XML, TEI et graphes pour le traitement, la visualisation et l'analyse des noms de poissons et créatures aquatiques dans le corpus Ichtya

Marie Bisson*†[1], Brigitte Gauvin*‡[2], Pierre-Yves Buard*§[1], and Barbara Jacob¶[3]

[1]Maison de la recherche en sciences humaines – Université de Caen Normandie, Centre National de la Recherche Scientifique : USR3486 – France

[2]Centre Michel de Boüard (CRAHAM) – CNRS : UMR6273, Université de Caen – Université de Caen Normandie, 14000, France

[3]Maison de la recherche en sciences humaines – Université de Caen Normandie, Normandie Université, Centre National de la Recherche Scientifique : USR3486 – France

Résumé

Présentation du programme de recherche
Le corpus Ichtya (B. Gauvin et T. Buquet resp.) est intégré au corpus Cahier depuis sa création. Il est articulé en 3 volets[1] :

- des éditions de textes multisupports avec traduction[2] ;

- la bibliothèque numérique Ichtya[3] ;

- un thesaurus numérique des noms de poissons et créatures aquatiques[4].

Le groupe Ichtya rassemble des chercheurs et ingénieurs du Centre Michel de Bouärd[5] et des ingénieurs du pôle Document numérique[6]. Les réalisations de ce groupe de travail sont le résultat d'un travail étroit entre les deux équipes.

Problématique de la communication

Notre communication se propose d'insister plus particulièrement sur la méthodologie appliquée au traitement de l'ichtyonymie dans le corpus. Il s'agit de montrer les liens étroits entre les technologies choisies, la matière scientifique et les problématiques de recherche qui ont conduit le groupe Ichtya à construire un outil de recherche particulier : le thesaurus des noms de poissons et créatures aquatiques.

Présentation de la bibliothèque

*Intervenant
†Auteur correspondant: marie.bisson@unicaen.fr
‡Auteur correspondant: brigitte.gauvin@unicaen.fr
§Auteur correspondant: pierre-yves.buard@unicaen.fr
¶Auteur correspondant: barbara.jacob@unicaen.fr

La bibliothèque Ichtya est une bibliothèque numérique qui rassemble des textes latins consacrés à l'ichtyologie qui furent publiés dans l'Antiquité, au Moyen Âge et à la Renaissance. Elle s'inspire de la *Bibliotheca Ichtyologica* de Peter Artedi (1705-1735) et a pour vocation de mettre en ligne et à disposition des lecteurs un corpus latin consacré au savoir ichtyologique.

Le corpus, entièrement encodé en XML-TEI, est composé actuellement de 21 textes (8 en ligne publique) et 5 traductions (2 en ligne publique). Ces textes, attribués ou anonymes, sont de nature hétérogène (textes sacrés, encyclopédies, dialogues, poèmes) et de longueur très variable (fragments, chapitres, livres entiers).

La bibliothèque a été conçue d'emblée comme un corpus que les outils numériques permettraient de valoriser et de dynamiser, d'où le choix du langage XML pour permettre une interopérabilité entre les différents volets. Deux objectifs présidaient à l'entreprise : le souhait d'étudier le lexique des noms de poissons latins à travers les siècles ; et la volonté de cerner avec précision comment se fait la transmission des savoirs de l'Antiquité jusqu'à la Renaissance dans un domaine précis. Les outils permettant ce type de valorisation ont donc été choisis ou élaborés dans cette perspective. Le second objectif a été rempli par l'indexation systématique des sources : pour chaque œuvre, les sources ont été identifiées segment par segment et le lecteur, au fil de sa lecture, voit inscrite en marge l'origine du passage concerné et peut afficher la totalité du texte source pour se livrer à une comparaison. Le premier objectif a été atteint grâce à la constitution d'un thesaurus indépendant des citations[7]. En adoptant les usages du numérique, on arrive ainsi à valoriser le corpus et offrir de nouveaux outils pour la recherche.

Une méthodologie particulière pour l'indexation

Le corpus de la bibliothèque Ichtya ayant été traité de manière à permettre l'étude des noms de poissons latins à travers les âges, un index était donc indispensable pour permettre aux visiteurs de faire une recherche par nom de poisson ou de créature aquatique et d'accéder immédiatement à tous les passages du corpus Ichtya dans lequel ce nom apparaît.

Pour la première édition de texte (*De piscibus* de l'*Hortus sanitatis*[8]), la méthodologie d'indexation retenue était traditionnelle : un marqueur normalisé précédait le terme rencontré[9].

Pour la bibliothèque Ichtya, au lieu d'être pensée à plat, texte après texte, terme après terme, la méthodologie d'indexation a été conçue sous forme de thesaurus indépendant[10].

Le thesaurus

Le thesaurus est construit en XML-TEI. Il est composé d'autant de fichiers XML (notices) que de formes latines (au nominatif) ou vernaculaires rencontrées dans le corpus de la bibliothèque Ichtya. Chaque notice présente toujours la référence précise à la source dans laquelle le terme apparaît. Cette indication de source s'accompagne, autant que possible, d'une ou plusieurs identifications et, pour les appellations latines et grecques, de la référence scientifique qui valide ces identifications. Ces identifications peuvent être accompagnées d'une note de commentaire. Les notices peuvent aussi présenter deux sortes de renvois sous forme de liens : d'une part à la forme principale en cas de paronymie, de variante orthographique ou de forme vernaculaire, indication qui figure en tête de la fiche, à la place de l'identification ; de l'autre aux autres termes désignant le même animal sous un autre nom. L'indexation par le biais du format XML permet de faire des liens directement d'une forme à l'autre. Ce thesaurus fournit un outil de première utilité pour l'étude des synonymies et polyonymies entre les noms de poissons dans les traités ichtyologiques.

Chaque forme de nom de poisson ou créature aquatique rencontré dans le corpus de la bibliothèque Ichtya fait donc l'objet d'une notice XML-TEI et chaque occurrence est liée à une notice du thesaurus[11].

Résultats textuels obtenus

Cette méthodologie nous a permis :

- de générer l'index de la bibliothèque Ichtya : l'index peut être mis à jour au fur et à mesure de son alimentation publique et permettre ainsi la consultation transversale de la bibliothèque Ichtya[12] ;

- de donner accès dynamiquement à une notice de thesaurus lors de la lecture d'un texte (chaque terme indexé est signalé par la couleur et la notice est accessible au clic) ;

- de proposer un site Thesaurus indépendant permettant d'accéder à l'ensemble des notices créées pour le corpus Ichtya dans son intégralité, via le sommaire, le moteur de recherche ou encore les liens internes entre chaque notice.

L'utilisation du langage XML à la fois pour les éditions, la bibliothèque Ichtya et le thesaurus de noms de poissons et créatures aquatiques, au moyen des recommandations de la *Text Encoding Initiative* (TEI), permet une interopérabilité totale. Chaque élément indexé dans une édition de texte, dans le corpus de la bibliothèque Ichtya ou dans l'index des zoonymies se trouve ainsi immédiatement traité et actif dans les données des deux autres supports : un nom de poisson indexé dans un texte de la bibliothèque se trouve immédiatement relié à une fiche dans le thesaurus.

Exploration graphique

La constitution indépendante du thesaurus nous a également permis de proposer une lecture du thesaurus sous forme de graphes dynamiques. À partir de l'encodage en arbre XML-TEI, des graphes ont pu être générés pour chaque notice, permettant de mettre en évidence les liens établis par les chercheurs : identification ; variantes graphiques ; notices en relation.

La chaîne de traitement développée s'appuie donc sur les informations scientifiques contenues dans les notices du thesaurus et en particulier sur les liens construits pendant les phases d'annotation. Le langage RDF permet d'exprimer explicitement ces liens sous la forme de relation entre les deux poissons (entre la notice de départ du lien et sa destination). Une fois le graphe RDF produit, un générateur de diagramme est appliqué pour produire une forme graphique intégrée au site du thesaurus. Cette intégration permet d'examiner un réseau à différentes échelles en proposant un système de zoom, de mettre en lumière un poisson et ceux qui lui sont directement liés ou encore de consulter une fiche depuis un diagramme.

La constitution de réseaux de notices permet aussi de modifier l'unité, ou le grain, de consultation. En effet, le lecteur peut examiner l'ensemble des poissons entretenant des relations de quelque nature que ce soit (traduction, variante, identification, etc.). Les diagrammes permettent en définitive de consulter l'ensemble d'un réseau de poissons liés les uns aux autres. Cette modification de l'unité de consultation à travers la mise à disposition de visualisations de réseaux de notices permet de faciliter l'étude de la circulation des savoirs : le fait de regrouper dans un même graphe " synoptique " toutes les notices connectées les unes aux autres peut permettre aux chercheurs d'identifier de nouveaux liens qui pourraient leur avoir échappé pendant les phases d'annotation.

D'un point de vue informatique et documentaire, il s'agit aussi d'explorer les limites du modèle de données arborescent (XML) à travers une expérimentation directe sur le terrain scientifique. On pourra alors évaluer les solutions de passage d'une organisation des données en arbre à un modèle plus expressif en tirant parti de la finesse de l'annotation mise en place par les spécialistes pendant les phases d'étude. En effet, le modèle de graphe RDF permet d'expliciter, en plus des parties de textes habituellement annotées en XML, les relations entretenues par ces différents textes. Là où le XML permet d'exprimer la seule relation *contient/est contenu par*, le RDF n'offre aucune limitation dans la caractérisation des relations entre les éléments qui compose un graphe. Il s'agit pour nous d'exploiter non seulement les

parties de textes annotées pendant le travail de recherche, mais aussi, et peut-être surtout, les relations tracées entre ces parties de textes en particulier à des fins de visualisation.

Par ailleurs, en articulant les visualisations produites à partir des graphes RDF et les interfaces textuelles exploitant des textes encodés en XML-TEI, nous présentons une solution tirant parti des deux modèles de données sans surcoût d'encodage manuel.

Enfin, cet enrichissement sémantique des données permettra, à terme, de proposer l'interrogation du corpus sur les relations existant entre les différents poissons constitutifs du corpus. L'exploitation des types de relations pourra, par exemple, permettre la création d'un sous-corpus composé uniquement des noms de poissons, latins et vernaculaires, se référant à un même animal, à travers les siècles et les langues.

Nous proposons ici une méthode d'exploitation pour l'enrichissement des interfaces de consultation. Il s'agit, à partir des instances XML, de reprendre l'analyse réalisée par les chercheurs pour en extraire le réseau sous-jacent et le rendre visualisable et manipulable en explicitant les relations implicites existant entre les notices de poissons. Notre méthode permet d'exploiter ce réseau en tant que tel du point de vue informatique et documentaire, ce qui est, en définitive, rarement le cas dans ce type de projet ; or, les intérêts pour la recherche et l'étude des textes anciens sont nombreux.

Notre communication se propose de présenter plus en détail :

- le corpus et les objectifs de recherche du groupe Ichtya en termes d'ichtyonymie ;

- les notices du thesaurus en TEI et l'établissement du lien fait entre elles et les textes de la bibliothèque Ichtya : on montrera l'apport de cette méthodologie pour encoder et analyser le corpus ;

- la génération des graphes dans l'objectif de proposer une visualisation non textuelle et ce que cela apporte d'un point de vue documentaire ;

- les principes d'articulation des modèles de données en fonction des besoins et sans annotation manuelle supplémentaire.

Bibliographie

Bisson Marie, Gauvin Brigitte et Jacob Barbara, *Environnement d'édition scientifique en XML-TEI utilisé dans le cadre du programme Ichtya pour encoder les compilations médiévales*, Documentation du Pôle Document numérique, 2020. Consultable en ligne : http://www.unicaen.fr/recherche/mrsh/

Bisson Marie, Gauvin Brigitte, Jacquemard Catherine, *Rédiger une notice pour le thesaurus des créatures aquatiques du corpus Ichtya*, Documentation du Pôle document numérique, 2018. Consultable en ligne : http://www.unicaen.fr/recherche/mrsh/sites/default/files/public/document_numeriqu

Bisson Marie , Goloubkoff Anne, " Les notices d'autorité en XML-TEI : un outil pour l'accroissement collaboratif de connaissances et l'indexation d'éditions de sources ", *Tabularia*, 2020, Les sources des mondes normands à l'heure du numérique. Consultable en ligne : https://doi.org/10.4000/tabularia.4176.

Buard, Pierre-Yves " Le réseau de la baleine ou la visualisation de l'histoire d'un texte ", dans *Inter litteras & scientias. Recueil d'études en hommage à Catherine Jacquemard*, éd. Brigitte Gauvin et Marie-Agnès Lucas-Avenel, Caen, Presses Universitaires de Caen (Miscellanea), 2019, p. 185-198.

Jacquemard Catherine, Gauvin Brigitte, Lucas-Avenel Marie-Agnès (ed.), avec la collaboration de C. Février et F. Lecocq, *HORTVS SANITATIS, Livre IV, Les poissons*, Caen, Presses universitaires de Caen (Fontes & Paginæ), 2013. Consultable en ligne : http://www.unicaen.fr/puc/sources/depiscib

Kummer, Robert, " Semantic Technologies for Manuscript Descriptions – Concepts and

Visions ", dans *Codicology and Palaeography in the Digital Age 2* éd. Franz Fischer, Christiane Fritze, et Georg Vogeler, Books on Demand, Norderstedt, 2010, p. 133-154.

Ceux-ci sont complétés par une bibliographie sur Zotero.org : https://www.zotero.org/groups/ichtya/items

La première édition a été publiée en 2013 : B. Gauvin, C. Jacquemard et M.-A. Lucas-Avenel (éd), *Hortus sanitatis : Livre IV, Les Poissons*, Caen, Presses universitaires de Caen (*Fontes et paginae*), 2013. Consultable en ligne : https://www.unicaen.fr/puc/sources/depiscibus/. Devraient suivre les éditions du livre 24 du *De animalibus* d'Albert le Grand et les livres 6 et 7 du *De natura rerum* de Thomas de Cantimpré.

La bibliothèque Ichtya est accessible à l'adresse suivante : https://www.unicaen.fr/ichtyalab/bibliotheque/accueil. Elle est partiellement publique.

Le thesaurus des poissons et créatures aquatiques est accessible en intégralité à l'adresse suivante : https://www.unicaen.fr/ichtyalab/thesaurus/accueil. Il est alimenté régulièrement au fur et à mesure de l'indexation du corpus de la bibliothèque Ichtya. Il comprend à la date de cette proposition de communication 2 290 entrées.

https://www.craham.cnrs.fr/.

http://www.unicaen.fr/recherche/mrsh/document_numerique.

Le répertoire constitué pour l'édition du *De piscibus* est consultable à cette adresse : https://www.unicaen.fr/puc/so

B. Gauvin, C. Jacquemard et M.-A. Avenel (éd), *Hortus sanitatis : Livre IV, Les Poissons*, Caen, Presses universitaires de Caen (*Fontes et paginae*), 2013. Consultable en ligne : https://www.unicaen.fr/puc/sources/depiscibus/.

Le *template* renseignant la forme normalisée avait été inséré devant chaque terme qu'on voulait retrouver dans l'index.

Après sa publication aux Presses universitaires de Caen, l'indexation XML du texte du *De piscibus* a donc dû être mise à jour pour intégrer la bibliothèque Ichtya.

Le nom du poisson dans le texte est encodé au moyen de l'élément **term**. L'élément est qualifié d'une valeur de langue (attribut @xml:lang) et d'une référence à sa notice (identifiant de la notice en valeur d'attribut @ref).

Voir ainsi l'index généré pour l'état actuel public de la bibliothèque Ichtya : https://www.unicaen.fr/ichtyalab/biblio

Mots-Clés: Modélisation de données, ichtyologie latine, XML, TEI, corpus de textes, thésaurus, transmission des textes, exploration de corpus