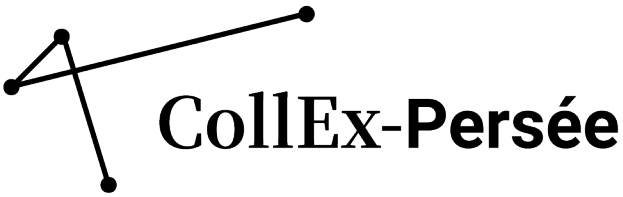
Fin du projet CollEx-Persée DISCO-LGE

A. BRENON D. VIGIER

21 juin 2021



1. Bilan

À l'heure de clore le projet DISCO-LGE, voici quelques éléments pour documenter l'état de réalisation auquel il est parvenu et compléter ainsi le premier apperçu donné dans le rapport précédent publié à l'automne 2020.

Nous tenons avant toute chose à remercier une fois de plus les partenaires scientifiques qui ont contribué à ce projet:

- **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

Nos remerciement chaleureux vont aussi tout particulièrement au CollEx-Persée pour son soutien financier et opérationnel tout à long du projet.

Les principes FAIR (*Findability*, *Accessibility*, *Interoperability*, *Reuse of digital*

assets) de science ouverte ont occupé une place centrale dans le projet DISCO-LGE et orienté plusieurs des choix que nous avons opérés dans le projet et se retrouvent nécessairement au long de ce bilan.

1.1 Objectifs

Pour rappel, **les objectifs** déclarés pour ce projet étaient au nombre de cinq :

1.
   - A. Rendre disponible sur Gallica une version numérisée de l'intégralité de la Grande Encyclopédie (LGE) 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 (EDdA), 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.

1.2 Exemplaire

Pour mémoire, l'exemplaire sélectionné par la BNF pour le présent projet est situé à la Bibliothèque de l'Arsenal à Paris.

Nous reproduisons ci-dessous en la complétant la notice n° **FRBNF41651490** du catalogue général de la BNF qui lui est consacré[1].

[1]Pour une synthèse concernant l'édition de l'ouvrage, on se reportera en particulier à C. Jacquet-Pfau (2015), « Élaboration et destinée d'une encyclopédie à la fin du XIXe siècle : les trente-et-un volumes de La Grande Encyclopédie. Inventaire raisonné des sciences, des lettres et des arts par une Société de savants et de gens de lettres (1885-1902) », revue *ELA*, n°177,

Type(s) de contenu et mode(s) de consultation : Texte noté : sans médiation

Titre(s) : La grande encyclopédie [Texte imprimé] : inventaire raisonné des sciences, des lettres et des arts par une société de savants et de gens de lettres / sous la direction de MM. Berthelot, Hartwig Derenbourg, D.-Camille Defus...[et al.][2]

Publication : Paris (61, rue de Rennes) : H. Lamirault, [puis] Société anonyme de la Grande Encyclopédie, 1885-1902

Impression : (37-Tours : Arrault et Cie)

Description matérielle : 31 vol. : fig. ; 31 cm

Note(s) : Cartes en couleurs, certaines dépliantes

Autre(s) auteur(s) :

- Berthelot, Marcellin (1827-1907). Directeur de publication
- Derenbourg, Hartwig (1844-1908). Directeur de publication
- Dreyfus, Camille (1851-1905). Directeur de publication

Secrétaire général :

André Berthelot (1862 1938)

L'ouvrage propose en outre 153 cartes géographiques hors-texte en six couleurs

1.3 État de réalisation

Voici l'état de réalisation final de ces objectifs du projet listés supra

Objectifs	État de réalisation	Commentaires
`1A`	Réalisé	La BNF a mis en ligne sur Gallica les 31 volumes numérisés.

p. 85-100.

[2] Entre onze et douze auteurs selon les volumes sont en réalité désignés sur cette notice par « et al. », et ainsi regroupés dans l'intitulé « Sous la direction de ».

Objectifs	État de réalisation	Commentaires
1B	Partiellement réalisé	Des versions de *LGE*, une en texte brut et une autre encodée en XML-TEI sont disponibles en téléchargement sur la plateforme Nakala mise à disposition par Huma-Num. Nous prévoyons encore d'autres améliorations à l'issue desquelles nous communiquerons une version stable à Ortolang.
1C	Réalisé	Nous avons finalement opté pour distribuer la version annotée avec le reste du corpus sur Nakala.
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 communication sur ce thème (The specificities of encoding encyclopedias : towards a new standard ?) a été présentée au colloque international ICHLL11 : 11th International Conference on Historical Lexicography and Lexicology qui s'est tenu en juin 2021.
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.

Objectifs	État de réalisation	Commentaires
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 mais n'a pas non plus été retenu. Ce projet trouve finalement une suite dans le projet GÉODE.

1.4 Productions

Directes Le premier de ces objectifs décrivait explicitement la production de documents numériques, de trois types suivant les différentes étapes considérées et correspondant respectivement aux objectifs 1A, 1B et 1C. L'ensemble de ces livrables est téléchargeable librement sur le site Nakala cité précédement enrichi de métadonnées conformément au principe de «trouvabilité» (*Findability*) cité plus haut.

Une version numérique des trente-et-un tomes de l'exemplaire fourni par la BnF

La réalisation de l'objectif `1A` s'est traduite par la renumérisation intégrale des 31 volumes de LGE opérée par notre partenaire la BnF. Cette version numérique se présente sous forme de fichiers dans deux formats différents.

- une version PDF: un fichier par tome d'environ 1.3 Go, contenant le texte brut positionné sur la page par-dessus la photo de la page correspondante de l'exemplaire papier
- une version XML ALTO: un fichier par page, dans un sous-répertoire séparé pour chaque tome. Chacun de ces dossiers pèse environ 300 Mo.

Les articles segmentés

Les versions précédentes comprennent tout le texte brut, regroupé en blocs géométriques de texte mais sans découpage en articles. Le premier apport du projet DISCO-LGE a été de fournir un découpage du texte en articles. Ces articles sont disponibles sous deux formats.

- un format textuel brut: il ne contient aucune balise et ne retient que la

mise en forme déjà présente dans les fichiers ALTO, c'est à dire le jeu sur la casse ainsi que les retours à la ligne

- un format en XML-TEI: il ajoute des métadonnées au fichier (référençant le projet, le tome dont est extrait le fichier ainsi que le titre de l'article), dont le contenu textuel est encodé en suivant les conventions décrites ci-dessous section 2.2.

Les articles annotés en morpho-syntaxe

L'outil d'annotation automatique PRESTO basé sur TreeTagger que nous utilisons permet de produire des fichiers dans lesquels à chaque mot est associé une étiquette morpho-syntaxique. Notre chaîne de traitement produit des fichiers dans différents formats:

- HTML
- Unitex
- ScienQuest
- TXM

Nous ne distribuons sur Nakala que la version pour TXM qui nous paraît la plus utile immédiatement dans la perspective d'une étude textométrique mais il est tout à fait possible de reproduire localement les autres versions manquantes au besoin à l'aide de la chaîne de traitement `PRESTO`, appliqués au fichiers au format brut issus de l'objectif 1B.

Indirectes À ces produits directs s'ajoutent plusieurs autres contingentes des choix d'implémentation auxquels nous avons dû procéder au cours du projet, pour atteindre les objectifs initiaux ou pour pallier aux imprévus qui ont surgi en confrontant nos projets aux contraintes techniques et qui n'avaient pas pu être prévus ni annoncés aux origines de DISCO-LGE. Ces productions contingentes de notre travail incluent une suite de scripts écrits pour fluidifier les interactions entre les différentes étapes de notre chaîne de traitement ainsi que des fichiers correctifs du contenu des pages au format ALTO, le tout contenu dans le dépôt logiciel `ProcessingLGE`.

- les scripts fournissent le liant nécessaire pour ajuster entre elles les différentes étapes et les différents formats du projet. Le premier dans l'ordre chronologique de la chaîne, `LGEprepareVolume.sh` permet d'obtenir nommage normalisé des fichiers sources communiqués par la BnF en vue de faciliter leur traitement automatique par `soprano`. Le deuxième, `LGEencode.sh`, simplifie l'appel à `soprano` en lui passant par défaut un

ensemble d'options raisonnables pour les besoins déclarés du projet, en référençant les fichiers correctifs. Le dernier, `LGEexportCorpus.sh`, crée les méta-données pour faire de la sortie de `soprano` un corpus au format attendu par la chaîne de traitement `PRESTO`.

- Les fichiers de correctifs se présentent sous la forme d'un fichier texte contenant des caractères explicables seulement par les bruits de l'OCR et qui peuvent être éliminés d'emblée des sources ainsi que de fichiers CSV, contenant une seule colonne répertoriant les identifiants des éléments `String` des fichiers ALTO que nous suggérons de considérer comme du bruit de l'OCR ou des éléments non-exploitables pour les besoins de notre étude (formules algébriques, partitions... que nous avons appelé «scories» au cours du projet). Ces listes sont distribuées avec les scripts ci-dessus et ont été créé manuellement à l'aide de l'outil `chaoui` en vue d'être ensuite appliqués au cours du processus automatique de nettoyage par `soprano` (c.f. infra).

Enfin, il faut compter au nombre de ces productions indirectes tous les supports de documentation qui ont été créés pour illustrer notre processus de travail et le rendre le plus *Accessible* possible.

- des wikis ont été créés pour décrire en détail le processus de travail attendu pour chacun des deux outils développés pour le projet, en sus d'une documentation classique explicitant le fonctionnement des différentes options
- le dépôt `ProcessingLGE` mentionné ci-dessus est une mise en pratique de «documentation par le code»: en fournissant ces scripts optionnels pour l'utilisation de la chaîne de traitement, nous suggérons une façon de combiner les différents éléments pour parvenir au but que nous nous sommes fixé et proposons une base solide comme point de départ pour réutiliser nos outils avec d'autres paramètres ou dans un contexte légèrement différent
- le dépôt Vue d'ensemble adopte un point de vue encore un peu plus abstrait de la technique puisqu'il ne contient pas de code mais seulement de la documentation: un guide d'utilisation de la chaîne de traitement.

2. Contributions

2.1 Scientifiques

Les résultats que nous venons d'évoquer ont fait l'objet d'une communication scientifique au cours du congrès ICHLL11 le 17 juin 2021. (développer le contenu,

faire le lien avec les objectifs initiaux du projet et montrer en quoi c'est un résultat qui a de la valeur au-delà de DISCO-LGE lui-même)

2.2 Schéma d'encodage pour les encyclopédies

Recommandation

Conformément au schéma d'encodage présenté dans le précédent rapport nous recommandons l'emploi du module *core* de la XML-TEI pour encoder les encyclopédies. Nous privilégions cette solution par rapport au fait d'éditer un schéma XML personnalisé se démarquant des règles de la TEI standard ou d'employer les éléments standards du module *dictionaries* dans un sens différent légèrement de leur sémantique initiale au nom du principe d'interopérabilité de la *FAIR science*.

Voici les conventions que nous proposons d'appliquer pour les encoder les articles d'une encyclopédies et que souhaitons à terme appliquer aux fichiers sortis par la chaîne de traitement:

- chaque article est inclus dans un élément `<div/>` dont l'attribut `xml:id` est unique dans le corpus
- 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 `head` ou `foot` 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/>`

Remarques

`<hi/>`

En pratique, il semble plus difficile qu'initialement prévu d'accéder à des informations typographiques sur les éléments textuels dans le format ALTO et donc de générer des balises `<hi/>` correctement. Il demeure possible de les réinejcter artificiellement et inconditionnellement dans les vedettes puisqu'on sait en parcourant les pages de l'œuvre que les vedettes sont toutes en majuscules et

en gras, mais cette information ne viendra pas d'un traitement des sources en ALTO, et, surtout, ne permettra pas d'obtenir des `<hi/>` aux autres endroits mis en reliefs par l'utilisation d'une typographie spéciale.

`<lb/>`

L'évidence n'avait pas été formulée précédement mais les débuts de ligne sont encodés comme il se doit avec l'élément `<lb/>` de la TEI. L'attribut `break` (avec la valeur `yes` ou `no`) permet de signaler si la nouvelle ligne sépare deux mots distincts ou s'il faut au contraire recoller les deux parties, ce qui est utile pour ensuite extraire le texte pour le segmenter en vue d'une analyse lexicale. La valeur d'un tel attribut est fortement corrélée à la présence d'un trait d'union mais hélas pas strictement équivalente, à cause du cas où un mot composé, comportant un trait d'union, . Le problème est donc plus complexe qu'il n'y paraît et serait sans doute mieux résolu dans des couches d'analyses ultérieures, notamment lexicale.

`<pb/>`

Notre encodage comprend aussi des éléments `<pb/>` («page begining») avec l'attribut `n` indiquant l'indice (démarrant à 0) de la nouvelle page (chaque page correspond à un fichier ALTO dans la source).

Structure

Le mécanisme de segmentation des articles est encore assez limité et ne permet pas de regarder leur structure interne, c'est pourquoi nous ne sommes pas encore en mesure de produire l'encodage des paragraphes (`<p/>`) et des renvois (`<xr><ref/></xr>`) dans les fichiers de sortie.

Encodage atteint

Voici donc un résumé de la convention actuellement suivie dans les fichiers produits par la chaîne de traitement:

- les articles sont chacuns inclus dans un élément `<div/>` dont l'attribut `xml:id` est unique dans le corpus
- la vedette de l'article est encodée dans un élément `<head/>`
- 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 `head` ou `foot` codent leur présence en haut ou en bas de la page)

- les débuts de lignes sont encodés par des balises `<lb/>`
- les débuts de pages sont encodés par des balises `<pb/>` dont l'attribut `n` contient l'indice de la page

La figure 1 illustre cet encodage par une sortie de la chaîne de traitement.

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
    <teiHeader>
        <fileDesc>
            <titleStmt>
                <title>La Grande Encyclopédie, T9, article CAN</title>
            </titleStmt>
            <publicationStmt>
                <publisher>CollEx-Persée DISCO-LGE</publisher>
                <date when="2021" />
            </publicationStmt>
            <sourceDesc>
                <bibl>
                    <title>La Grande Encyclopédie</title>
                    <publisher>H. Lamirault</publisher>
                    <date from="1885" to="1902" />
                </bibl>
            </sourceDesc>
        </fileDesc>
    </teiHeader>
    <text>
        <body>
            <div xml:id="can-0">
                <lb /><head>CAN</head> É Fl Cl ER. Nom vulgaire du Cassia fistula L., de
                <lb />la famille des Légumineuses-Cæsalpiniées (V. Casse).
            </div>
        </body>
    </text>
</TEI>
```

Figure 1: Article Canéficier du volume 9 de LGE

2.3 Solution de segmentation et d'encodage d'articles

Ce qui avait été initialement présenté comme une solution «de secours» dans les précédentes versions de ce rapport est désormais une suite logiciel stable que nous distribuons parmi les résultats du projet. Pour rappel, cette suite consiste en deux logiciels principaux ainsi que les scripts décrits plus-haut pour faciliter l'emploi de la chaîne de traitement et favoriser ainsi sa réutilisation.

- `soprano` est un logiciel en ligne de commande conçu pour nettoyer les fichiers ALTO encodant les pages d'une encyclopédie et pouvoir les segmenter en articles. Les articles identifiés peuvent ensuite être sauvegardés sous forme de fichiers, sous deux formats différents: texte brut ou XML-TEI suivant le schéma d'encodage décrit ci-dessus. Parmi les opérations de nettoyage, `soprano` permet d'éliminer des caractères identifiés comme étant

des bruits de l'OCR et de corriger le positionnement et le découpage des blocs de texte réalisé par l'OLR. Le comportement des différents réglages permettant cela est documenté sur son wiki.

- `chaoui` est une interface sous forme d'une application web qui permet d'afficher un rendu de fichiers au format XML ALTO. Elle permet de charger un ou plusieurs fichiers, d'afficher les valeurs de confiance de l'OCR (attribut `WC` des éléments `String` du format ALTO pour «word confidence») ainsi que de mettre en évidence les blocs de texte (éléments `TextBlock` de ce même format). Une documentation illustrée de captures d'écran décrit plus en détail ses différentes fonctionnalités.

3. Suites

Au-delà, le projet en ayant atteint son objectif de fournir une première version encodée en XML-TEI de La Grande Encyclopédie trouve son prolongement naturel dans le projet GÉODE auquel il fournit une des quatre encyclopédies du corpus d'étude.