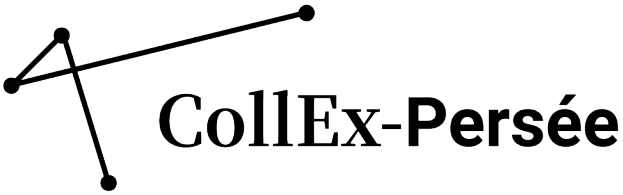
Mises à disposition de La Grande Encyclopédie — Premiers résultats

A. BRENON D. VIGIER

02 octobre 2020



1. Rappel du contexte du projet

Le projet Collex-DISCO-LGE s'inscrit dans un dispositif de recherche plus vaste, conduit au laboratoire ICAR, et qui vise à étudier avec les méthodes et les outils des sciences du langage et de la linguistique de corpus outillée, les changements majeurs intervenus dans le discours encyclopédique en France entre 1751 (parution du premier tome de l'*Encyclopédie ou Dictionnaire raisonné des arts, des sciences et des métiers* de Diderot, d'Alembert et Jaucourt) et nos jours. Dans ce dispositif figurent, à la date de rédaction de ce rapport, deux autres projets financés en cours : GÉODISCO et GÉODE. Pour conduire à bien cette étude sur les encyclopédies françaises du XVIIIe s. à nos jours, nous avons déterminé un corpus d'étude formé de sept productions éditoriales :

- **L'Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers, par une société de gens de Lettres.** 1751-1772. 17 volumes de textes et 11 volumes de planches.
  Supplément de l'Encyclopédie. 1776-1780 4 volumes + 1 volume de planches + 2 volumes de tables
- **L'Encyclopédie Méthodique par ordre de matières par une Société de gens de Lettres, de Savans et d'Artistes.** 1782-1832, plus de 200 volumes.

- **Le Grand Dictionnaire Universel du XIXe s.** (P. Larousse) 17 volumes. 1866-1876, 1878, 1888.
- **La Grande Encyclopédie. Inventaire raisonné des sciences, des lettres et des arts par une société de savants et de gens de lettres.** 31 volumes, 1885-1902.
- **Encyclopédie française**, 21 volumes,1934-1966.
- **Encyclopædia Universalis**, version 2018
- **Wikipédia** version juillet 2018

L'objectif central du projet COLLEX-PERSÉE a consisté à rendre accessible aux internautes et aux chercheurs une version XML-TEI enrichie linguistiquement d'un des ouvrages inclus dans ce corpus et qui constitue une étape majeure dans l'histoire de l'encyclopédisme en France : *La Grande Encyclopédie* [*LGE*].

Ce monument encyclopédique, le dernier d'un siècle que Pierre Larousse a qualifié dans la préface du *Grand Dictionnaire universel du XIXe siècle* de «siècle des dictionnaires» – au sens générique – paraît entre 1885 et 1902, d'abord en fascicules, puis en volumes (trente et un au total) . L'œuvre a été réalisée sous la direction d'une douzaine de personnalités dont l'ordre alphabétique fait apparaître en première position le nom de Berthelot, «sénateur, membre de l'Institut», grand chimiste. On y trouve aussi de nombreuses personnalités de tous les domaines - M. Barrès, R. de Gourmont, M. Mauss, F. Brunetière, F. Buisson, O. Reclus, M. Berthelot et ses quatre fils, E. Jobbé-Duval, C.-V. Langlois, Reinach. . . - leurs contributions constituant aujourd'hui des références solides appréciées d'un grand nombre de chercheurs. Le projet se réfère explicitement au premier «monument» du genre, l'*Encyclopédie* de Diderot et d'Alembert, se proposant de la réactualiser, et s'inscrit dans le mouvement positiviste puis scientiste du XIXe siècle. Reflet de la science en mouvement et mise en texte d'un projet encyclopédique portant sur les lettres, les arts et les sciences dites aujourd'hui exactes, *LGE* se situe au terme d'une période qui a vu naître le début de l'expansion des sciences et des techniques – accompagnée de l'émergence de la terminologie (Jacquet-Pfau C. (2007), Langages, n° 68, p. 24-38) – en même temps que leur vulgarisation. Elle donne un précieux état des lieux des différents domaines des sciences, abordés sous l'angle des connaissances et du progrès, pour un public souhaité le plus large possible sans toutefois faire de concessions au sérieux scientifique de l'entreprise. Les articles, simples notices ou véritables monographies (voir p. ex. air (52 pages), chemin de fer (28 pages), franc-maçonnerie, France (91 p.)).

Les partenariats scientifiques noués pour la réalisation de ce projet ont été :

- **Bibliothèque Nationale de France** (BNF): E. Bermès, J.-P. Moreux
- **Institut National de recherche en sciences et technologies du numérique** (INRIA), équipe Almanach : L. Romary, M. Khemakhem
- **Laboratoire LITT&ARTS**, Univ. Stendhal, Grenoble : G. Williams
- **Laboratoire PRAXILING** (U. P. Valéry, Montpellier) : S. Diwersy, H. Bohbot
- **Laboratoire LATTICE** (ENS Ulm / U. Paris 3, Paris) : I. Galleron

2. Etat de réalisation des objectifs et des livrables déclarés dans le projet déposé

Les objectifs déclarés étaient au nombre de cinq :

1.
   - A. Rendre disponible sur Gallica une version numérisée de l'intégralité de la Grande Encyclopédie en mode texte;
   - B. Rendre disponible sur l'Equipex ORTOLANG une version nue et une version étiquetée de *LGE*.
   - C. rendre disponible sur un serveur de l'ENS une version pouvant être chargée dans la plateforme textométrique TXM.
2. Participer à la mise au point d'une chaîne de traitement automatisée pour l'encodage XML-TEI des textes encyclopédiques et leur enrichissement linguistique.
3. Participer aux travaux en cours dans le cadre de l'initiative TEI vers une personnalisation du schéma TEI pour les textes encyclopédiques.
4. Réaliser une étude textométrique pilote qui identifiera - en recourant aux méthodes de la statistique textuelle et à une approche combinant *corpus driven* & *corpus based* - certaines spécificités remarquables du discours encyclopédiques dans *LGE* par contraste avec l'*Encyclopédie* Diderot & d'Alembert, l'Encyclopædia Universalis et Wikipédia.
5. Préparer le dépôt en mars 2019 auprès de l'ANR (porteur : D. Vigier) et du FNS (Porteuse : Pr. Rossari, Uni. Neuchâtel) d'un projet de recherche international pluridisciplinaire de type PRCI.

L'état de réalisation de ces objectifs à la date de rédaction de ce rapport est le suivant

Objectifs	État de réalisation	Commentaires
`1A`	Partiellement réalisé	La BNF a mis en œuvre une nouvelle numérisation de l'intégralité des volumes sous forme de fichiers PDF océrisés. Ces fichiers devraient à terme remplacer les fichiers existants sur Gallica.
`1B`	Partiellement réalisé	Une première version de *LGE* encodée en XML-TEI est actuellement disponible et téléchargeable sur le site du projet. Lorsque cette version aura été améliorée (courant 2021), nous la communiquerons à Ortolang

Objectifs	État de réalisation	Commentaires
1C	Partiellement réalisé	Une première version de *LGE* annotée est actuellement disponible sous licence libre et téléchargeable sur le site du projet.
2	Réalisé	Un prototype de cette chaîne est actuellement disponible sous licence libre et téléchargeable sur le site du projet.
3	Réalisé	Nous avons proposé et utilisé un schéma d'encodage conforme XML-TEI mais situé en dehors du module de dictionnaires TEI. Une proposition de communication sur ce thème (The specificities of encoding encyclopedias : towards a new standard ?) a été soumise et acceptée pour le colloque international ICHLL11: 11th International Conference on Historical Lexicography and Lexicology dont la date a été déplacée de juin 2020 à juin 2021 pour cause de Covid19.
4	Non réalisé	À l'heure actuelle, cette étude n'a pas été conduite. Nous sommes en effet parvenus à disposer d'une première version de *LGE* disponible pour import dans TXM qu'à la toute fin du projet.
5	Réalisé	Le dépôt d'un PRCI franco-Suisse (acronyme : DISCO) a été effectué en mars 2019 mais n'a pas été retenu par le jury pour être financé. Un nouveau dépôt a été accompli en mars 2020. La décision est attendue pour octobre 2020.

3. Détail des étapes franchies dans le projet, des difficultés rencontrées, des solutions de secours adoptées et des réalisations accomplies.

3.1. Numérisation & Océrisation(/OCRisation) par la BNF

La BnF a procédé à la numérisation des volumes de LGE au cours de l'été 2019. Les tomes ont été transmis sous forme d'archives .zip qui contiennent à la fois

une version au format PDF sous la forme d'un fichier unique et une version au format ALTO sous la forme d'un répertoire contenant un fichier XML-ALTO par page. Les fichiers PDF contiennent l'image de toutes les pages scannées, ainsi que le texte qui a été OCRisé, ce qui rend le texte sélectionnable dans les logiciels destinés à la consultation de ce format et permet également la recherche de chaînes de caractères. À chaque page du fichier PDF d'un tome correspond un fichier ALTO qui lui ne contient que du texte mais inclut une référence à chaque fois qu'une illustration est détectée dans la page. Les images de ces illustrations pourraient être obtenues à partir des photos des pages contenues dans le PDF et des coordonnées de la référence à l'illustration mais ne sont pas présentes dans les archives. Le contenu du texte est le même dans le PDF et dans les pages ALTO correspondantes, même si le séquençage diffère très légèrement : les «mots» dans le PDF (les instructions atomiques d'affichage de chaînes de caractères) et dans l'ALTO (le contenu de l'attribut `CONTENT` des balises `<STRING/>`) ne sont pas en correspondance biunivoque.

3.2. Mise au point d'une chaîne automatique d'encodage à partir des fichiers ALTO

Module 1 : nettoyage des PDF livrés par la BNF. Stratégie adoptée, logiciels créés, livrables. La qualité de ces fichiers est globalement bonne, les problèmes existant se répartissent essentiellement en 3 catégories :

- erreurs de lectures des caractères : on trouve dans les pages un certain nombre de faux-positifs, l'OCR ayant «reconnu» des caractères là où il n'y avait qu'une courbe dans une gravure ou des taches sur une page. Il y a également des caractères mal reconnus, le plus souvent à cause d'un écart angulaire trop élevé par rapport à la verticale lors de la prise de vue de la page mais aussi parfois à cause d'un cisaillement «en profondeur» des caractères vers le milieu des tomes où la page s'arrondit trop vers la reliure centrale
- erreurs de géométrie : les pages de LGE sont assez fournies en illustrations qui viennent perturber le flot du texte dans la page. On trouve des illustrations en marge du texte sur une demi-colonne, le texte se poursuivant à côté, des illustrations qui interrompent complètement la colonne ainsi que des illustrations pleine page qui barrent les deux colonnes de texte. De ce fait les colonnes de textes ne constituent pas toujours un bloc unique au sens de ALTO (il n'y a de toute façon pas de notion de bloc dans PDF), celles interrompues par des illustrations sont coupées en plusieurs blocs. Cela n'est pas une erreur de géométrie en soi mais peut causer des problèmes à la détection du sens de lecture de la page. Enfin, et encore une fois souvent à cause de distorsions visuelles causées par la photographie, les distances et les tailles relatives des blocs peuvent être mal évaluées, provoquant le groupement au sein d'un même bloc d'éléments normalement distincts, typiquement des éléments péritextes indiquant le numéro de la page en cours ou l'intervalle de vedettes qu'elle couvre se retrouvant inclus

comme première ligne de la colonne de texte démarrant immédiatement sous eux.

- erreurs de sens de lecture : le sens de lecture «naturel» attendu par un humain n'est pas toujours facile à déterminer pour l'algorithme implémenté dans l'OCR utilisé par la BnF, il arrive qu'un bloc soit positionné «avant» un autre alors qu'un lectorat humain les aurait plutôt lu dans l'autre ordre. Une remarque à ce sujet, si les pages apparaissent comme des objets en deux dimensions, rendues visuellement comme telles par les lecteurs de PDF, et si les fichiers ALTO contiennent bien des indications de position en abscisse et en ordonnée, en revanche les fichiers sont stockés linéairement sur un disque et possèdent un début et une fin. Il existe un sens de parcours linéaire du fichier, lorsque l'on veut retranscrire le texte qu'il contient (l'objet textuel étant naturellement en une seule dimension). Si les blocs en ALTO pourraient tout à fait être placés dans un ordre arbitraire et parcourus dans le bon sens grâce aux indications des propriétés `IDNEXT` des balises `<TextBlock/>`, en revanche leur contenu implique un sens de lecture à l'intérieur du bloc, et le format PDF, étant de toute façon dépourvu de blocs, contient pour chaque page une liste d'instructions d'affichage dans un certain ordre donné. Ainsi, un PDF apparaissant «dans l'ordre» dans un lecteur (parce que l'on voit les mots positionnés à l'endroit attendu et que l'on arrive à lire une phrase cohérente du fait de cette géométrie) peut en réalité être dans le désordre (parce que l'instruction d'affichage du dernier mot par exemple n'apparaît pas en dernier mais à un tout autre endroit dans la liste d'opération à effectuer pour le rendu de la page).

Ces erreurs ont des conséquences variables pour le traitement qui nous intéresse et des effets de cascades s'observent parfois : une erreur de lecture de caractère a un effet sur la géométrie détectée pour le bloc le contenant, ce qui à son tour peut avoir des répercussions sur le sens de lecture de la page.

La différence de structure du contenu des pages entre les formats PDFs et ALTO explique la différence de difficulté à corriger ces différentes erreurs entre les deux formats. À cette différence, s'ajoute le fait que le format PDF est un format binaire, qui demande l'écriture de code complexe pour être lu jusqu'à faire apparaître les instructions d'affichage dont il est question ci-dessus, et ce d'autant plus s'il s'agit également de modifier cette liste d'instructions pour créer un nouveau fichier PDF (beaucoup d'instructions ont des effets de bords ce qui les rend peu déplaçable sans risquer de changer leur sens et donc possiblement le contenu du PDF).

Module 2 : encodage XML-TEI. Comme rapporté dans notre bilan d'étape, le manque de souplesse du module ad-hoc pour les dictionnaires du standard TEI nous a incités à nous tourner vers un encodage plus générique n'utilisant pas les balises de ce module dont les contraintes de schéma XML ne nous paraissaient pas refléter les réalités des structures trouvées dans les pages de LGE (très simplement, le module est fait pour des dictionnaires, qui

ne contiennent pas de longs développements très structurés parfois en plusieurs parties et sous-parties sur un nombre conséquent de niveaux et associent à chaque vedette une définition clairement identifiable — ce qui est peu le cas dans un discours encyclopédique où du contexte historique précède parfois un terme à définir, ou bien où un article est consacré entièrement à la vie d'une personne, dont elle relate plusieurs aspects, et où l'on se demande bien ce qui, dans tout le texte, constituerait la «définition» de ce personnage).

L'encodage choisi est parfaitement compatible avec le schéma TEI standard, pour une compatibilité entière et se décrit sommairement ainsi :

- les articles sont chacuns inclus dans un élément `<div/>`
- la vedette de l'article est encodée dans un élément `<head/>` dans laquelle la mise en forme est rendue en utilisant les balises habituelles comme `<hi/>`
- les paragraphes sont représentés par des éléments `<p/>`
- les illustrations donnent lieu à des balises `<figure/>` dans lesquelles sont inclues des balises `<figDesc/>`
- les éléments péritextes sont encodés par des balises `<fw/>` dont l'attribut `type` (pouvant valoir `header` ou `footer` codent leur présence en haut ou en bas de la page)
- enfin les renvois, élements essentiels du discours encyclopédiques sont représentés par des `<xr/>` dont le mot «cible» du renvoi est placé au sein d'une balise `<ref/>`

Logiciel GROBID : fonction dans la chaîne et difficultés rencontrées

Le logiciel GROBID devait initialement assurer l'encodage des articles à partir des fichiers PDFs livrés par la BnF. L'entraînement a donné initialement de bons résultats, mais assez rapidement un seuil de qualité n'a pas pu être franchi à cause des différents types d'erreurs décrits plus haut. Les scories de textes imaginés par l'OCR se retrouvaient inclus dans le texte de sortie, perturbant parfois le modèle entraîné jusqu'à l'empêcher de reconnaître correctement les limites des articles. Les problèmes d'ordre de lecture ont également eu un impact fort, puisque les éléments péritextes fondus apparaissant entre les deux colonnes de texte se retrouvaient insérées au milieu du texte, ajoutant du contenu indésirable à cet endroit.

Cet outil ne permet pas de corriger ces problèmes et nos partenaires sur le projet qui l'ont conçu n'ont pas trouvé de solutions pour améliorer la qualité des PDFs. Nous avons donc développé à ICAR une suite d'outils pour essayer de régler cette question. Le premier de ces outils a été la librairie Hufflepdf, dont le but était de pouvoir accéder au texte des PDFs, de montrer les instructions d'affichage dans l'ordre où elles apparaissent dans les fichiers, et, à terme de pouvoir corriger leur contenu, à l'aide de règles automatiques ou de manière supervisée. En effet, la taille importante des fichiers et des questions de licence rendaient impossibles en pratique la réalisation de ces tâches de manière entièrement manuelle dans un logiciel comme Abby FineReader (non seulement le temps de traitement des 31 tomes de ~1200 pages aurait été trop important, mais de plus le volume

de mémoire consommé pour la simple ouverture des fichiers rendait le système instable et quasi-inutilisable). De plus des structures apparaissant sur certaines pages (systématiquement des cartes géographiques en pleine page) provoquaient des erreurs fatales du logiciel, et ce de manière reproductible, lors de l'étude préliminaire qui a été réalisée en corrigeant les scories du premier tome à l'aide de ce logiciel. Ces structures n'ont pas été identifiées, mais il semble que le contenu des pages du PDF du premier tome au moins (ex. la carte p. 216) aient un contenu incompatible avec la librairie utilisée par Abby FineReader pour décoder le PDF.

Avec le temps, la perspective de réussir à modifier le contenu des PDF grâce à Hufflepdf est devenue de plus en plus incertaine : difficile d'écrire des règles entièrement automatique, mais la conception d'une interface homme-machine pour corriger manuellement les problèmes aient aurait nécessité énormément de temps, sans parler du temps de traitement manuel lui-même.

Solutions de secours : Stratégie adoptée, logiciels créés, livrables.

Le premier tournant dans la stratégie du projet a été de concevoir une interface de visualisation et de correction du contenu des fichiers ALTO : chaoui. En effet, les fichiers ALTO étant des fichiers textuels dans un dialecte du XML, une abondance d'outil permet de modifier et d'interagir avec leur contenu. L'outil `chaoui` permet ainsi d'effectuer un rendu des pages en XML-ALTO similaire à ce qu'un lecteur de PDF affiche pour leur contrepartie dans ce format et de sélectionner des zones pour marquer leur contenu comme «à détruire».

L'idée initiale était de fournir de cette manière des listes de «mots» identifiées par un opérateur humain comme scories pour semi-automatiser les corrections par Hufflepdf : leur suppression aurait ensuite été faite automatiquement à partir de cette liste obtenue manuellement par un programme utilisant la librairie Hufflepdf. Cette solution s'est avérée irréalisable à cause de la très légère différence de séquençage entre PDF et ALTO (et c'est la tentative d'écrire un tel outil qui a permis de révéler cette différence entre les deux formats). Puisqu'il n'y a aucune base commune d'identifiant permettant de mettre en relation les mots du PDF et de l'ALTO, la seule solution était de les lire dans l'ordre en parallèle dans les deux fichiers à la fois et de vérifier leur correspondance mot-à-mot. La différence de découpage entre les mots rend cela très impraticable voir impossible. De plus, la difficulté à définir le déplacement d'instructions dans un PDF a rendu encore plus inenvisageable les corrections, automatiques ou non, sur l'ordre de lecture.

C'est pour cette raison qu'un deuxième logiciel a vu le jour pour essayer d'effectuer ces traitements sur les fichiers ALTO : soprano. Il permet de filtrer les mots relevés comme scories dans `chaoui` ainsi que de corriger l'ordre des blocs en appliquant son propre algorithme de positionnement basé sur un découpage géométrique de la page.

Enfin, c'est ce même outil auquel a été ajouté les fonctionnalités de découpage de la séquence des pages en articles et d'encodage en XML-TEI des articles trouvés.

État actuel de la chaîne

Les fonctionnalités principales de `soprano` sont complètement implémentées : la suppression des scories identifiées manuellement, la correction de la géométrie de la page ainsi que de son ordre de lecture et le découpage du flot du texte en articles distincts. Cet élément constitue le premier maillon d'une chaîne de traitement transformant les fichiers ALTO du corpus en fichiers XML-TEI encodant les articles extraits de La Grande Encyclopédie.

Il est complété par un ensemble de scripts qui permettent respectivement

- de simplifier la structure du corpus distribué par la BnF (`LGEprepareVolume.sh`)
- de faire appel à `soprano` en détectant automatiquement les filtres disponibles pour le tome ciblé et en passant les options correspondante à `soprano` (`LGEencode.sh`)
- de générer les métadonnées permettant de transformer le répertoire de sortie contenant les articles en un corpus au format attendu par la chaîne de traitement PRESTO (`LGEexportCorpus.sh`)

Cette autre chaîne constitue le deuxième segment des traitements effectués sur le corpus en vue de son intégration à des outils de textométrie. Idéalement, un dernier script aurait permis de lier les deux segments afin d'obtenir un traitement unique de bout en bout, partant des archives pour obtenir les articles annotés morpho-syntaxiquement et encodés en XML-TEI, mais l'absence d'un installeur pour la chaîne `PRESTO` complique son appel depuis un répertoire arbitraire en ne désolidarisant pas entièrement le logiciel de son dépôt.

Un tel script rassemblant l'ensemble des deux segments de traitement devrait ou bien effectuer des calculs lourds sur les chemins relatifs entre le répertoire contenant l'état initial de `LGE` et celui hébergeant le code de `PRESTO`, procéder à des copies supplémentaire et changer de répertoire en cours d'exécution, ou bien encore forcer l'ensemble de l'exécution depuis le répertoire de `PRESTO`. Aucune de ces trois solutions n'est franchement plus satisfaisante du point de vue de l'ingénierie logicielle que de simplement décrire ces deux segments et de laisser l'opérateur ou l'opératrice les appeler librement en fonction des étapes qu'il ou elle souhaite atteindre et de l'état du corpus déjà disponible.

3.3. Production d'une première version structurée et encodée en XML-TEI de LGE

Les tomes ont donc été d'abord extraits des archives communiquées par la BnF puis leur structure simplifiée à l'aide de `LGEprepareVolume.sh`. Cette simplification consiste en un parcours et une édition de lien et est très rapides, de l'ordre de quelques secondes. L'extraction des archives elle-même a été l'étape la plus lourde, nécessitant quelques minutes de calcul pour chacune des archives d'environ 1.3Go par volume. Ces opérations ont été répétées pour les 31 tomes.

Ces fichiers extraits ont ensuite été traités manuellement à l'aide de `chaoui` pour produire les filtres de scories des tomes ainsi que repérer les caractères

non-textuels apparaissant de manière erronnée sur les pages et constituant un «bruit» dans le signal. Cette phase, empirique par nature, s'est développée dans le temps et a été l'occasion d'une boucle entre le développements des codes sources de `chaoui` et `soprano` et l'activité d'élimination des scories. Il s'agit bien entendu d'un coût temporel unique qui représente la plus grande part de celui de ce projet, et n'est pas à repayer à chaque fois que les données structurées et encodées en XML-TEI de `La Grande Encyclopédie` sont produites.

Le calcul effectif, une fois les filtres disponibles prend environ 3mn par tome. Le programme `soprano` occupe environ 500Mo maximum en mémoire au cours de son traitement, ce qui a permis de manière pratique de paralléliser le calcul, en lançant des instances de la commande `LGEencode.sh` sur plusieurs tomes du corpus simultanément. Le temps total théorique de traitement de 93mn (31 * 3mn) a ainsi été ramené à une trentaine de minutes sur un système avec un processeur multi-cœurs.

3.4. Production d'une première version annotée de LGE importable dans TXM

La chaîne de traitement `PRESTO` manipule pour l'instant essentiellement du texte brut. C'est à dire qu'elle accepte en entrée des fichiers dans différents encodages, dont la XML-TEI, mais qu'elle les débarrasse de leurs balises avant de les passer dans le logiciel `TreeTagger` puis de réassembler un jeu minimal de balises pour mettre la sortie dans un format XML-TEI «plat» (une balise `<w/>` autour de chaque mot mais plus de structure profonde au-dessus) compatible avec TXM.

Elle contient pour cela différents scripts de pré-traitement pour s'adapter aux différents formats et aux différentes conventions d'encodage des différentes sources du projet pour lequelle elle a été créée. Il aurait fallu potentiellement créer un tel script pour nos choix spécifiques d'encodage afin de tenir à l'écart le contenu des balises `<pb/>` ou `<fw/>` par exemple, dans lesquelles se trouve du contenu intéressant méritant d'être encodé mais qui ne doit pas apparaître dans le texte des articles proprement dits dont il ne fait pas partie. Au lieu de cela, puisque `soprano` permet de sélectionner les types de balises à inclure dans la sortie et de produire ou bien du XML ou bien du texte brut, un deuxième calcul du corpus a été effectué, comme le précédent et dans des conditions de performances proches des résultats donnés ci-dessus afin de disposer directement d'une version en texte du corpus.

Le répertoire obtenu a ensuite été mis au format attendu par la chaîne `PRESTO` en extrayant des métadonnées minimales. Pour l'instant, celles-ci contiennent seulement la liste des fichiers et des données constantes sur tout le corpus comme le titre de l'œuvre et sa source. Il reste un travail conséquent à fournir pour extraire les métadonnées propres à chaque article — auteur, domaine de connaissance — à partir d'une grammaire plus fine pour analyser les articles qui permettrait d'identifier ces informations lors de leur encodage.

Enfin, ce corpus a été traité de la manière habituelle avec la chaîne `PRESTO` en

passant un profil par défaut.

4. Perspectives de travail au-delà du projet

Si les besoins de ce projet ont amené au développement d'un certain nombre d'outils, en revanche la portée de ces outils dépasse celle du projet. Le programme `chaoui` par exemple permet de manière tout à fait générale de visualiser et d'éditer des fichiers au format ALTO, pas spécifiquement ceux de *La Grande Encyclopédie*.

Le manque de temps n'a pas toujours permis de séparer proprement l'algorithme général applicable à toute donnée en entrée des valeurs tout à fait contingentes de ce corpus particulier. Les outils produits pourraient être rendus plus directement utiles et profitables à la communauté de la recherche en passant du temps pour bien séparer ces deux aspects afin d'obtenir les outils les plus généraux possibles.

Cette généralisation ouvrirait la voie à un réemploi des outils du projet pour les appliquer à d'autres œuvres proches comme par exemple *Le Grand Dictionnaire Universel* mentionné plus haut afin de permettre entre autre l'étude comparative de ces deux œuvres.