Ann. Soc. Géol. Nord, T. 31 (2e série), p. 101-108, 2024

CoESciTer. UN PROJET DE TRANSCRIPTION ET D'ÉDITION NUMÉRIQUE D'UN CORPUS DE SCIENCES DE LA TERRE

CoESciTer. A project of transcription and digital edition of an Earth Sciences Corpus

Milène Mallevays

Service Commun de Documentation de l'Université de Lille, F-59000 Lille, France
coesciter@univ-lille.fr

Geoffrey Haraux

Service Commun de Documentation de l'Université de Lille, F-59000 Lille, France
geoffrey.haraux@univ-lille.fr

Jessie Cuvelier

CNRS, Univ. Lille, UMR 8198 Evo-Eco-Paléo, F-59000 Lille, France
jessie.cuvelier@univ-lille.fr

Résumé

Le projet CoESciTer, programmé de 2022 à 2024, a eu pour objectif la transcription et l'édition d'un corpus de manuscrits scientifiques en Sciences de la Terre, incluant des notes de cours de professeurs et d'élèves, avec une priorité sur les documents rédigés par Abel Briquet. La technologie HTR (*Handwritten Text Recognition*), via les logiciels Transkribus et eScriptorium, a été employée pour automatiser ce processus. Des modèles personnalisés ont été développés pour améliorer la précision des transcriptions, particulièrement pour les manuscrits dont l'écriture était complexe. Le texte issu des transcriptions a été encodé au format XML-TEI, facilitant ainsi leur utilisation et accessibilité. En complément de la transcription, le texte a été enrichi pour corriger certaines fautes d'orthographe et signaler les abréviations utilisées dans le texte original des notes de cours. La création d'un index permettant de naviguer dans les transcriptions par termes et notions scientifiques est un autre objectif du projet. Un script de reconnaissance a balisé automatiquement les termes scientifiques dans les fichiers XML-TEI, en utilisant le vocabulaire des « Sciences de la Terre » du « Thésaurus de la science ouverte » développé par l'INIST-CNRS. L'utilisation de la plateforme Heurist a permis ainsi de configurer une base de données et d'intégrer des filtres sophistiqués pour la recherche. Cet index comprend plus de 5 600 occurrences de termes scientifiques, dont 530 termes uniques, offrant une ressource précieuse pour les chercheurs en histoire des sciences. Le projet CoESciTer a de cette manière combinée des technologies avancées et des méthodes d'édition rigoureuses pour produire une ressource accessible et utile, contribuant significativement à l'étude de l'enseignement des sciences géologiques à la Faculté des Sciences de Lille.

Abstract

*The CoESciTer project, scheduled from 2022 to 2024, aimed to transcribe and edit a corpus of scientific manuscripts in Earth Sciences, including lecture notes from teachers and students, with priority on the documents written by Abel Briquet. HTR (*Handwritten Text Recognition*) technology, via Transkribus and eScriptorium software, was used to automate this process. Custom templates were developed to improve the accuracy of transcriptions, particularly for manuscripts with complex handwriting. The text from the transcriptions was encoded in XML-TEI format, thus facilitating their use and accessibility. In addition to the transcription, the text has been enriched to correct certain spelling errors and point out the abbreviations used in the original text of the course notes. The creation of an index allowing you to navigate the transcriptions by scientific terms and notions is another objective of the project. A recognition script automatically tagged scientific terms in the XML-TEI files, using the "Earth Sciences" vocabulary from the "Open Science Thesaurus" developed by INIST-CNRS. The use of the Heurist platform made it possible to configure a database and integrate sophisticated filters for search. This index includes more than 5,600 occurrences of scientific terms, including 530 unique terms, providing a valuable resource for researchers in the history of science. The CoESciTer project has in this way combined advanced technologies and rigorous editing methods to produce an accessible and useful resource, contributing significantly to the study of geological science teaching at the Faculty of Sciences of Lille.*

INTRODUCTION

Le Service Commun de Documentation (SCD) de l'Université de Lille porte le projet CoESciTer avec l'UMR CNRS 8198 Evolution, Ecologie, Paléontologie (CNRS/Université de Lille) et le Centre François Viète d'épistémologie et d'histoire des sciences et des techniques (Université de Nantes – Université de Caen). Ce projet a pour but de constituer un corpus en ligne de sources manuscrites sur l'histoire de l'enseignement des sciences de la Terre au XIXe et XXe siècles.

En effet, cette période voit se développer des chaires de minéralogie et de géologie dans les universités françaises en région, suite à l'établissement des premières chaires parisiennes et des cartes géologiques départementales françaises (Savaton, ce volume). Ces disciplines ont rapidement suscité l'intérêt des notables et des administrations provinciales car

la géologie au sens large était cruciale pour le développement économique, industriel et agricole. Par conséquent, les facultés des sciences établies en province à partir des années 1850 ont souvent inclus une chaire de géologie. Par exemple, une chaire de minéralogie et de géologie a été créée en 1857, à l'Université de Lille, entraînant un développement significatif de l'enseignement de cette discipline lors de la nomination de Jules Gosselet en 1864 (Cuvelier, ce volume). Ainsi les professeurs et les étudiants ont pris des notes de cours pour respectivement les donner ou les recevoir, manuscrits conservés jusqu'à nos jours. D'où l'initiative de constituer un **Co**rpus sur l'**E**nseignement des **Sci**ences de la **Ter**re (**CoESciTer**).

Le projet CoESciTer est financé par le GIS CollEx-Persée pour une durée de deux ans, de septembre 2022 à septembre 2024 et s'appuie sur un partenariat fort avec des sociétés savantes, concernées par la valorisation, la diffusion et l'histoire des sciences de la Terre : la Société géologique du Nord (SGN), la Société géologique de France (SGF) et le Comité français d'histoire de la géologie (Cofrhigéo). Les bibliothèques du Muséum National d'Histoire Naturelle et de l'Institut de France sont également partenaires du projet en numérisant des documents issus de leurs collections pour enrichir le corpus, principalement constitué de sources issues du SCD de l'Université de Lille.

Le corpus est composé de notes de cours en sciences de la Terre (minéralogie, géologie, paléontologie, géographie physique) rédigées par un professeur en amont pour préparer son cours, ou de notes prises par des étudiants qui ont suivi les cours en question. Les carnets ou liasses intégrés au corpus sont conservés au Lilliad learning centre Innovation de l'Université de Lille (entité du SCD de l'Université de Lille), au sein du fonds donné à l'Université par la Société géologique du Nord (Delrue *et al.*, 2021), à la bibliothèque de paléontologie de l'UMR CNRS 8198 Evolution, Ecologie, Paléontologie, à la bibliothèque de l'Institut de France et à la bibliothèque du Museum National d'Histoire Naturelle.

Dans un premier temps, le projet a démarré par une enquête visant à identifier des documents similaires dans les bibliothèques de l'enseignement supérieur et de la recherche et les bibliothèques territoriales, dans les dépôts d'archives publiques et dans les musées d'histoire naturelle en région. À l'exception de deux institutions (l'Université de Paris-Saclay et l'Ecole Normale Supérieure de Paris), l'enquête n'a pas permis de signaler d'autres fonds remarquables de même typologie.

Dans un second temps, les documents du corpus ont été numérisés, et les images mises à disposition dans la bibliothèque numérique patrimoniale LillOnum. Seule une partie du corpus a été transcrite et encodée en XML-TEI : la mise au point d'une méthodologie fonctionnelle a occupé l'essentiel du temps dévolu au projet. L'objectif était de réaliser des transcriptions au moyen d'outils de transcription automatisée de manuscrits (HTR), d'encoder les textes au format XML-TEI et de mettre en ligne ces textes structurés accompagnés d'un index des termes scientifiques pour faciliter leur exploitation. L'appropriation de ces technologies, pas encore employées majoritairement au sein des bibliothèques de l'Université de Lille, était un élément central du projet. C'est pourquoi une ingénieure d'étude a été recrutée spécifiquement pour le projet afin de coordonner le travail de transcription et l'appropriation des outils de transcription HTR ainsi que d'apporter ses compétences en matière d'encodage XML-TEI et d'identification automatisée des termes scientifiques pour la création d'un index. Elle s'est chargée de la mise au point de la méthodologie et des scripts employés ainsi que de la mise en ligne des textes encodés et de la base de données qui permet de faire fonctionner l'index.

Cet article a pour objectif de présenter les enjeux de la réalisation de la création de la première édition numérique du corpus CoESciTer. Nous discuterons du processus de transcription des manuscrits du corpus, de l'encodage de ces derniers en XML-TEI et enfin de la création de l'index permettant de filtrer le corpus par termes techniques et notions associées présents dans les transcriptions.

MATÉRIEL D'ÉTUDE

Le corpus est constitué de notes de cours, soit de notes préparatoires écrites par des enseignants, soit des notes d'étudiants prises pendant le cours.

La partie la plus conséquente du corpus est constituée de 21 carnets de notes prises par Abel Briquet (1874-1952) lorsqu'il assistait aux cours de géologie, minéralogie, géographie et paléontologie de la Faculté des Sciences de l'Université de Lille. Les carnets en question ont été rédigés entre 1901 et 1904. Les cours suivis ont été donnés par Edouard Ardaillon (1867-1926), Charles Barrois (1851-1939), Henri Douxami (1871-1913) et Jules Gosselet (1832-1916). Tous ces documents ont été signalés dans Calames (https://calames.abes.fr/pub/#details?id=FileId-4756) et numérisés dans la bibliothèque numérique patrimoniale LillOnum (https://lillonum.univ-lille.fr/s/lillonum/item-set/1613332). Conservés par les bibliothèques de l'Université de Lille au Lilliad learning centre Innovation, ces carnets sont entrés dans les collections lors du dépôt de la bibliothèque de la Société géologique du Nord à l'Université, transformé ensuite en don en 2009 (Delrue *et al.*, 2021). Ce fonds d'intérêt majeur pour l'histoire des sciences de la Terre bénéficie du label national « Collection d'excellence » attribué par le GIS CollEx-Persée aux fonds documentaires d'intérêt national pour la recherche. Un carnet correspondant aux notes prises par Abel Briquet lors d'un cours de géologie de Jules Gosselet (Annexe 1) a été entièrement transcrit et encodé au format XML-TEI. Le résultat a été mis en ligne sur le site CoESciTer (https://coesciter-lillonum.univ-lille.fr/).

Deux liasses de notes manuscrites datées de 1871-1872 et 1880-1881 de Charles Barrois et de Jules Gosselet, conser-

vées à la bibliothèque de Paléontologie de l'UMR Evolution, Ecologie, Paléontologie (Locatelli, 2017) ont été également numérisées et mises en ligne dans la bibliothèque numérique patrimoniale LillOnum. Un cours de minéralogie de Jules Gosselet a été intégralement transcrit et encodé (Annexe 1).

Deux établissements partenaires du projet ont également contribué au corpus. La bibliothèque de l'Institut de France a numérisé un cours de Constant Prévost (1787-1856) pris en note par Henri Doniol (1818-1906). Il a été mis en ligne sur la bibliothèque numérique de l'Institut (https://bibnum.institutdefrance.fr/ark:/61562/bi25383) et visible aussi au sein du corpus sur LillOnum au moyen du protocole IIIF. La bibliothèque du Museum National d'Histoire Naturelle a numérisé deux manuscrits issus du fonds d'archives Marcellin Boule (1861-1942) : un cours de géologie qu'il a donné comme professeur à Clermont-Ferrand en 1890 et un cours de paléontologie de Louis Lartet (1840-1899) qu'il a suivi comme étudiant à Toulouse entre 1880 et 1882. Ces deux manuscrits seront prochainement disponibles via la bibliothèque numérique LillOnum. Ils sont signalés dans Calames avec le reste du fonds Marcellin Boule.

La transcription et l'étude ci-dessous portent sur les manuscrits de géologie au sens large enseignée à l'Université de Lille (Annexe 1).

TRANSCRIPTION DU CORPUS

Dans le contexte du projet, la transcription d'une partie du corpus est une étape prioritaire. Il est nécessaire de connaître le contenu du corpus avant de commencer cette étape car il influence la sélection des documents manuscrits à prioriser pour la transcription. Le corpus CoESciTer contient deux types de documents, des notes de cours de professeurs et des notes de cours d'élèves.

Très rapidement, il a été constaté que les manuscrits rédigés par les étudiants étaient majoritaires dans le corpus par rapport à ceux des professeurs, et qu'Abel Briquet, alors élève à l'université de Lille, en était à 80 % l'auteur (Fig. 1).

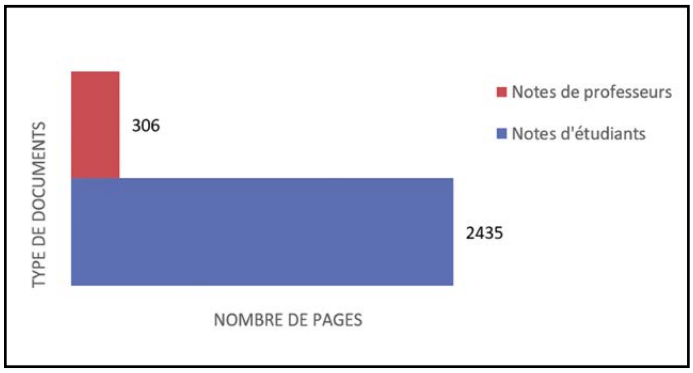

Fig. 1. Répartition du nombre de pages du corpus par type de document (notes de professeurs / notes d'étudiants).
Fig. 1. Distribution of the number of pages in the corpus by type of document (teacher notes / student notes).

De plus, la qualité de l'écriture des scripteurs influence le temps et la difficulté de réalisation des transcriptions. Une comparaison entre les manuscrits de notes de professeurs et de notes d'étudiants a révélé des différences très visibles. Les notes de professeurs sont propres et structurées, l'écriture est lisible et stable. Les documents sont rédigés avec des phrases complètes, il y a peu de ratures, d'abréviations et de fautes d'orthographe (Fig. 2A). À l'inverse, les notes d'étudiants sont difficilement lisibles à cause d'une écriture rapide nécessaire à la prise de notes, une absence de phrase complète et une grande présence d'abréviations, de ratures et de fautes d'orthographe (Fig. 2B). Tous ces éléments ralentissent le déchiffrage, la lecture, la compréhension du texte et donc de sa transcription.

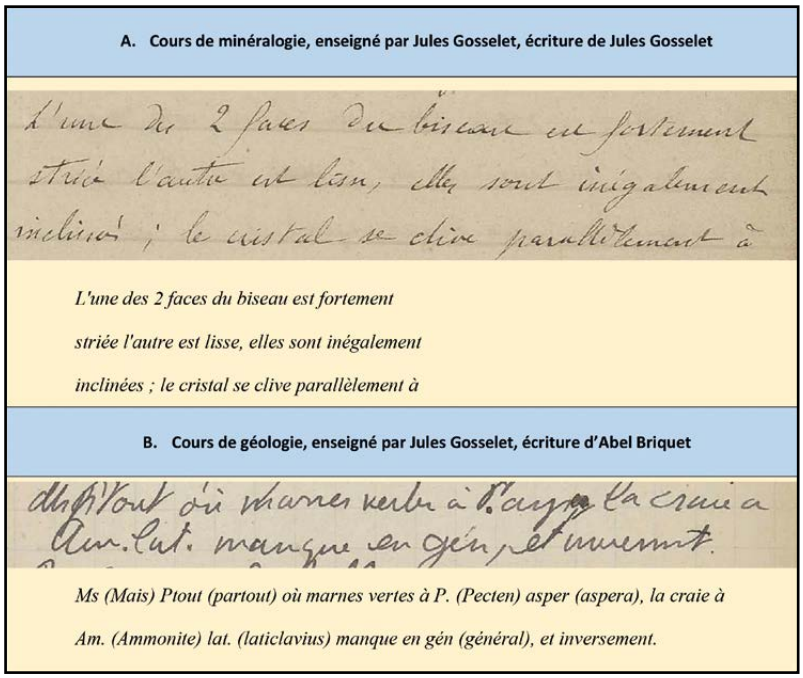

Fig. 2. Extrait de cours de minéralogie et de géologie (texte original et transcription de l'extrait), écrit par un enseignant (Jules Gosselet) en A et un étudiant (Abel Briquet) en B.
Fig. 2. Extract from mineralogy and geology courses (original text and transcription of the extract), written by a teacher (Jules Gosselet) in A and a student (Abel Briquet) in B.

En présence de documents aussi spécifiques (abréviations, ratures, lexiques scientifiques, erreurs, etc.), comment les transcrire de façon efficace pour rendre disponibles ces informations à la communauté scientifique ?

L'un des objectifs principaux du projet CoESciTer consistait à exploiter l'HTR (*Handwritten Text Recognition*), une technologie spécialisée dans la transcription automatique des écritures manuscrites. Dans le cadre de ce projet, deux logiciels de HTR ont été utilisés : eScriptorium (webographie : eScriptorium), accessible grâce au consortium CREMMA, et Transkribus (webographie : Transkribus).

Transkribus, développé dans les années 2010 et depuis 2019 par la READ COOP, ainsi qu'eScriptorium, créé par SCRIPTA PSL en 2019, offrent des solutions avancées pour la transcription automatique de documents. eScriptorium utilise une interface graphique permettant d'exploiter le logiciel de transcription automatique KRAKEN, développé en 2015 par Benjamin Kiessling. Les deux logiciels sont accessibles gratuitement pour la transcription manuelle. Cependant, Transkribus devient payant pour l'utilisation des modèles de transcription automatique, bien que des crédits gratuits soient offerts à l'ouverture d'un compte, permettant une utilisation initiale sans frais. En revanche, eScriptorium, en tant que logiciel open source, reste entièrement gratuit, y compris pour l'utilisation de ses modèles de transcription automatique.

L'utilisation de l'HTR vise à faciliter la transcription rapide et en grande quantité de documents manuscrits. Toutefois, il est essentiel de préciser que les technologies HTR ne sont pas encore capables de transcrire des manuscrits avec un bon niveau de qualité, sans relecture humaine de contrôle. Ces techniques ont également tendance à avoir plus de difficultés à donner un résultat correct quand l'écriture n'est pas parfaitement claire et stable.

Le fonctionnement général des logiciels de HTR nécessite plusieurs étapes. Tout d'abord, une image est fournie au logiciel, qui identifie les zones de texte présentes. Ensuite, les caractères de ces zones sont extraits et comparés à des exemples connus en mémoire. Le logiciel détermine alors la correspondance la plus proche et produit les caractères numériques correspondants, créant ainsi le résultat de la transcription automatique. Par conséquent, le lancement d'une transcription automatique peut produire des résultats différents selon le modèle de transcription automatique sélectionné. Cette intelligence artificielle s'appuie sur les correspondances images/caractères numériques qu'elle a préalablement apprises et stockées en mémoire.

Les logiciels HTR proposent des modèles publics, disponibles gratuitement, pour tous les utilisateurs. Ils ont soit été créés par les administrateurs des logiciels, soit ils ont été rendus publiques par des utilisateurs des logiciels. Les modèles proposés par les administrateurs sont entraînés avec énormément de données et d'écritures différentes. Nos premières tentatives de transcription automatique ont été menées sur Transkribus avec un de leurs modèles publics. Ce modèle indiquait un pourcentage d'erreur prévisionnel de 7 à 8 %, mais utilisé sur les manuscrits rédigés par Abel Briquet, le pourcentage d'erreur réel tournait autour du 20 %. Ce résultat, bien qu'attendu, compte tenu du peu de stabilité dans l'écriture d'Abel Briquet était décevant. Il était donc impossible de poursuivre le projet avec un modèle aussi peu performant sur l'écriture du scripteur qui représente plus de 80 % de notre corpus. La relecture et la correction des transcriptions automatiques auraient pris plus de temps qu'une transcription manuelle.

Néanmoins, il est possible de créer des modèles personnalisés avec les logiciels HTR, ce qui permet d'adapter les correspondances images / caractères numériques à des documents spécifiques. Ainsi, des modèles basés sur l'écriture des scripteurs, notamment Abel Briquet, ont été développés pour surpasser les modèles publics en performance.

Ce processus de création implique l'entraînement d'un modèle de transcription à partir de transcriptions manuelles fournies au logiciel. Initialement, les images sont manipulées pour délimiter les zones et lignes de texte à transcrire, en assurant l'ordre logique des paragraphes et des lignes sur le document manuscrit. Une fois les préparatifs terminés, des transcriptions manuelles préalablement réalisées, sont intégrées dans le logiciel pour commencer l'entraînement du modèle.

En raison de la nature longue et fastidieuse du processus de transcription manuelle, des efforts ont été faits pour assurer la production de transcriptions de qualité pour l'entraînement des modèles. Cela inclut au moins une transcription et une relecture par deux personnes différentes, avec au moins une personne experte ou qualifiée en sciences de la Terre. Dans certains cas, une deuxième relecture est réalisée pour minimiser les erreurs et assurer la clarté des documents transcrits.

La documentation de Transkribus recommande entre 5 000 et 15 000 mots transcrits pour lancer un entraînement (Webographie : Transkribus, préparation des données) et précise également qu'un modèle peut commencer à être considéré de qualité lorsque son taux d'erreur est inférieur à 10 % (Webographie : Transkribus, CER). Pour les deux extrémités de la fourchette de mots deux modèles ont été créés et obtenus, respectivement, 43 % et 24 % d'erreur avec l'écriture d'Abel Briquet (Fig. 3).

Ces taux sont encore beaucoup trop élevés pour considérer ces modèles de transcription comme efficaces. À la fin du projet, en entraînant le modèle avec la totalité des transcriptions manuelles d'Abel Briquet, soit plus de 27 000 mots transcrits, un taux d'erreur de 10 % a été atteint. En parallèle de l'entraînement du modèle de l'écriture d'Abel Briquet, un modèle pour l'écriture de Jules Gosselet, présente dans des notes de professeur, a été créé. En quelques mois, là où plus d'un an a été nécessaire pour entraîner un modèle correct pour Abel Briquet, et avec seulement 4 000 mots transcrits, le modèle de transcription spécifique pour l'écriture de Jules

Gosselet a été entraîné avec juste 6 % d'erreur (Fig. 3). La différence dans la qualité de l'écriture des deux scripteurs est significative et a beaucoup joué sur la réalisation de leur modèle respectif.

Scripteur	Nombre de mots	Taux d'erreur
Abel Briquet	5 200 mots	43 %
Abel Briquet	14 500 mots	24 %
Abel Briquet	27 000 mots	10 %
Jules Gosselet	4 000 mots	6 %

Fig. 3. Taux d'erreur pour trois modèles de l'écriture d'Abel Briquet, avec un nombre croissant de mots enregistrés dans le logiciel, et comparaison des performances avec des taux d'erreur acceptables (inférieurs à 10 %), développés pour deux scripteurs différents (Abel Briquet et Jules Gosselet). Modèles développés sur le logiciel Transkribus.
Fig. 3. Error rates for three models of Abel Briquet's writing, with an increasing number of words recorded in the software, and comparison of performance with acceptable error rates (less than 10%), developed for two different writers (Abel Briquet and Jules Gosselet). Models developed on Transkribus software.

ENCODAGE EN XML-TEI

Une fois les transcriptions réalisées, il est nécessaire de les encoder dans un format de fichier spécifique avant de pouvoir générer les pages web et les mettre en ligne. Le format XML-TEI a été utilisé à cet effet. Le XML (*Extensible Markup Language*) est un langage d'encodage de données et la TEI (*Text Encoding Initiative*) est une norme internationale de XML, principalement utilisée dans les Humanités Numériques. L'encodage de nos données de cette manière facilite leur utilisation par quiconque souhaite consulter les transcriptions, les réutiliser dans d'autres projets ou continuer les transcriptions dans le cadre du projet CoEScTier.

En plus de la transcription des manuscrits du corpus, une édition du contenu est également réalisée. Cela inclut l'encodage des abréviations (graphies courte et longue), des fautes d'orthographe sur les termes scientifiques et sur les noms propres (graphies erronée et corrigée), des mots illisibles ou indéchiffrables, mais également d'autres éléments plus visuels comme les soulignements, les mots barrés mais lisibles, les caractères en indice et exposant. Cependant, étant donné la quantité des erreurs de casse et de ponctuations ainsi que toutes les fautes d'orthographe présentes dans les notes d'étudiants, il n'a pas été possible de les encoder dans les fichiers XML-TEI, faute de temps. Cette édition du contenu est intégrée aux fichiers XML-TEI des transcriptions, permettant ainsi de recenser des informations sur la forme et le fond des transcriptions originales et éditées et de les stocker dans un même fichier. Les fichiers XML-TEI des transcriptions actuellement en ligne sont disponibles sous licence CC BY 4.0 sur le site CoESciTer dans l'onglet « Mode d'emploi ».

L'édition numérique du projet propose deux versions des transcriptions du corpus, une imitative qui va représenter les manuscrits tels qu'ils ont été rédigés par nos scripteurs et une éditée qui est légèrement modifiée pour proposer une lecture alternative des manuscrits permettant une lecture plus fluide et compréhensible pour les utilisateurs du site (Fig. 4). Ces derniers peuvent choisir quelle version de la

Fig. 4. Comparaison des éditions numériques des transcriptions sur le site web CoESciTer : en haut, version éditée et en bas, version originale.
Fig. 4. Comparison of digital editions of the transcriptions on the CoESciTer website: top, edited version and bottom, original version.

transcription ils souhaitent afficher grâce à des boutons en haut de chaque page web contenant une transcription. D'autres fonctionnalités sont présentes sur le site dans l'onglet « Mode d'emploi », comme l'explication de certaines mises en forme dans les pages web des transcriptions, l'accès aux images numérisées des manuscrits ou encore le téléchargement des fichiers XML-TEI.

Une présentation du projet, ainsi que le lien vers le site CoESciTer (Webographie : CoESciTer) est disponible via LillOnum, la bibliothèque numérique de l'université de Lille, dans l'onglet « CoESciTer » (Webographie : LillOnum). À la fin du projet, en juin 2024, 241 pages transcrites sont accessibles aux historiens des sciences pour étudier l'enseignement professé en géologie et en minéralogie à la Faculté des Sciences de Lille.

INDEX DU CORPUS

Un objectif supplémentaire du projet était de créer un index permettant de filtrer les transcriptions par termes et par notions scientifiques associées pour faciliter la recherche. À cette fin, l'encodage des termes scientifiques dans les fichiers XML-TEI s'est avéré nécessaire. Un script de reconnaissance a été développé pour baliser automatiquement ces mots avec leurs notions associées. Pour fonctionner, le script a besoin d'une liste de mots à rechercher dans les fichiers XML-TEI. L'enjeu principal de la création de l'index était la création de cette liste. Deux approches ont été envisagées : créer cette liste à partir du texte lui-même ou utiliser une liste préexistante.

La première option impliquait un travail manuel substantiel et la supervision d'un expert en géologie, assurant ainsi l'exhaustivité des termes techniques dans l'index. Cependant, les ressources nécessaires pour cette approche étaient insuffisantes. La solution adoptée a été d'utiliser le vocabulaire des « Sciences de la Terre » du « Thésaurus de la science ouverte » développé par l'INIST-CNRS, fournissant une liste d'environ 10 000 termes avec leurs notions scientifiques associées (Webographie : Thésaurus Loterre). Même si certains termes largement utilisés aujourd'hui ou historiques ne figurent pas dans la liste du thésaurus et par conséquent sont absents de l'index, l'utilisation de cette liste a permis de produire une première version de l'index, qui peut être modifiée et améliorée lors d'une reprise du projet CoESciTer.

L'intégration de l'encodage des termes de l'index dans les fichiers XML-TEI a permis la création d'une page web dédiée à l'indexation. L'objectif principal était de mettre en place des filtres pour trier les données par terme, notion, nom de la leçon, professeur, scripteur et matière. Malgré les efforts pour développer un index interne, les contraintes de temps pour l'autoformation et l'incertitude quant à la réalisation d'un résultat satisfaisant ont conduit à opter pour une solution externe : l'utilisation de Heurist, une plateforme reconnue pour ses capacités de création, de stockage et de publication de bases de données (Webographie : Heurist). Cette plateforme a permis de configurer la base de données pour l'index et de mettre en œuvre les filtres requis sur la page web. La facilité d'utilisation de Heurist a facilité la réalisation d'un index répondant précisément à nos besoins, sans rencontrer de difficultés significatives.

Une explication de fonctionnement de l'index et de ses filtres est disponible dans l'onglet « Mode d'emploi » du site CoESciTer. L'index actuellement disponible en ligne répertorie plus de 5 600 occurrences de termes scientifiques différents. Ce total inclut 530 termes scientifiques uniques, indiquant que certains termes se trouvent plusieurs fois dans l'index en raison de leur fréquence d'utilisation dans les documents transcrits.

CONCLUSION

Le projet CoESciTer a nécessité un effort méthodique visant à transcrire et à éditer un corpus varié de manuscrits scientifiques, incluant à la fois les notes de professeurs et d'étudiants. L'application de technologies avancées telles que l'HTR a permis d'aborder la transcription de documents manuscrits malgré les défis présentés par la diversité et la complexité de l'écriture. Les résultats obtenus avec les modèles HTR, bien que prometteurs, ont nécessité une adaptation spécifique aux caractéristiques individuelles des scripteurs, soulignant l'importance de la personnalisation des modèles pour optimiser la précision des transcriptions de notre corpus.

L'encodage des données en XML-TEI a facilité la gestion et la publication des transcriptions, permettant une accessibilité de ces manuscrits aux chercheurs et au grand public intéressés par l'histoire des sciences, en leur fournissant une version éditée plus lisible, sans mettre de côté la transcription originale. L'indexation des termes scientifiques a enrichi cette accessibilité en offrant des outils de recherche sophistiqués, malgré les contraintes initiales rencontrées lors de la création de la liste de termes.

La mise à disposition des transcriptions et de l'index sur la plateforme CoESciTer représente une contribution significative à la communauté scientifique, offrant une ressource précieuse pour étudier le développement de l'enseignement des sciences de la Terre à la Faculté des Sciences de Lille. L'ensemble du projet souligne l'importance de l'intégration de technologies innovantes associées à des ressources humaines compétentes dans le domaine des études historiques et des Humanités Numériques, ouvrant ainsi de nouvelles perspectives pour la recherche et l'enseignement dans ce domaine spécialisé.

Remerciements : Les auteurs souhaitent remercier particulièrement François Guillot, Marie Antoine-Hennion, Emmanuelle Fournel et Léo Collinet pour avoir passé de longues heures à transcrire, relire et aider à l'encodage, le consortium CREMMA pour l'ouverture des accès à escriptorium, Olivier Vermaut pour la mise en ligne des images numérisés du corpus, Célia Guerinaud pour l'installation des pages web CoESciTer sur LillOnum, Laetitia Bossart pour la signalisation de ce fonds d'archives sur Calames et Jérémie Berthe et son équipe pour les numérisations du corpus, ainsi que les agents du Service Commun de Documentation qui ont pris part de près ou de loin à ce projet. Les auteurs sont reconnaissants au GIS CollEx-Persée qui a financé le projet et à tous les partenaires de nous avoir permis de mener ce projet à son terme.

BIBLIOGRAPHIE

Cuvelier J. (2024). L'enseignement de la géologie *sensu lato* à l'Université de Lille de 1858 à 1939. *Annales de la Société Géologique du Nord*, **2e série, 31** : 115-130.

Delrue L., Cuvelier J., Laden S. & Crepin B. (2021). Déchiffrer la Société géologique du Nord en escaladant les rayons de sa bibliothèque : histoire et analyse du fonds documentaire. *Annales de la Société Géologique du Nord*, **2e série, 28** : 57-92.

Locatelli E. (2014). La bibliothèque recherche des sciences de la Terre de l'Université de Lille au fil du temps : historique du patrimoine, un fonds au service de la communauté scientifique. *In* : Blieck A. & De Baere J.-P. (dir.), La Société géologique du Nord et l'histoire des sciences de la Terre dans le nord de la France. *Mémoires de la Société géologique du Nord*, **17 :** 151-173.

Savaton P. (2024). La contribution des chaires de « géologie » des facultés des sciences au développement et à la diffusion des sciences géologiques au XIXe siècle. *Annales de la Société Géologique du Nord*, **2e série, 31** : 153-160.

Webographie

ESCRIPTORIUM : https://escriptorium.inria.fr/

TRANSKRIBUS : https://www.transkribus.org/

TRANSKRIBUS, préparation des données : https://help.transkribus.org/data-preparation

TRANSKRIBUS, CER : https://help.transkribus.org/character-error-rate-and-learning-curve

THÉSAURUS Loterre : https://skosmos.loterre.fr/26L/fr/

HEURIST : https://heuristnetwork.org/

LILLONUM : https://lillonum.univ-lille.fr/s/lillonum/page/accueil

COESCITER : https://coesciter-lillonum.univ-lille.fr/

ANNEXE

Liste des carnets et des cours, transcrits et encodés au format XML-TEI, accessibles sur LillOnum (https://coesciter-lillonum.univ-lille.fr/)

Nom de la leçon	XML-TEI	Professeur	Scripteur	Date	Intervalle de page	Nombre de pages
Le Jurassique	Géologie, cours de M. Gosselet I.	Jules Gosselet	Abel Briquet	1901-1902	p. 4-6	2
Le Lias	Géologie, cours de M. Gosselet I.	Jules Gosselet	Abel Briquet	1901-1902	p. 7-11	5
Jurassique moyen Dogger	Géologie, cours de M. Gosselet I.	Jules Gosselet	Abel Briquet	1901-1902	p. 12-23	6
Jurassique supérieur Malm	Géologie, cours de M. Gosselet I.	Jules Gosselet	Abel Briquet	1901-1902	p. 24-73	25
Crétacique inférieur	Géologie, cours de M. Gosselet I.	Jules Gosselet	Abel Briquet	1901-1902	p. 74-83	5
Cénomanien	Géologie, cours de M. Gosselet I.	Jules Gosselet	Abel Briquet	1901-1902	p. 82-89	4
Turonien	Géologie, cours de M. Gosselet I.	Jules Gosselet	Abel Briquet	1901-1902	p. 88-95	4
Sénonien	Géologie, cours de M. Gosselet I.	Jules Gosselet	Abel Briquet	1901-1902	p. 94-99	3
Crétacique du bassin du Mons	Géologie, cours de M. Gosselet I.	Jules Gosselet	Abel Briquet	1901-1902	p. 98-109	6
Crétacé en dehors du Nord	Géologie, cours de M. Gosselet I.	Jules Gosselet	Abel Briquet	1901-1902	p. 111-133	12
Fossiles vertébrés du secondaire	Géologie, cours de M. Gosselet I.	Jules Gosselet	Abel Briquet	1901-1902	p. 134-145	6
Formations continentales prétertiaires	Géologie, cours de M. Gosselet I.	Jules Gosselet	Abel Briquet	1901-1902	p. 146-155	5
Tertiaire	Géologie, cours de M. Gosselet I.	Jules Gosselet	Abel Briquet	1901-1902	p. 156-201	23
Oligocène	Géologie, cours de M. Gosselet I.	Jules Gosselet	Abel Briquet	1901-1902	p. 203-213	6
Néogène	Géologie, cours de M. Gosselet I.	Jules Gosselet	Abel Briquet	1901-1902	p. 214-221	4
Tableau d'ensemble du tertiaire	Géologie, cours de M. Gosselet I.	Jules Gosselet	Abel Briquet	1901-1902	p. 222-225	2
Tertiaire d'Europe	Géologie, cours de M. Gosselet I.	Jules Gosselet	Abel Briquet	1901-1902	p. 227-235	5
Minéralogie	Minéralogie	Jules Gosselet	Jules Gosselet	1871-1872	p. 1-40	40
Orpiment	Minéralogie	Jules Gosselet	Jules Gosselet	1871-1872	p. 1-11	11
Minéralogie	Minéralogie	Jules Gosselet	Jules Gosselet	1871-1872	p. 1-40	40
Cours de Minéralogie	Minéralogie	Jules Gosselet	Jules Gosselet	1871-1872	p. 1-27	27