondance, est abordée plus loin : p. 14 pour le théâtre, p. 15 pour la correspondance, p. 16 pour la poésie..

2 Tête de division : <head>

L'élément <head> permet , au sein d'une <div>, d'encoder le titre de la division s'il existe. Prenons par exemple le poème « Levant romantique » dans le recueil *Des stances oubliées* de Jean-Joseph Rabearivelo. C'est le premier poème du recueil.

Nous écrivons le titre du poème comme suit :

¹⁵ <http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-fw.html>

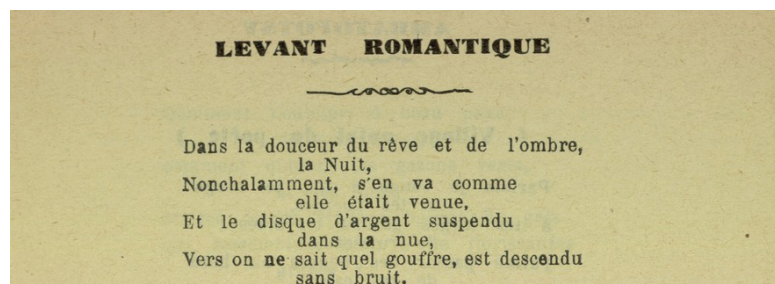
¹⁶ http://www.bvh.univ-tours.fr/XML-TEI/manuelTEIrenaissance3_2012.pdf (p. 23)

¹⁷ Pour les ajouts en marge de haut et de bas de page, voir les explications p. 23.


```

<div>
<head rend="center">
LEVANT ROMANTIQUE <lb/>
</head>
...
</div>

```



3 Paragraphe, alinéa et espace

a) Paragraphe

Le projet Bentham propose un bouton pour encoder les paragraphes. Pour cela, il suffit donc de sélectionner le texte du paragraphe et de cliquer sur : **P**

Le texte sélectionné sera alors entouré des balises : `<p> </p>`

Chaque paragraphe du manuscrit sera donc encodé de cette manière.

b) Alinéa

Le projet Bentham ne propose pas de bouton pour encoder les alinéas. Mais le manuel d'encodage des Bibliothèques Virtuelles Humanistes propose un attribut à ajouter au paragraphe où un retrait de première ligne a été effectué.

Il s'agit de l'attribut `@rend` avec la valeur "indent" : `<p rend="indent">`

L'alinéa fera l'objet d'un bouton à part entière :

c) Espace `<space>`

En parallèle des alinéas de paragraphes, il peut arriver qu'il y ait un espace blanc à l'intérieur d'un paragraphe (et non en début comme l'alinéa) ou entre des lignes de texte par exemple. Il est possible de reproduire cet espace dans la transcription grâce à l'élément `<space>` associé aux attributs `@unit` et `@quantity`.

Pour créer un espace de trois lignes, on pourra utiliser : `<space quantity="3" unit="lignes">`

Pour créer un espace de cinq centimètres, on utilisera : `<space quantity="5" unit="cm">`

Nous créerons un bouton `<space>` aux valeurs non remplies : donc celles-ci à la main.

`<space quantity="" unit="">`

On complètera

4 Centrer, aligner, justifier

Nous garderons les fonctionnalités « centrer », « aligner à gauche », « aligner à droite » et « justifier » qui existent dans la barre d'outil par défaut dans le logiciel Omeka et qui nous serviront à mettre en page le texte tapuscrit ou le manuscrit.

5 Texte théâtral

a) Acte, scène

Pour encoder l'acte ou la scène, au théâtre, on utilise la balise précédemment expliquée p. 12.

Le texte théâtral est généralement structuré en actes et en scènes. Actes et scènes feront partie de divisions :

`<div type="acte">` pour les actes ; `<div type="scène">`. Par exemple, pour l'acte I, on aura : `<div n="1" type="acte">`.

b) Réplique et locuteur

Le manuel des Bibliothèques Virtuelles Humanistes explique très bien l'emploi des éléments `<sp>` (pour *speech* ou réplique) et `<speaker>` (pour locuteur) dans les textes théâtraux :

```
<sp who="#amour">
  <lb/><speaker><name type="allegorie">AMOUR</name></speaker>
  <ab>Qui est cette folle qui me...</ab>
</sp>
<sp who="#folie">
  <lb/><speaker><name type="allegorie">FOLIE</name></speaker>
  <ab>Tu ne m'usses pù empêcher,...</ab>
</sp>
```

L'élément `<sp>` utilise l'attribut `@who` pour différencier les locuteurs.

L'encodage du texte théâtral est assez structuré : nous devons d'abord ouvrir une `<div>` acte puis une `<div>` scène, puis, à chaque changement de locuteur, le nom du locuteur et sa réplique sont tous deux inclus dans la balise `<sp>` :

```
<sp who="#nom_du_locuteur">nom du locuteur + réplique</sp>
```

À l'intérieur du `<sp>` englobant, les Bibliothèques Virtuelles Humanistes utilisent l'élément `<speaker>` pour indiquer le nom du locuteur tel qu'il apparaît physiquement dans le texte (en majuscules s'il est en majuscules).

Pour encoder le texte de la réplique, les BVH encodent avec l'élément `<ab>`. Mais, si l'on a déjà utilisé l'élément `<sp>`, il ne semble pas nécessaire d'encoder le texte de la réplique une seconde fois. Il ne nous semble pas non plus utile, dans le cadre de notre projet, d'encoder deux fois le nom du locuteur : à la fois dans la balise `<sp>` et dans la balise `<speaker>`. Nous créerons donc un bouton `<sp>` mais pas de bouton `<speaker>`.

En revanche, si la réplique présente un tiret long, le bouton développé par le projet Bentham est d'une grande aide pour coder celui-ci en HTML.

c) Didascalies

Comme l'explique le manuel des Bibliothèques Virtuelles Humanistes, on encode les indications de mise en scène (didascalies) grâce à l'élément `<stage>`

Voici un exemple tiré du manuel des BVH :

```
<stage>Folie se fait invisible, tellement, qu'Amour ne la peut assener.</stage>
```


6 Correspondance épistolaire

a) Support

On peut préciser le type de division (ici le support) dans lequel on se trouve dans le cadre d'une correspondance : une lettre, une carte postale, etc. Ces données sont renseignées dans le formulaire Dublin Core et apparaîtront dans le balisage du fichier final. Pour construire une table des matières, il faut indiquer tous les changements de divisions, y compris pour les lettres, et indiquer le début et la fin de la lettre grâce à une balise ouvrante et une balise fermante.

Dans le cas d'une lettre avec plusieurs pages, l'ouverture de la division se fera dans la transcription de la première page, et la fermeture dans la transcription de la dernière page :

```
<div type="lettre">    </div>
```

b) Sceau <stamp>

L'élément <stamp> permet d'encoder le sceau (celui des services de la censure durant la guerre d'Espagne par exemple). À l'origine, le Consortium TEI a commencé ses travaux sur l'encodage des livres anciens ; il n'avait donc pas prévu que l'élément <stamp> puisse encoder un timbre. Nous pouvons néanmoins adapter cette balise XML-TEI à notre besoin et encoder le timbre postal de la manière suivante :

```
<stamp type="postal">
```

S'il s'agit de censure, on pourra encoder le sceau de la manière suivante :

```
<stamp type="censure">
```

Le bouton que l'on emploiera aura la forme d'un timbre postal ou tout simplement : <stamp>

Le bouton placera les balises <stamp type=""> que l'on complétera la valeur à la main.

c) Date et lieu de rédaction <dateline>

Une lettre commence souvent par une ligne où figurent la date et/ou le lieu de rédaction de celle-ci. L'élément <dateline> permet de préciser ces informations. <dateline>, comme son nom ne l'indique pas, permet donc aussi d'indiquer le lieu.

Par exemple : <dateline>Paris 1^{er} Novembre 1954</dateline>

d) Formule d'appel ou formule finale <salute>

La formule d'appel (par exemple « Cher untel, ») ainsi que la formule finale de politesse (par exemple « Amicalement, ») sont encodées chacune grâce à l'élément <salute> :

Par exemple :

```
<salute>Cher untel,</salute>
```

```
<salute>Amicalement,</salute>
```

e) Signature

Lorsqu'une figure une signature en bas d'une lettre, on l'encode grâce à l'élément `<signed>`

L'élément `<signed>` permettra d'encoder toute signature figurant sur un manuscrit. Cet élément ne concerne donc pas seulement les lettres.

Si l'auteur de la lettre signe de son nom on cliquera d'abord sur le bouton `<signed>` puis sur le bouton qui permet d'encoder le nom : `<persName>`.

Par exemple :

```
<signed>
  <persName>
    <forename>Albert</forename>
    <surname>Camus</surname>
  </persName>
</signed>
```

f) Début et fin de la lettre `<opener>` `<closer>`

L'élément `<opener>` permet d'encoder tout ce qui se situe avant le cœur de la lettre (date et lieu de rédaction ainsi qu'une formule d'appel). L'extrait suivant tiré des *guidelines* du Consortium TEI nous en donne un exemple :

```
<opener>
<dateline>
<name type="place">Dorking, Surrey,</name>
<date>July 30th, 1849</date>
</dateline>
<salute>To <persName>Franklin Blake, Esq.</persName> Sir, —</salute>
</opener>
```

L'élément `<closer>` permet d'encoder tout ce qui se situe après le texte de la lettre (formule finale, signature).

Nous créerons donc deux boutons : `<opener>` et `<closer>`

7 Poésie

a) Poème et recueil

Chaque poème sera encodé comme `<div type="poème">`. Si d'autres structures sont présentes (par exemple le recueil), on utilisera également la balise `<div>` et on précisera le type. Par exemple, dans le cas du recueil : `div type="recueil">`.

b) Strophe et vers

L'élément `<lg>` (*linegroup*) permet d'encoder les strophes en poésie. Nous créerons donc le bouton <lg>

Au sein de la strophe, il est recommandé par le Consortium TEI, d'utiliser l'élément `<l>` et non l'élément unique `<lb/>` pour encoder les vers (*line*). C'est aussi le choix que nous ferons. Nous créerons donc un bouton <l>

8 Texte en prose : roman, essai, presse, etc.

Les types de divisions (par chapitre pour le roman ou l'essai, par article pour la presse, par poème pour la poésie, etc.) seront renseignés dans le formulaire Dublin Core. En revanche, la division devra néanmoins être encodée dans la transcription.

9 Liste, colonne et tableau

Le manuel des Bibliothèques Virtuelles Humanistes explique de manière claire et illustrée d'exemples l'encodage des listes, colonnes et tableaux¹⁸. Nous résumons ici ce que nous en retenons pour notre projet.

a) Liste `<list>`

L'élément `<list>` sert à encoder des listes simples. Nous créerons donc un bouton Liste pour encoder une liste simple. Une fois le texte de la liste sélectionné, le bouton placera une balise ouvrante `<list>` et une balise fermante `</list>` au début et à la fin de la liste.

Chaque élément de la liste est encodé grâce à une balise `<item>`. Chaque élément de la liste (item) est précédé d'une puce de forme variable. Nous créerons donc également un bouton Item.

La liste peut avoir un titre. Si c'est le cas, celui-ci s'encode grâce à l'élément Head qui donnera aussi lieu à un bouton.

Remarque : `<head>` n'a rien à voir avec l'en-tête des métadonnées du document ; `<head>` signifie ici « titre ».

b) Colonne `<cb/>`

Les colonnes s'encodent grâce à une balise unique `<cb/>` qui place un saut de colonne (*column break*). Nous créerons donc un bouton Colonne.

Les sauts de colonnes seront utiles dans tout texte écrit sur plusieurs colonnes : les articles de presse (Cf. le corpus de José Mora Guarnido dans les manuscrits hispaniques), certains poèmes (Cf. les poèmes du recueil *Presque-Songes* de Jean-Joseph Rabearivelo dans les manuscrits francophones), etc.

c) Tableau `<table>`

Créer un tableau en XML-TEI est plus complexe que de créer des colonnes ou des listes en HTML.

Voici comment la structure de l'encodage se présente :

¹⁸ http://www.bvh.univ-tours.fr/XML-TEI/manuelTEIrenaissance3_2012.pdf (p. 26 à 33)

```

<table>
<row cols="nombre_de_colonnes"
<cell>Contenu de la première cellule</cell><cell>Contenu de la seconde cellule</cell>
</row>
<row cols="nombre_de_colonnes"
<cell>Contenu de la première cellule</cell><cell>Contenu de la seconde cellule</cell>
</row>
</table>

```

Les balises <table>, <row> et <cell> ne servent à encoder que les tableaux faits par l'auteur ou les tableaux figurant sur un manuscrit. Selon leur fréquence, nous pourrions créer les boutons correspondants. L'équivalent en HTML de <row> est <tr> et de <cell> est <td>. La création de tableau existe aussi dans la barre Tiny MCE d'Omeka. Il est donc légitime de se demander quelle solution privilégier entre le XML-TEI et le HTML lorsque les deux langages proposent une solution différente pour encoder un même phénomène. Nous privilégierons, en général, le XML-TEI, qui est souvent plus précis que le HTML pour encoder des phénomènes liés à l'encodage de manuscrits. Mais, dans le cas précis du tableau, nous pourrions envisager de préférer l'encodage en HTML, qui est plus simple et tout aussi précis.

10 Abréviations

Le projet Bentham n'aborde pas l'encodage des abréviations.

En revanche, les Bibliothèques Virtuelles Humanistes proposent, dans leur manuel d'encodage, d'utiliser les balises <abbr> (pour l'abréviation) et <expan> (pour écrire la forme désabrévée du mot). Ces deux balises doivent être entourées de la balise <choice>, comme dans l'exemple qui suit tiré du manuel des Bibliothèques Virtuelles Humanistes¹⁹

```

<choice>
  <abbr>AM.</abbr>
  <expan>AMOUR</expan>
</choice>

```

La balise <expan> sert à faciliter la recherche et l'exploitation en recréant la forme complète du terme.

Nous créerons donc trois boutons :

<choice>

,

<abbr>

et

<expan>

11 Erreur et correction <sic> <corr>

Nous proposons d'encoder les erreurs d'orthographe grâce à la balise <sic> et d'en donner une correction grâce à la balise <corr> .

Cela peut s'avérer très utile dans l'exploitation des textes. En effet, si l'on veut faire une recherche sur le mot qui a été orthographié d'une manière inhabituelle. Par exemple, « génèse » pour « genèse » : on encodera « génèse » avec <sic> parce que ce n'est pas l'orthographe habituelle, et on encodera « genèse » avec <corr> pour que l'on puisse faire des recherches sur ce mot sans avoir à connaître l'orthographe spécifique de l'auteur :

¹⁹ Cf. le manuel d'encodage TEI – Renaissance et temps modernes des BVH, p. 36. (<http://www.bvh.univ-tours.fr/XML-TEI/index.asp>)

```
<choice>
  <sic>g  n  se</sic>
  <corr>gen  se</corr>
</choice>
```

Remarque : le terme *sic* ne sera pas visible   cran. Le terme corrig   peut, lui,   tre visible sous forme d'infobulle si on le souhaite, mais cela n'est pas n  cessaire.

C)    Indexation

Des balises nous permettent d'encoder les noms de personnes et de lieux afin de pouvoir les rechercher facilement dans le texte.

1 Noms de personnages ou de personnes

L'  l  ment <persName>, qui est g  n  rique, nous permet d'encoder n'importe quel nom de personnes (r  elles ou fictives).

Si le pr  nom et le nom de la personne sont indiqu  s et que l'on souhaite l'encoder, on utilisera l'  l  ment englobant <persName>,    l'int  rieur duquel on utilisera l'  l  ment <forename> pour le pr  nom et <surname> pour le nom de famille.

Par exemple :

```
<persName>
<forename>Mouloud</forename>
<surname>Feraoun</surname>
</persName>
```

Remarque : le « n » de « persName » prend une majuscule alors que le « n » de <forename> et de <surname> prend une minuscule.

Nous cr  erons donc les trois boutons suivants :

<persName>

<forename>

<surname>

Pour les noms englobant plusieurs personnes (par exemple des habitants), il existe l'  l  ment <orgName> (pour « organisation » ou « groupe »). Nous cr  erons donc   galement le bouton :

<orgName>

2 Noms de lieux

Les noms de lieux sont   galement encodables gr  ce aux   l  ments <placeName> et <geogName>.

L'  l  ment <placeName> permet de d  signer ville, village, pays, r  gion. On pr  cisera le type de lieu gr  ce    l'attribut @type et    la valeur d'attribut correspondante.

Nous cr  erons un bouton <placeName> assez englobant qui placera les balises :

```
<placeName type="type de lieu">Lieu</placeName>
```

On précisera ensuite le type de lieu à la main dans la balise.

L'élément <geogName> permet d'encoder les types de lieux naturels (montagnes, fleuves, etc.).

Nous créerons donc un bouton

<geogName>

D) Transcription

1 Soulignement, double soulignement, surlignement

a) Soulignement simple

Lorsqu'un mot du manuscrit est souligné, le transcripteur sélectionne le mot et clique sur le bouton  qui se chargera d'entourer le mot des balises XML-TEI suivantes :

<hi rend="underline"></hi>

En HTML, le soulignement se fait avec la balise <u>. Il est donc légitime de se poser la question de quel encodage choisir, puisque, dans ce cas précis, l'encodage en HTML semble aussi précis que l'encodage en XML-TEI.

b) Double soulignement

Le projet Bentham ne propose pas de bouton de soulignement double. Comme il n'existe pas de balise TEI pour encoder ce phénomène, nous pouvons décider soit de souligner d'un trait simple grâce aux balises <hi rend="underline"></hi>, soit d'utiliser les balises HTML suivantes de double soulignement (mais qui ne fonctionnent plus dès que le texte est sur plusieurs lignes)²⁰ :

Mot ou fragment souligné doublement

On pourra préférer d'encoder le double soulignement comme un soulignement simple ou comme soulignement double selon le choix qui sera fait pour chaque projet de transcription spécifique ; ces options sont en cours de discussion.

c) Surlignement

On parle de surlignement lorsque l'auteur a surligné du texte avec un marqueur fluorescent. Le surlignement n'a pas été prévu par le projet Bentham mais il est encodable grâce à une balise en langage HTML.

Par exemple, pour le surlignement en jaune (*yellow*) :

Nous créerons un bouton surlignement qui placera les balises

Il faudra alors taper à la main le nom de la couleur dans la balise : par exemple « blue » (pour bleu).

²⁰ Cf. <http://jsfiddle.net/cKNP4/>

2 Ajout, suppression, substitution et surcharge

a) Texte barré ou suppression

Nous entendons par « texte barré ou suppression » tout texte barré qui reste néanmoins lisible. En effet, le texte illisible barré est traité plus loin, dans la partie consacrée au « passage difficile à lire », p. 26.

A priori nous prévoyons d'utiliser la même balise pour encoder le texte supprimé par l'auteur, qu'il soit barré, raturé, haché, rayé d'une croix, etc. puisqu'il s'agit d'un même phénomène : la suppression. Là où le texte a été supprimé, le projet Bentham propose de sélectionner le texte en question et d'appuyer sur le bouton :  (pour *strikethrough* ou « barré »).

Ce bouton, que nous conservons, permettra d'entourer le texte supprimé des balises TEI suivantes : `<del></del>`.
deletion ou « suppression »).

En HTML, il existe, parallèlement à la balise `<del>`, la balise `<strike>`. Nous avons privilégié des balises XML-TEI dans la mesure où elles existaient parce qu'elles emploient un vocabulaire plus précis en matière de phénomène à encoder par rapport au HTML. `<del>` est reconnu par le Consortium TEI ainsi que par la dernière version du HTML5, tandis que `<strike>` n'est pas reconnu en HTML5. Nous préférons donc la balise `<del>` à la balise `<strike>`.

b) Ajout supralinéaire

Le bouton utilisé par le projet Bentham pour encoder les ajouts supralinéaires est : 

Ce bouton encode l'ajout avec les balises XML-TEI suivantes : `<add> </add>`

Ci-dessous un exemple d'ajout supralinéaire, concernant le mot « originally », tiré d'une transcription du projet Bentham²¹ :

Religion is ^{originally} an object by itself.

Le Consortium TEI propose les balises `<add>` avec un attribut : `<add place="above"></add>`

Cette proposition apporte une précision intéressante. Par conséquent, si nous souhaitons garder le bouton développé par le projet Bentham, nous le modifierons de telle sorte qu'il intègre l'attribut `@place` et sa valeur "above" en suivant les recommandations du Consortium TEI.

c) Ajout infralinéaire

Il semblerait que les manuscrits de Jeremy Bentham ne comportent pas d'ajouts infralinéaires (en dessous de la ligne de base) puisqu'il n'existe pas de bouton spécifique aux ajouts infralinéaires dans leur boîte à outils. Nous aurions donc pu choisir d'encoder les ajouts infralinéaires comme des ajouts supralinéaires en suivant le projet Bentham.

Nous préférons cependant créer un bouton qui nous permette de positionner l'ajout en position infralinéaire (en dessous de la ligne de base), si l'ajout est à cette position dans le manuscrit.

Nous créerons donc un bouton 

²¹ Cf. <http://www.transcribe-bentham.da.ulcc.ac.uk/td/JB/072/056/004>

Celui-ci permettra d'encoder l'ajout infralinéaire avec les balises XML-TEI préconisées par le Consortium TEI, qui propose les balises suivantes :

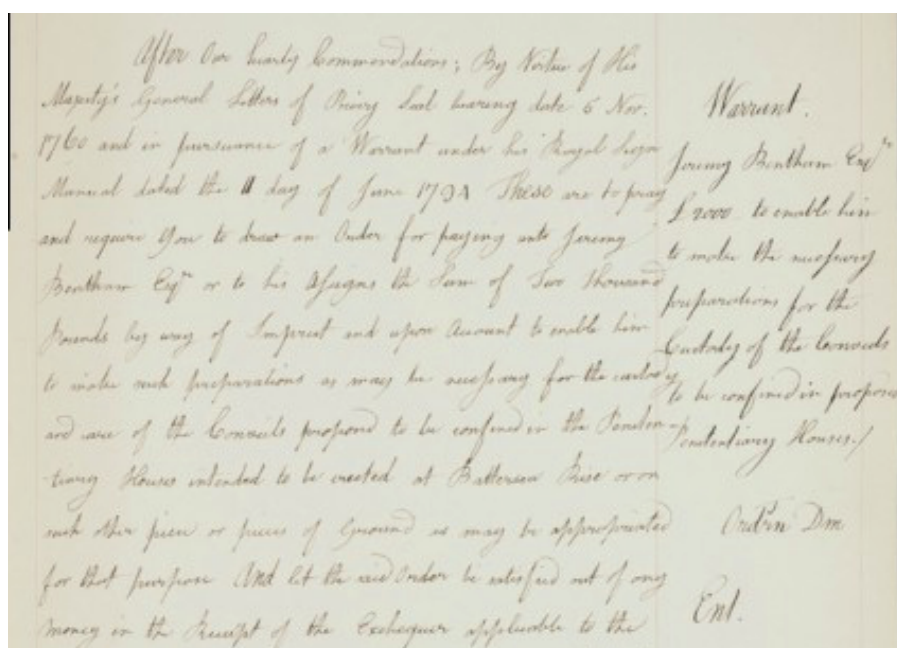
`<add place="below"></add>`

Remarque : nous excluons d'encoder l'ajout supralinéaire avec le bouton qui encode l'exposant (lettre ou chiffre).

d) Ajout en marge de droite

Les copies d'écran ci-dessous²² sont tirées du projet Bentham : la première capture d'écran représente une page de manuscrit numérisé (à droite de l'écran dans leur interface de transcription) ; la deuxième capture d'écran représente la page de transcription correspondant à la page de manuscrit mentionnée (en regard de l'image du manuscrit, à gauche de l'écran dans leur interface).

Sur la page de manuscrit, une marge (très discrète) a été tracée à droite, probablement par Bentham lui-même, et l'auteur y a inscrit du texte :



Dans la transcription, telle qu'on la voit sur la deuxième copie d'écran, la marge, encodée grâce aux balises `<note></note>` se matérialise par du texte encadré dans lequel il n'y a pas de retour à la ligne et qui vient se placer juste avant le premier mot du corps de texte :

²² Cf. <http://www.transcribe-bentham.da.ulcc.ac.uk/td/1B/115/089/002>

Warrant. Jeremy Bentham Esq^r £2000 to enable him to make the necessary preparations for the Custody of the Convicts to be confined in proposed Penitentiary Houses. / Ord^rn Dm Ent. After Our hearty Commendations ; By Virtue of His Majesty's General Letters of Privy Seal bearing date 5 Nov. 1760 and in pursuance of a Warrant under his Royal Sign Manual dated the 11 Day of June 1794 These are to pray and require You to draw an Order for paying unto Jeremy Bentham Esq^r or to his Assigns the Sum of Two thousand Pounds by way of Imprest and upon Account to enable him to make such preparations as may be necessary for the custody and care of the Convicts proposed to be confined in the Penitentiary

Ce rendu ne nous semble pas satisfaisant pour encoder les transcriptions de notre projet puisqu'il oublie la position spatiale de la marge. La position de ces ajouts doit être respectée car elle témoigne de phénomènes importants survenus durant le processus de création. Ces ajouts peuvent se situer : à droite, à gauche, en haut ou en bas. L'ajout en marge peut donc se trouver à ces quatre positions.

Or, le Consortium TEI propose d'encoder de façon précise le lieu où se trouve l'ajout grâce à un attribut que l'on associe à la balise <add>

Pour préciser qu'il s'agit d'une marge, les balises à placer sont les suivantes : `<add place="margin">`

Ensuite, pour indiquer qu'il s'agit d'une marge de droite, on complète la valeur "margin" en précisant "right" dans le cas de l'ajout de marge de droite :

`<add place="margin right">`

C'est donc cette balise, reconnue par le Consortium TEI, que nous recommandons. Nous créerons le bouton : 

Enfin, nous nous servirons de la balise <note>, que le projet Bentham utilise pour encoder les ajouts en marge, pour encoder des notes d'apparat critique, comme nous le verrons plus loin, p. 29.

e) Ajout en marge de gauche

La balise préconisée par le Consortium TEI s'écrit selon le même modèle :

`<add place="margin left">`

C'est donc la balise que nous utiliserons et nous créerons le bouton :



f) Ajout en marge de haut de page

Le Consortium TEI propose les balises suivantes :

`<add place="top">`

Nous utiliserons les balises recommandées par le Consortium TEI : l'élément <add> associé à l'attribut @place et à la valeur "top". Nous créerons donc le bouton :



g) Ajout en marge de bas de page

Pour encoder l'ajout en marge de bas de page, le Consortium TEI recommande les balises :

`<add place="bottom">`

Nous suivons ces recommandations et créerons, pour placer ces balises, le bouton :



h) Placer très précisément l'ajout ?

Il existe un moyen de placer l'ajout précisément dans la page grâce à une indentation dont on peut préciser l'importance en termes chiffrés et la localisation dans la page en termes de positionnement en hauteur et en longueur. C'est la solution qui avait été retenue par le projet "Prototype Proust". Cette solution nous semble trop technique face au nombre de corpus que nous avons à traiter. Bien sûr, si un tel niveau de précision reste possible dans le cas de certains manuscrits, nous proposons, pour notre projet, un outillage léger pour transcrire rapidement des grandes masses de manuscrits.

i) Substitution

Le projet Bentham gère les substitutions comme des suppressions combinées aux ajouts supralinéaires.

Les balises qui servent à encoder le texte supprimé sont : `<del></del>`

Et les balises qui servent à encoder le texte qui le remplace sont : `<add></add>`

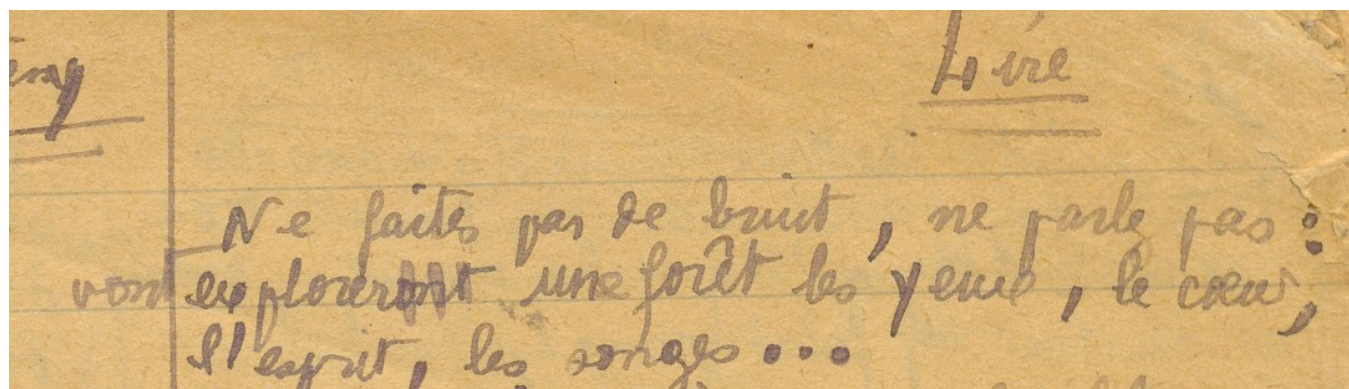
L'exemple suivant montre comment les substitutions sont représentées dans le projet Bentham²³ :

Tothill Fields, ~~and the said Lot B~~ containing in a low along the
said Tothill Fields by estimation ~~925~~ ¹⁵⁰⁰ feet or thereabouts,

Derrière cette représentation visuelle, l'encodage qui a été effectué est le suivant :

```
Tothill Fields, <del>and</del> <add>the said Lot B</add> containing in a low along the <lb/>  
said Tothill Fields by estimation <del>925</del> <add>1500</add> feet or thereabouts, <lb/>
```

Nous utiliserons cette manière d'encoder la substitution en tâchant de préserver un maximum d'informations. En effet, dans l'exemple suivant tiré du poème « Lire » du recueil *Presque-Songes* de Jean-Joseph Rabearivelo, la forme verbale « exploreront » a été remplacée par « vont explorer ». Mais l'auteur n'a pas barré le mot « exploreront » en entier : il a uniquement raturé la terminaison « ont » et a ajouté « vont » dans la marge de gauche :



²³ Cf. <http://www.transcribe-bentham.da.ulcc.ac.uk/td/JB/107/187/002>

L'encodage le plus simple est bien d'encoder « exploreront » comme suppression () et « vont explorer » comme ajout (<add>). Mais le résultat n'est pas entièrement satisfaisant comme le montre ce test d'encodage du recueil *Presque-Songes*²⁴ :

```
<head>LIRE</head>
<lg>
<l>Ne faites pas de bruit, ne parlez pas :</l>
<l>vont explorer <del rend="simple" source="#Msl">exploreront</del>
une forêt les yeux, le cœur,</l>
```

LIRE

Ne faites pas de bruit, ne parlez pas :
vont explorer ~~exploreront~~ une forêt les yeux, le cœur,
l'esprit, les songes

L'encodage de la surcharge nécessite davantage de précision : tout d'abord, la balise doit s'appliquer uniquement sur l'élément barré (« ont »). Ensuite, seul « vont » doit figurer en ajout (<add>). Enfin, afin de placer l'ajout en marge gauche, l'attribut @place et sa valeur "margin left" complèteront l'élément <add> comme nous le montre le nouveau test d'encodage effectué²⁵ :

```
<head>LIRE</head>
<lg>
<l>Ne faites pas de bruit, ne parlez pas :</l>
<l><add place="margin left">vont</add>explorer<del>ont</del>
une forêt les yeux, le cœur,</l>
<l>l'esprit, les songes
```

LIRE

Ne faites pas de bruit, ne parlez pas :
vont explorer ~~ont~~ une forêt les yeux, le cœur,
l'esprit, les songes

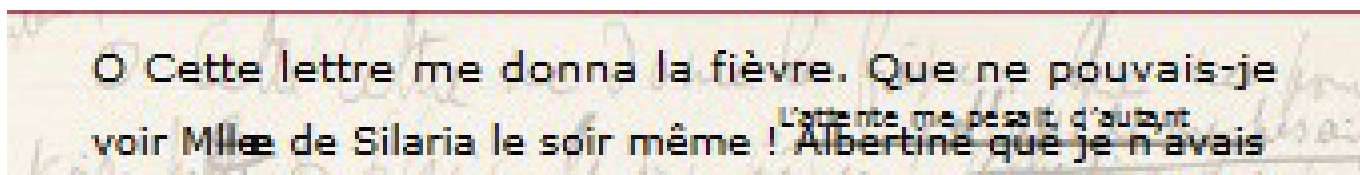
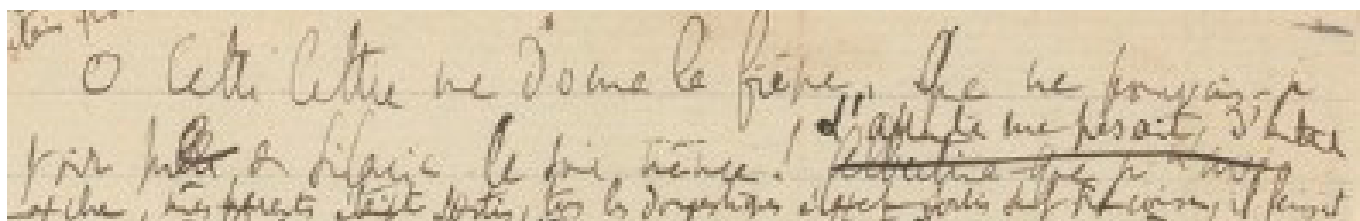
Dans ce nouveau test, la substitution est correctement encodée : la suppression porte uniquement sur la terminaison « ont », et « vont » est bien encodé comme ajout en marge de gauche. Le test de visualisation ayant été effectué avec l'outil TEI Boilerplate qui utilise une feuille de style XSLT par défaut, l'attribut @place et sa valeur "margin left" n'ont pas été correctement pris en compte. Ce défaut pourra être corrigé en modifiant la feuille XSLT de TEI Boilerplate ou un autre outil de visualisation XML-TEI. Le visualisateur TEI Boilerplate pourrait être utilisé à profit. Cette solution est donc à tester.

Nous venons de voir des exemples de substitution, où un élément est ajouté pour en remplacer un autre, c'est-à-dire où un ajout est combiné à une suppression.

j) Surcharge

Voyons maintenant un cas de substitution plus complexe, où l'ajout est écrit par-dessus la suppression. On parle alors de surcharge.

Dans l'exemple suivant, tiré du projet Prototype Proust, un « e » est écrit par-dessus les lettres « lle » du mot « mlle », au début de la deuxième ligne de l'extrait de manuscrit :



²⁴ Encodage effectué par Lydie Danjean dans un fichier XML consacré à *Presque-Songes* en 2014

²⁵ Test d'encodage effectué par Cécile Brémon sur un seul fragment du fichier XML de *Presque-Songes* le 29/04/2015

Dans cet exemple, l'encodage permet donc de montrer qu'un « e » a été écrit sur la terminaison de « Mlle ». Ce « e » constitue un ajout (<add>). À l'élément <add> a été ajouté un attribut @place dont la valeur est "superimposed" (superposé) :

```
<line rend="indent:170">voir M<del>lle</del><add place="superimposed">e</add> de  
Silaria le soir même ! <add place="above" rend="indent:490">L'attente me  
pesait, d'autant</add><del>Albertine que je n'avais</del></line>
```

Nous utiliserons cet attribut et cette valeur et créerons donc un bouton pour l'ajout superposé que l'on pourra appeler plus simplement

surcharge

et qui placera les balises suivantes :

```
<add place="superimposed">Ajout superposé</add>
```

Cet ajout superposé vient donc en complément de la balise pour encoder la surcharge. Reprenons l'exemple précédent :

```
M<del>lle</del><add place="superimposed">e</add>
```

Par conséquent, pour encoder une surcharge, nous utiliserons d'abord le bouton de suppression puis le bouton de surcharge.

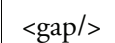
3 Passage difficile à lire

Le projet Bentham utilise la formule *illegible text* (texte illisible) pour exprimer la difficulté à lire le texte. Le projet Bentham divise cette catégorie *illegible text* en deux sous-catégories : *undeleted illegible text* (texte illisible non barré) et *deleted illegible text* (texte illisible barré). Lorsque l'on n'est pas sûr de la transcription d'un texte (souvent parce que le texte est difficile à lire), le projet Bentham propose la catégorie *questionable reading* (littéralement « lecture à débattre »).

Voici les trois solutions que le projet Bentham Project propose en termes de boutons et de balises et que nous validons :

a) *Undeleted illegible text* : Texte illisible non barré / texte non lu non barré

Dans le cas de texte illisible non barré, le bouton que le projet Bentham propose est le suivant : 

Ce bouton se traduit par la balise unique suivante : 

La balise unique <gap/> permet de placer des points de suspension encadrés par des crochets. Ce signe nous servira à représenter tout mot ou passage écrit illisible, sans que celui-ci soit barré.

b) *Deleted illegible text* : Texte illisible barré / texte non lu barré

Dans le cas de texte illisible barré, le projet Bentham propose la combinaison de deux boutons. Pour cela, il faut réaliser les trois manipulations suivantes :

on clique sur  , on sélectionne la balise <gap/> qui s'est installée, puis on clique sur  et on obtient : <gap/>

Visuellement, on obtient des points de suspension entre crochets, le tout barré : 

Nous utiliserons également cette solution pour rendre le texte illisible barré, que ce soit un mot ou plusieurs mots.

c) *Questionable reading* : « Lecture à débattre » / lecture hypothétique / lecture conjecturale

Dans le cas d'une lecture « à débattre » / lecture hypothétique / lecture conjecturale, le projet Bentham propose d'utiliser le bouton suivant : 

Ce bouton se traduit par les balises XML-TEI suivantes : `<unclear></unclear>` (littéralement « peu clair »)

Nous reproduisons un exemple de *questionable reading* tiré du projet Bentham²⁶ à propos du mot « Roods ». La première capture d'écran correspond à la visualisation et la seconde capture d'écran correspond à l'encodage :



`fifty three Acres two <unclear>Roods</unclear> and fourteen Perches<lb/>`

Le mot « Roods » a été proposé par le transcripateur mais celui-ci n'est pas sûr de son choix de transcription. Il l'indique donc par l'élément `<unclear>` qui, à la visualisation, vient souligner le mot (ou le groupe de mots) de pointillés et placer un point d'interrogation entre crochets à la droite du mot pour représenter l'incertitude. Ces éléments sont grisés et ne peuvent donc pas être confondus avec le texte de l'auteur, ce qui nous satisfait donc pleinement. Par conséquent, nous adoptons cette balise.

4 Instrument d'écriture (crayon à papier, feutre, stylo, plume, etc.)

On peut se servir de la balise unique `<handShift/>` et de son attribut `@medium` pour encoder le type d'instrument utilisé pour l'écriture : le crayon à papier, le feutre, le stylo, la plume, etc. :

```
<handShift medium="crayon à papier"/>
<handShift medium="feutre"/>
<handShift medium="stylo"/>
<handShift medium="plume"/>
<handShift medium="stylo-plume"/>
```

(Liste non exhaustive)

Remarque : la balise `<handShift/>` se place à l'endroit où il y a un changement. On peut néanmoins en placer une première balise au début de la transcription de la première page du manuscrit afin de spécifier le type d'instrument dès le début du document.

Nous créerons un bouton



qui placera la balise vide

```
<handShift medium="type d'instrument"/>
```

On complétera ensuite la valeur d'attribut à la main ou grâce à une liste déroulante qui comprendra les différents instruments d'écriture.

²⁶ Cf. <http://www.transcribe-bentham.da.ulcc.ac.uk/td/JB/107/187/002>

5 Couleur d'écriture

La barre d'outils comportera une palette de couleurs en HTML pour permettre de modifier la couleur du texte afin d'être fidèle à l'apparence du texte du manuscrit.

Dans le cas d'un texte bleu, la barre d'outils par défaut dans e-Man propose l'encodage suivant :

```
<span style="color: #0000FF;"> Texte bleu cyan</span>
```

Nous conserverons donc le bouton de la barre Tiny MCE : 

6 Langue

Le fichier XML commencera par la ligne

```
<?xml version="1.0" encoding="UTF-8"?>
```

 afin de résoudre les problèmes d'accents et d'encodage des caractères.

Dans le cas de textes d'écrivains qui évoluent dans différents environnements linguistiques (lorsque, par exemple, le français côtoie une autre langue – malgache, arabe, etc.), il peut arriver qu'au sein d'un texte figurent des mots écrits dans l'autre langue.

Pour encoder une langue autre que la langue principale, le projet Bentham propose le bouton : 

(pour *foreign*, mais nous créerons un autre bouton pour éviter l'ambiguïté avec fr=français.)

Par exemple :

```
<foreign>
```

 (qui permettra de placer les balises XML-TEI suivantes :

```
<foreign xml:lang="nom de la langue"> Texte</foreign>
```

La langue sera complétée à la main entre les guillemets de la balise :

```
<foreign xml:lang="mg"> mot en langue malgache</foreign>
```

On ne pourra, en revanche, pas faire de liste déroulante des différentes langues parce qu'il y a trop de cas de figure possibles. Ces valeurs seront rajoutées à la main.

7 Changement de scripteur

On encode le changement de scripteur (ou changement de main) grâce à la balise unique `<handShift/>` et à son attribut `@scribe`.

Nous créerons donc un bouton pour encoder le changement de scripteur :

Ce bouton placera une balise unique :

```
<handShift medium="scribe"/>
```



8 Campagnes d'écriture

Les campagnes d'écriture correspondent aux différents changements qui sont intervenus sur une page de manuscrit à différentes étapes de révision.

Le projet Prototype Proust propose l'élément `<change>` pour encoder ces campagnes d'écriture. Ces étapes dans l'écriture sont numérotées dans l'ordre chronologique. Dans le projet Prototype Proust, l'élément `<change>` est associé à l'élément `<zone>`. Cependant, ces zones ne peuvent accueillir des paragraphes mais seulement des `<line>` comme en poésie. Nous proposons donc d'utiliser l'attribut `@change`.

Le plus simple, pour notre projet, sera de numéroté (à partir de 01) les différentes campagnes d'écriture grâce à

l'attribut @type et la valeur d'attribut "num"²⁷ (toute valeur numérale).

Nous créerons donc un bouton **Campagnes** qui permettra de placer les balises suivantes :

```
<change type="Campagne 01"></change>
```

Remarque : il est nécessaire de définir au préalable les campagnes d'écriture afin de pouvoir les encoder.

E) Notes et commentaires

1 Commentaires du transcripteur (invisibles pour l'internaute)

Le bouton que propose le projet Bentham pour ajouter un commentaire est le suivant :



Les balises HTML que ce bouton introduit sont les suivantes :

```
<!--Ceci est un commentaire-->
```

Ces balises permettent de commenter la transcription de manière invisible pour l'internaute. Ces commentaires seront utiles pour que chaque transcripteur puisse documenter sa transcription et pour que l'informaticien puisse documenter le code.

2 Notes critiques courtes (visibles au survol)

Nous utiliserons l'élément <note> pour encoder des notes critiques courtes qui correspondront à l'apparat critique des transcriptions. Dans ces notes, le transcripteur pourra, par exemple, indiquer que le passage concerné est :

- une intervention éditoriale,
- une citation ou référence bibliographique,
- la raison pour laquelle un mot est illisible ou difficile à lire (autre que barré) : tache d'eau, trou, etc.
- etc.

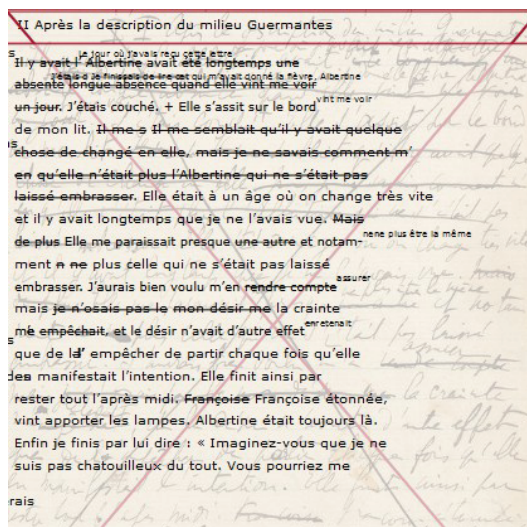
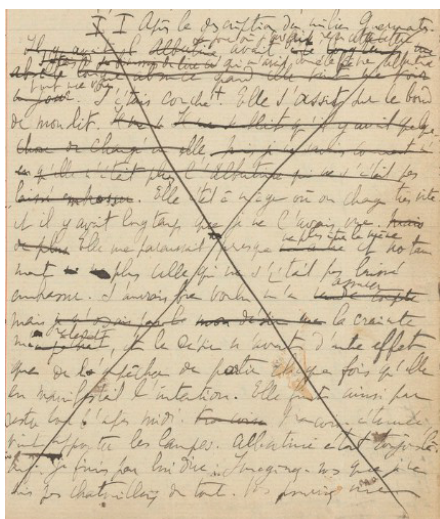
Nous utiliserons donc le bouton  du projet Bentham qui encadrera la note des balises ouvrante <note> et fermante </note>.

Ces notes critiques seront courtes et visibles dans une infobulle au survol du passage annoté.

Le transcripteur pourra simplement écrire la note entre la balise ouvrante <note> et la balise fermante </note> au sein de la transcription. Cette situation peut concerner un passage dont on n'a pas pu encoder toutes les subtilités ou l'ajout d'informations nécessaires pour la poursuite de la lecture. Il y a des phénomènes qu'on ne peut encoder, d'où l'utilité de la balise <note>.

Dans l'exemple suivant, le projet Proust a réussi à encoder la croix de Saint-André (voir la capture d'écran de droite) :

²⁷ « <num> (numéral) contient un nombre écrit sous une forme quelconque » (Cf. les *guidelines* du Consortium TEI : <http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-num.html>)



L'encodage de la croix de Saint-André demande une technicité que nous ne pouvons mettre en œuvre vu l'ampleur de notre projet. Nous indiquerons donc ce procédé au sein d'une note critique courte qui se verra au survol : par exemple « paragraphe marqué d'une croix ».

III. Conclusion

Après avoir trouvé la solution pour publier les images numérisées des documents et manuscrits des fonds d'archives des écrivains contemporains issus de domaines hispaniques et francophones notamment, la barre d'outils que nous proposons est une solution pour simplifier la tâche du transcripteur soucieux d'utiliser un outil simple pour se concentrer sur l'analyse et la lecture du texte. La barre d'outils que nous lui proposons est un tableau de bord fonctionnel. Contrairement au projet Bentham, le public visé est un public scientifique, qui veut être précis dans la réalisation de sa transcription en vue de construire une étude critique ou une édition numérique ou/et papier. C'est pour cela que nous n'ouvrons pas la transcription aux visiteurs du site.

À l'issue de cette étude de faisabilité, l'accès au code développé par le projet Bentham et les différents tests effectués donne les moyens de choisir l'interface dans laquelle nous implanterons la barre d'outils (e-Man ou wiki de transcription ?) et de créer les boutons décrits dans ce guide. Le choix devra être effectué à partir des tests informatiques poussés, aussi bien sur la barre outils et sur la liaison entre Omeka et MediaWiki.

Une fois les boutons créés, nous nous proposons d'en conserver la maniabilité grâce à des regroupements en six ensembles, autour d'une lecture logique. Ces cinq ensembles ont été décrits dans les six points de ce guide :

A/ Structure physique

B/ Structure logique

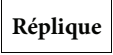
C/ Indexation

D/ Transcription

E/ Notes et commentaires

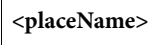


TABLEAU RÉCAPITULATIF DES BALISES

	Phénomène à encoder	Barre d'outils de transcription : Boutons Bentham Project ²⁸ , ou à créer		Balises XML-TEI ou HTML correspondantes	
Structure physique	Caractères spéciaux	Esperluette		&	HTML
		Tiret long		—	HTML
	Indice et exposant mathématiques	Exposant		<hi rend="superscript"></hi>	XML-TEI
		Indice		<hi rend="subscript"></hi>	XML-TEI
	Retour à la ligne			<lb/>	XML-TEI
	Coupure de mot en fin de ligne (césure)			<lb rend="hyphen"/>	XML-TEI
	Saut de page			<pb/>	HTML
	En-tête et pied de page de documents imprimés	En-tête		Par exemple, le titre courant en haut à gauche : <fw place="top-left" type="head"></fw> Valeurs acceptées : left, right, center	XML-TEI
		Pied de page		Par exemple, le numéro de page, en bas à droite : <fw place="bottom" type="pageNum"></fw>	XML-TEI
Structure logique	Division (deux boutons)	<div> ouvrante		Par exemple, la <div> ouvrante pour un chapitre : <div type="chapitre">	XML-TEI
		</div> fermante		S'applique à tout type de division : </div>	XML-TEI
	Tête et fin de division	<head> (Titre de division, par exemple centré)		<head rend="center"></head>	XML-TEI
		<trailer> (Mention explicite de fin de division, par exemple centré)		<head rend="center"></head>	XML-TEI
	Paragraphe et alinéa	Paragraphe			HTML
		Paragraphe avec alinéa		<p rend="indent"></p>	XML-TEI
	Texte théâtral	Réplique et locuteur		<sp who="#Nom du locuteur">Nom du locuteur + réplique</sp>	XML-TEI

28 Présentation du Transcribe Bentham Project (en anglais) : <http://blogs.ucl.ac.uk/transcribe-bentham/>

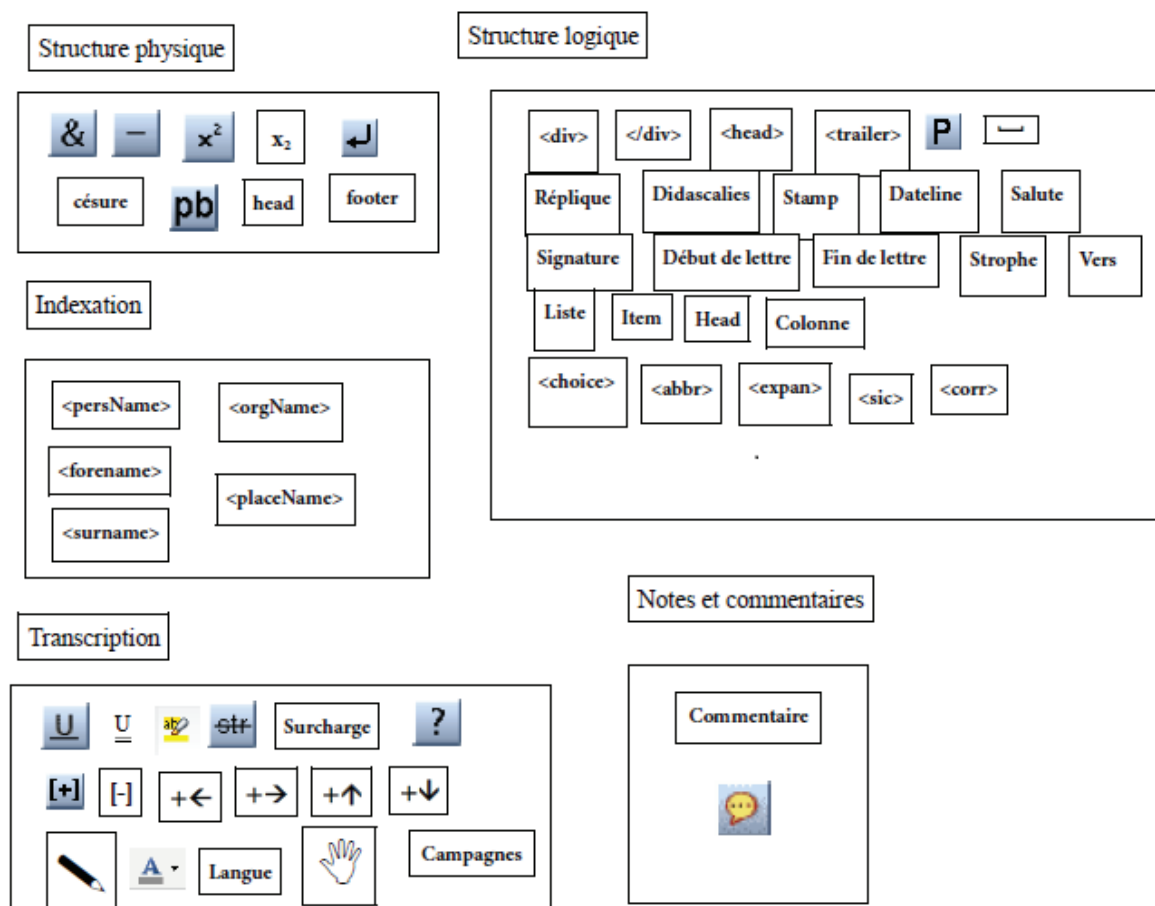
		Didascalies	Didascalies	<stage></stage>	XML-TEI
	Correspondance épistolaire	Sceau ou timbre	Stamp	<stamp type="censure"></stamp>	XML-TEI
		Date et lieu de rédaction	Dateline	<dateline></dateline>	XML-TEI
		Formule d'appel ou formule finale	Salute	<salute></salute>	XML-TEI
		Signature	Signature	<signed></signed>	XML-TEI
		Début et fin de la lettre	Début de lettre	<opener></opener>	XML-TEI
			Fin de lettre	<closer></closer>	XML-TEI
	Poésie	Strophe et vers	Strophe	<lg></lg>	XML-TEI
			Vers	<l>	XML-TEI
	Liste, colonne et tableau	Liste	Liste	<list></list>	XML-TEI
			Item	<item></item>	
			Head	<head></head>	
		Colonne	Colonne	<cb/>	HTML
	Abréviation	Ensemble	<choice>	<choice></choice>	XML-TEI
		Abréviation	<abbr>	<abbr></abbr>	
		Désabréviation	<expan>	<expan></expan>	
	Erreur et correction	Ensemble	<choice>	<choice></choice>	XML-TEI
		Orthographe inhabituelle	<sic>	<sic></sic>	
		Correction	<corr>	<corr></corr>	
Indexation	Noms de personnages ou de personnes	Nom de personne ou de personnage	<persName>	<persName></persName>	XML-TEI
		Prénom	<forename>	<forename></forename>	XML-TEI
		Nom de famille	<surname>	<surname></surname>	XML-TEI
		Nom de groupe	<orgName>	<orgName></orgName>	XML-TEI

Transcription	Noms de lieux	Nom de lieu		<code><placeName type="à compléter"></placeName></code>	XML-TEI
	Soulignement et surlignement	Soulignement		<code><hi rend="underline"></hi></code>	XML-TEI
		Double soulignement		<code></code>	HTML
		Surlignement		Par exemple, jaune : <code></code>	HTML
	Suppression	Suppression		<code></code>	HTML
	Ajout	Ajout supralinéaire		<code><add place="above"></add></code>	XML-TEI
		Ajout infralinéaire		<code><add place="below"></add></code>	XML-TEI
		Marge à gauche		<code><add place="margin left"></add></code>	XML-TEI
		Marge à droite		<code><add place="margin right"></add></code>	XML-TEI
		Haut de page		<code><add place="top"></add></code>	XML-TEI
		Bas de page		<code><add place="bottom"></add></code>	XML-TEI
	Substitution		  +  ou 	(Voir aux boutons correspondants : suppression et ajout)	XML-TEI
	Surcharge			<code><add place="superimposed"></add></code>	XML-TEI
	Passage illisible ou difficile à lire	Texte illisible (ou non lu) barré	 	<code> <gap/> </code>	XML-TEI + HTML
		Questionable reading (Lecture « à débattre », lecture conjecturale, lecture hypothétique)		<code><unclear></unclear></code>	XML-TEI
	Instrument d'écriture			<code><handShift medium="crayon à papier"/></code>	XML-TEI

	Couleur			La couleur sera automatiquement définie par la palette Texte bleu cyan	HTML
	Langue	Langue côtoyant la langue principale dans le texte	Langue	<foreign xml:lang=""></foreign>	XML-TEI
	Changement de scripteur	Changement de main (Préciser : auteur, éditeur, membre de la famille)		<handShift medium="scribe"/>	XML-TEI
	Campagnes d'écriture		Campagnes	<p type="Campagne 01">	XML-TEI
Notes et commentaires	Notes et commentaires	Commentaires invisibles	Commentaire	<!--Commentaire invisible-->	HTML
		Notes critiques courtes (apparat critique)		<note></note>	XML-TEI

PROPOSITION DE PRÉSENTATION DE BARRE D'OUTILS

Nous faisons ici une proposition de présentation de notre barre d'outils XML-TEI. Le nom des boutons et l'image qui leur sont associés pourront être modifiés là où une amélioration aura pu être trouvée. Les boutons provenant de la barre du projet Bentham ont été conservés, ainsi que certains boutons de la barre Tiny MCE d'Omeka. Une solution doit être trouvée pour permettre un gain de place : réduction de la taille des boutons, ouverture d'un menu, etc.





Documentation :

- ✚ les *guidelines* (recommandations) du *Transcribe Bentham Project* (*University College London*) (http://www.transcribe-bentham.da.ulcc.ac.uk/td/Help:Transcription_Guidelines, mai 2010)
- ✚ les *guidelines* (recommandations) du Consortium *Text Encoding Initiative* (TEI) (<http://www.tei-c.org/index.xml>)
- ✚ les *guidelines* (recommandations) du projet d'archives manuscrites éthiopiennes EMA
- ✚ le manuel d'encodage des Bibliothèques Virtuelles Humanistes (Tours) (<http://www.bvh.univ-tours.fr/XML-TEI/index.asp>, 25 juillet 2012)
- ✚ le manuel d'encodage pour le projet *Desgodets* : (<http://www.desgodets.net/files/desgodetsDocumentation.pdf>, 17 mai 2013)
- ✚ le manuel d'encodage de Bénédicte Vauthier pour l'édition numérique des manuscrits de travail de *Paisajes después de la batalla* de Juan Goytisolo (<http://fliphtml5.com/xocf/iucx>, 2013)
- ✚ les *guidelines* (recommandations) pour l'encodage des archives de Walt Whitman (http://www.whitmanarchive.org/mediawiki/index.php/Whitman_Encoding_Guidelines)
- ✚ la liste de balises établie par Sylvie Giraud, pour l'encodage des manuscrits de Flaubert (équipe Flaubert) le 11 février 2015





Recommandations pour le traitement génétique des documents dans e-Man

Recommandations rédigées par Cécile Brémon¹
Sous la responsabilité de Richard Walter²



Version n°3 du 29 mai 2015

Nous listons ici différentes hypothèses de travail sur l'encodage des états génétiques d'un document. Ensuite, nous proposons des solutions pour la constitution du dossier génétique et la définition des relations entre documents.

Pour l'analyse génétique, il faut se reporter au document « Édition numérique des avants-textes » de Fatiha Idmhand et Claire Riffard.

I. Encoder les variantes génétiques dans un état ?

Voici des hypothèses pour réfléchir aux différentes modalités de transcription des évolutions du texte à partir d'une unité de base : le premier avant-texte.

1 Encoder les variantes du texte dans la transcription ?

Notre réflexion porte sur l'encodage des mouvements entre le manuscrit transcrit, l'édition définitive et les autres avant-textes de l'œuvre, ce qui reviendrait à encoder le dossier génétique de l'œuvre concernée. Pour cela, nous proposons d'adapter des balises qui servent normalement à encoder la régularisation d'une orthographe ancienne dans un texte. En effet, nous ne rencontrons pas ce phénomène dans nos corpus. Nous pourrions donc emprunter ces balises pour nous en servir autrement.

En effet, il existe une balise <orig> pour *original form* qui permet d'encoder la forme originelle d'un mot ou d'une expression et d'en proposer sa régularisation. Les éléments <orig> et <reg> sont contenus dans l'élément <choice> comme le montre l'exemple suivant extrait du manuel des Bibliothèques Virtuelles Humanistes :

```
<choice>
  <orig>EVRRES</orig>
  <reg>EUVRES</reg>
</choice>
```

¹ cecile.bremon@gmail.com

² richard.walter@ens.fr

Nous proposons d'adapter cette balise <orig> à la forme d'un terme ou d'un passage dans le manuscrit qu'on transcrit, d'indiquer, grâce à l'élément <reg> la forme de ce terme dans tel ou tel autre avant-texte, et de compléter l'élément <reg> d'un attribut @type et d'une forme codifiée (« avant-texte ») et d'une valeur « numérale ». On peut très bien imaginer une suite de « régularisations » sous la forme originelle au sein des balises <choice>.

Nous créerons donc trois boutons : <choice> , <orig> et <reg> .

Le terme de « régularisation » ne correspond pas exactement à ce que nous voulons montrer mais « original » ou « originel » correspond à l'idée du point de départ. Or, seuls <reg> et <corr> peuvent s'employer avec <orig>. Cependant, nous utilisons déjà <corr> dans le sens de correction orthographique. Nous adaptons donc la balise <reg>, définie pour encoder une orthographe régularisée, pour encoder les variantes d'un texte au sein des avant-textes successifs.

Par exemple :

```
<choice>
  <orig>Passage du 1er manuscrit, avant-texte 1 [Avant-texte 1]</orig>
  <reg>Variante dans l'avant-texte 2 [Avant-texte 2]</reg>
  <reg>Variante dans l'avant-texte 3 [Avant-texte 3]</reg>
</choice>
```

À l'écran, les différentes versions du texte, encodées comme <reg>, apparaîtraient au survol dans des infobulles. L'internaute verrait alors les différentes variantes qui existent dans le dossier génétique. La place de ces variantes dans le dossier génétique serait explicitée en ajoutant, par exemple [Avant-texte 2] pour indiquer que la variante se trouve dans l'avant-texte 2 du dossier génétique.

Remarque : l'affichage des variantes que l'on propose suppose de connaître, au préalable, l'ordre chronologique des variantes dans le processus de création. Par ailleurs, cet affichage ne peut pas concerner des corrections sur de grandes parties de texte comme, par exemple, le déplacement ou la suppression de chapitres entiers. Dans le cas de suppression dépassant une page, celle-ci sera indiquée en notes, dans ce que nous avons appelé les notes critiques courtes. En définitive, l'encodage des variantes du texte ne pourrait concerner que les variantes à l'intérieur d'une page transcrite. La granularité de cet encodage doit être définie en fonction des cas de figure (par exemple dans le cas d'une variante se trouvant à cheval sur deux pages).

À noter : selon le choix fait au niveau de l'encodage, l'ordre peut être inversé : par exemple, si l'on choisit d'encoder le deuxième avant-texte, celui-ci sera considéré comme <orig> dans cette transcription et les variantes du premier avant-texte seront encodées avec <reg>, de même que les variantes des avant-textes ultérieurs. Ainsi de suite pour chaque transcription. Cela risque de poser des problèmes de décalage et de gestion de contenus sur le long terme.

2 Transcrire les variantes dans le premier état de l'œuvre ?

Nous proposons ici des réflexions qui vont peut-être changer l'optique du projet qui, au départ, a été conçu pour la

visualisation d'un état du texte pour le faire évoluer vers une édition des variantes d'un manuscrit.

Nous souhaiterions encoder les changements entre l'état du manuscrit qu'on transcrit (qui correspond à l'image du manuscrit qu'on a en regard dans l'interface de transcription dans e-Man) et les états ultérieurs du texte, jusqu'à l'édition définitive du texte. Pour cela, le manuscrit à transcrire doit correspondre au premier état du texte.

Les liens entre manuscrit transcrit, édition définitive et autres avant-textes de l'œuvre se font par un renvoi aux manuscrits numérisés qui se trouvent dans le dossier génétique de l'œuvre sur la plate-forme e-Man.

La solution de partir de l'édition définitive (comme dans une édition génétique papier) a été évoquée. Mais, tout l'intérêt d'une bibliothèque virtuelle de manuscrits est, selon nous, de présenter comme texte encodé un manuscrit et non une édition définitive imprimée. En effet, si le choix de la transcription de l'édition définitive est pertinent dans le cas d'une édition papier (moindre coût, plus grande clarté), l'outil informatique permet d'encoder et de comparer les différents états du texte et, en particulier, le chercheur peut décider de travailler à partir d'un état donné du texte. C'est une véritable expérience de découverte des différents états du texte que nous souhaitons proposer.

Notre projet de transcription est inédit puisque, s'il existe bien des exemples de transcriptions encodées de manuscrits d'écrivains – des carnets de Proust (Prototype Proust³), de ceux de Flaubert (Optima⁴, ITEM), des manuscrits de Jeremy Bentham (*Transcribe Bentham Project*) – ces manuscrits renfermaient les phénomènes à encoder au sein d'un seul état du texte. Nous avons parfois huit états du texte différents. On peut se poser la question de la faisabilité d'encoder les huit états avant-textes de la même édition définitive. Cela peut être possible en dupliquant le texte du premier état et en encodant tous les états. Nous pouvons aussi choisir d'encoder toutes les variantes à partir du premier état du texte. Pour cela, il faudrait être certain du premier état du texte. Nous devons vérifier quelle est la solution la plus rentable économiquement et la plus solide techniquement. Pour cela, il nous faut effectuer des tests. Le débat est ouvert !

II. Le plug-in Collection Tree : un outil pour constituer le dossier génétique ?

1 Fonctionnement général

Collection Tree permet de classer les collections les unes dans les autres en affectant des collections comme « collections parentes ». En fait, le lien créé n'est qu'un moyen de classement : les contenus n'appartiennent toujours qu'à une seule collection. On peut utiliser cette fonction de collection parente en créant une métacollection vide dans laquelle on regroupe plusieurs collections.

Le plug-in Collection Tree ajoute un onglet « Hiérarchie de la collection » mais ne range pas les contenus dans l'onglet « Contenus » (qui n'est pas modifié).

Dans « Hiérarchie de la collection », on voit uniquement la hiérarchie des collections : les contenus des collections n'apparaissent pas.

2 Tests

Nous avons effectué un test avec quatre avant-textes du *Fils du pauvre* de Mouloud Feraoun que nous synthétisons sous la forme du tableau suivant :

³ http://research.cch.kcl.ac.uk/proust_prototype/

⁴ <http://www.unicaen.fr/puc/images/preprint0062011.pdf>

Niveau 1	Collection	Mouloud Feraoun
Niveau 2	Sous-collection	Dossier génétique du Fils du pauvre
Niveau 3	Contenus	Avant-texte 1 – Le Fils du pauvre ⁵ Avant-texte 2 – Le Fils du pauvre Avant-texte 3 – Le Fils du pauvre Avant-texte 4 – Le Fils du pauvre
Niveau 3 bis	Sous-sous-collection	Sources exogénétiques du Fils du pauvre (cartes postales, articles de presse, etc.)

Les sources exogénétiques comprennent plusieurs documents, d'où la nécessité de créer une collection pour les y rassembler.

Cette hiérarchie est représentée sous forme de liste à puces dans le plug-in Collection Tree comme le montre la capture d'écran ci-dessous :



Les avant-textes (contenus de collection) ne sont pas visibles dans la « Hiérarchie de la collection ». Mais, si l'on clique sur la collection du dossier génétique du *Fils du pauvre*, on peut accéder aux contenus.

Au même niveau que les avant-textes, nous avons créé une sous-sous-collection « Sources exogénétiques du Fils du pauvre ». Celle-ci, contrairement aux contenus (avant-textes), apparaît bien dans la « Hiérarchie de la collection ». Les avant-textes, eux, ne sont pas visibles dans la « Hiérarchie de la collection ».

3 Avantages/inconvénients

Par conséquent, Collection Tree a le grand avantage de permettre de gérer les collections, mais il a le désavantage de ne gérer que les collections et non les contenus⁶, même quand ceux-ci sont au même niveau qu'une collection (par exemple, les sources exogénétiques).

Dans le modèle que nous exposons, le manuscrit est indiqué comme avant-texte (suivi d'un chiffre) dès son titre afin de pouvoir le distinguer des autres avant-textes de l'œuvre et classer tous les avant-textes par ordre chronologique (selon l'ordre établi par le chercheur).

Si ce modèle était adopté, cela impliquerait de revoir les titres des avant-textes dans e-Man.

⁵ Chaque avant-texte correspond à un seul document. Il existe en réalité huit avant-textes du *Fils du pauvre*. Nous pourrions donc appliquer ce modèle aux huit avant-textes.

⁶ Dans l'état actuel d'e-Man, on peut accéder rapidement à tous les contenus (documents) d'une collection auteur.

Une autre solution serait de créer une collection pour chaque manuscrit donné (chaque avant-texte), mais cela impliquerait une double manipulation sur tous les contenus, ce qui représente beaucoup de temps pour peu de bénéfice. Il est à noter que cette solution représenterait un nombre énorme de créations de collections.

Remarque : un document ne peut pas être dans deux collections en même temps à moins d'être dupliqué. (Copie d'un document dans une autre collection). Ainsi, on garde la logique de la boîte d'archive réelle : un document est dans une seule boîte d'archives à la fois.

III. Le plug-in Item Relations : un outil pour créer des relations entre les documents

Les relations permettent de rendre le dossier génétique plus dynamique. En cela, elles complètent et enrichissent l'apport du plug-in Collection Tree. En effet, les relations permettent de préciser la place du document par rapport à un autre document et de créer un lien dynamique permettant d'accéder à l'autre document. Il est à noter que les relations créées sont à double sens.

1 Solution 1

• Proposition

Avant-texte 1 ⁷ – Le Fils du pauvre	Est en relation avec	Avant-texte 2 – Le Fils du pauvre
Avant-texte 2 – Le Fils du pauvre	Est en relation avec	Avant-texte 1 – Le Fils du pauvre

Avant-texte 1 – Le Fils du pauvre	Est en relation avec	Dossier génétique du Fils du pauvre
Dossier génétique du Fils du pauvre	Est en relation avec	Avant-texte 1 – Le Fils du pauvre

• Problèmes

- Si la relation « est en relation avec » a l'avantage de fonctionner dans les deux sens, elle reste néanmoins imprécise : dans ce premier exemple, on utilise « est en relation avec » pour qualifier tous les types de relations.
- On souhaite mettre en relation des items (des documents tels que les avant-textes) avec des collections (comme le dossier génétique ou toute autre collection dont dépendrait un document appartenant déjà à une collection auteur ou à une collection « sources exogénétiques »). Or, on ne peut indiquer qu'un **identifiant d'item** (par exemple : l'id « 38 » dans « items/show/38 ») et non un **identifiant de collection** (par exemple : l'id « 8 » dans « collections/show/8 ») au niveau de l'interface de saisie du plug-in « Item Relations » :

⁷ Le texte final de tous ces avant-textes est l'édition définitive que, pour l'instant, nous ne publions pas dans e-Man.

Item Relations

Here you can relate this item to another item and delete existing relations. For descriptions of the relations, see the [Browse Vocabularies](#) page. Invalid item IDs will be ignored.

Sujet	Relation	Object	Supprimer
This Item	est en relation avec ▼	Item ID 	n/a

- On voit clairement, dans la capture d'écran ci-dessus, que seul un identifiant d'item (Item ID) est accepté comme objet de la relation. On ne peut donc créer de relations qu'entre deux items et non entre un item et une collection supplémentaire.

• Développements

- En ce qui concerne l'affichage des relations entre items, la solution 1 est simple et n'entraîne aucun développement particulier en matière d'affichage de la relation. En revanche, elle reste imprécise.
- Pour résoudre le problème d'identifiant, nous proposons de créer un item (sans fichier) avec le titre de chaque collection à mettre en relation avec d'autres items de d'autres collections. Ainsi, nous pourrions mettre en relation des items présents dans les sources exogénétiques avec une autre collection grâce à l'id item ainsi créé pour cette collection « sources exogénétiques ». Cela nous servirait dans le cas que nous exposons plus loin de la lettre d'Albert Camus présente dans une collection Albert Camus et étant en relation avec les sources exogénétiques d'une des œuvres de Mouloud Feraoun (exemple fictif).

2 Solution 2

• Proposition

Le premier tableau propose de mettre en lumière la relation d'antériorité d'un avant-texte par rapport à un autre :

Avant-texte 1 – Le Fils du pauvre	Est antérieur à	Avant-texte 2 – Le Fils du pauvre
Avant-texte 2 – Le Fils du pauvre	Est postérieur à	Avant-texte 1 – Le Fils du pauvre

Le deuxième tableau propose de mettre en évidence l'appartenance d'une lettre à une autre collection que sa collection d'appartenance initiale. Imaginons que nous ayons rangé cette lettre dans une collection Albert Camus⁸. Nous souhaitons maintenant créer une relation entre la lettre d'Albert Camus et la collection « sources exogénétiques du *Fils du pauvre* de Mouloud Feraoun » (exemple fictif) :

Lettre d'Albert Camus (appartenant à la collection Albert Camus)	Fait partie de	Collection Mouloud Feraoun
Dossier génétique du Fils du pauvre	Contient	Avant-texte 1 – Le Fils du pauvre

⁸ En réalité, celle-ci appartient déjà à la collection Mouloud Feraoun dans e-Man.

• Problèmes

- Les propositions que nous faisons posent un problème à l'affichage des relations comme illustré dans les deux tableaux suivants :

Avant-texte 1 – Le Fils du pauvre	Est antérieur à	Avant-texte 2 – Le Fils du pauvre
Avant-texte 2 – Le Fils du pauvre	Est antérieur à	Avant-texte 1 – Le Fils du pauvre

Lettre d'Albert Camus	Fait partie de	Collection Mouloud Feraoun
Collection Mouloud Feraoun	Fait partie de	Lettre d'Albert Camus

- Dans les deux exemples ci-dessus, les relations ne s'affichent pas correctement : la première ligne des deux tableaux correspond à l'item courant (« avant-texte 1 ») dont on a précisé la relation avec un autre item (« avant-texte 2 » ou « dossier génétique »). La deuxième ligne des deux tableaux (en gras) correspond à l'affichage que l'on obtient automatiquement dans l'item « avant-texte 2 » lorsque l'on renseigne la relation dans l'item « avant-texte 1 ».
- Par ailleurs, comme dans la solution 1, on souhaite mettre en relation des items (la lettre d'Albert Camus dans notre exemple) avec des collections (la collection « sources exogénétiques du *Fils du pauvre* de Mouloud Feraoun », toujours dans notre exemple). Le problème réside, comme expliqué dans la solution 1, dans le fait qu'il ne soit pas possible, à première vue, d'associer un identifiant de collection à un item.

• Développements

- La solution 2 est plus fine que la solution 1 puisqu'elle précise l'ordre des éléments dans la relation. Cependant, elle entraîne des développements importants en matière d'affichage : par défaut, les intitulés de la relation sont les mêmes dans l'item courant (Avant-texte 1 – Le Fils du pauvre) et dans l'item appelé (Avant-texte 2 – Le Fils du pauvre). Si la phrase « l'avant-texte 1 est antérieur à l'avant-texte 2 » est bien correcte, la phrase « l'avant-texte 2 est antérieur à l'avant-texte 1 » ne l'est plus.
- Par conséquent, le défi qu'il faudra relever pour l'affichage des relations est de transformer, à l'affichage, les termes servant à qualifier la relation : « est antérieur à » dans l'item courant pour qu'il devienne « est postérieur à » dans l'item appelé. On peut arriver à cela grâce à une série de conditions qui, si elles sont nombreuses, reposent toutefois sur un principe simple : **si** $A > B$, **alors** $B < A$. (par exemple « A est antérieur à B donc B est postérieur à A »).

Si ces conditions sont mises en œuvre, il nous faudra, par exemple, prévoir l'affichage suivant :

- Quand B est 1^{er} dans la relation sujet-relation-objet, il nous faut afficher « est antérieur à »
- Quand B est 3^e dans la relation sujet-relation-objet, il nous faut afficher « est postérieur à ».

Ces conditions peuvent donc s'exprimer d'une façon simple. Toutefois, pour pouvoir définir tous les types de relations dans e-Man, il nous faudra établir la liste complète des types de relations avec leurs correspondances.

Les illustrations présentes sur la couverture de ce rapport sont toutes tirées de la plate-forme e-Man, plus précisément des plate-formes « Manuscrits francophones » et « Manuscrits hispaniques » :



Portrait de Mouloud Feraoun, plate-forme « Manuscrits francophones »



Portrait de Delmira Agustini, plate-forme « Manuscrits hispaniques »



Portrait de Jean-Joseph Rabearivelo, plate-forme « Manuscrits francophones »



Portrait de Carlos Denis Molina, plate-forme « Manuscrits hispaniques »



Portrait de José Mora Guardino, plate-forme « Manuscrits hispaniques »