Le corpus Ichtya en XML-TEI et graphes

Traiter, visualiser et analyser les noms de poissons et créatures aquatiques

Marie BISSON(1), Pierre-Yves BUARD(2),
Brigitte GAUVIN(3), Barbara JACOB(4)

(1)(2)(4) Pôle Document numérique, Maison de la recherche en sciences humaines,
USR 3486 Université de Caen Normandie, France, CNRS
(3) Centre Michel de Boüard, CRAHAM, UMR 6273
Université de Caen Normandie, France, CNRS

1 Le corpus Ichtya et les objectifs de recherche

La bibliothèque Ichtya, comme le programme Ichtya auquel elle appartient, est tout à la fois le produit d'un rêve et celui de la sérendipité, le fruit de la rencontre entre l'œuvre d'un savant du XVIII^{e} siècle et du langage XML-TEI.

Il y a une quinzaine d'années, quelques enseignants-chercheurs latinistes de l'université de Caen se sont réunis autour de l'édition commune d'une encyclopédie latine de la fin du Moyen Âge, intitulée l'*Hortus sanitatis*. Il s'agissait d'en donner une édition classique, accompagnée d'une traduction et d'un commentaire. Pour des raisons diverses, qui n'étaient pas toutes d'ordre scientifique, notre choix s'est porté sur la partie consacrée aux poissons. Or, au fur et à mesure que nous avancions dans notre travail et que nous progressions dans l'identification des sources de cette encyclopédie, il est apparu de manière évidente que l'*Hortus sanitatis* était le dernier maillon d'une chaîne de transmission qui remontait à Aristote via les encyclopédistes de l'Antiquité latine, de l'Antiquité tardive et du Moyen Âge. En d'autres termes, identifier et commenter les sources de l'*Hortus sanitatis*, c'était identifier et commenter des fragments de toutes les encyclopédies qui avaient conduit à celui-ci.

C'est également en avançant dans ce travail que nous avons compris que l'*Hortus* n'était finalement qu'un énorme centon, une immense banque de citations réorganisées thématiquement puis alphabétiquement (1066 pour le *Livre des poissons*). La mise

en évidence de cette structure en segments nous a encouragés à proposer à la fois une édition numérique et une édition papier. Un encadrement solide de la part des ingénieurs du Pôle Document numérique, une écoute attentive de nos besoins et de nos objectifs, un accompagnement technique sans faille, un zeste de maïeutique et la promesse que, grâce au langage XML, nous ne ferions qu'une fois le travail pour aboutir à une édition sur double support achevèrent de nous convaincre.

Tandis que l'édition numérique prenait forme, un autre projet naissait : puisque le travail effectué sur l'*Hortus*, indexation des noms comme des citations, était directement réutilisable, pourquoi ne pas tenter de concrétiser le rêve de Peter Artedi (1705-1735), fondateur de l'ichtyologie moderne, qui dans sa *Bibliotheca Ichthyologica* a recensé toutes les œuvres et parties d'œuvres qui traitaient d'ichtyologie, depuis l'Antiquité jusqu'à son époque ? Alors que Peter Artedi s'était arrêté à la liste commentée des ouvrages, le numérique nous permettait de reconstituer la bibliothèque complète, avec mise à disposition des œuvres dans leur intégralité, en l'augmentant de toutes sortes d'outils de consultation déjà mis en œuvre dans l'édition numérique de l'*Hortus sanitatis*. Ce fut la base du programme Ichtya, fondé par Catherine Jacquemard.

Ce programme comporte, outre une bibliographie interactive consultable sur Zotero[1], trois grands volets : des éditions de texte bilingues, sur le modèle de l'*Hortus sanitatis* (Jacquemard, Gauvin et Lucas-Avenel, 2013), au nombre desquelles figureront les livres 6 et 7 du *Liber de natura rerum* de Thomas de Cantimpré ; un thesaurus des noms de poissons latins[2] ; et une bibliothèque numérique, la bibliothèque Ichtya[3], sur laquelle nous allons nous pencher pour commencer.



Fig. 1 : Page d'accès au corpus latin de la bibliothèque Ichtya

L'objectif de la bibliothèque Ichtya n'est pas d'élaborer ou de réunir des éditions critiques, dont certaines existent d'ailleurs déjà, mais de mettre en ligne, à la disposition de la communauté de chercheurs s'intéressant au savoir ichtyologique, un corpus de

[1] [https://www.zotero.org/groups/356871/ichtya/library], consulté le 27 juillet 2021.
[2] [https://ichtya.unicaen.fr/lab/thesaurus/accueil.html], consulté le 27 juillet 2021.
[3] [https://ichtya.unicaen.fr/lab/bibliotheque/accueil.html], consulté le 27 juillet 2021.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

textes latins consacrés à ce sujet, indexés et balisés, sous forme de site Web interrogeable par un moteur de recherche. Un ensemble de textes, allant du I^er^ au XVII^e^ siècle, a d'abord été répertorié, océrisé et pré-encodé en XML-TEI au CRAHAM et au pôle Document numérique de la MRSH. Ce corpus est constitué de textes antiques (Pline l'Ancien, Ausone, Ovide, Solin, Basile de Césarée, Ambroise de Milan, Isidore de Séville), médiévaux (Alexandre Nequam, Barthélemy l'Anglais, Thomas de Cantimpré, Albert le Grand, Vincent de Beauvais, Arnold de Saxe, Marcus d'Orvieto) et modernes (l'*Hortus sanitatis*, Carolus Figulus, Pierre Belon, Peter Artedi).

Le corpus réuni dans la bibliothèque, bien qu'homogène par sa thématique, est évidemment hétérogène pour tous les autres critères : il mêle œuvres complètes, livres ou chapitres d'œuvres, selon l'origine des textes ; par ailleurs, si la plupart de ceux-ci relèvent du genre encyclopédique, on trouve aussi des poèmes (Ovide, Ausone) ou des dialogues (Figulus[4]) ; enfin, certaines des œuvres ont été étudiées et éditées, d'autres ne le sont pas. L'ensemble de la bibliothèque représente un volume textuel de plus de 95 000 mots.

Ce corpus a ensuite été traité de manière à répondre à un double objectif : créer un lexique des noms de poissons latins à travers les siècles et comprendre avec précision comment se fait la transmission des savoirs au fil du temps.

Il a donc été décidé d'indexer en priorité dans les textes les noms de poissons et de créatures aquatiques, afin de permettre aux utilisateurs de faire une recherche par nom, d'accéder immédiatement à tous les passages du corpus Ichtya dans lequel le nom apparaît, et d'en connaître les variantes et les synonymes. Pour le chercheur qui lit un texte dans la bibliothèque Ichtya, les noms de poissons indexés se distinguent du reste du texte par leur couleur. Lorsqu'il clique dessus, la fiche d'identification de l'animal apparaît sans qu'il ait besoin de sortir du texte. Nous développerons ce point plus loin.

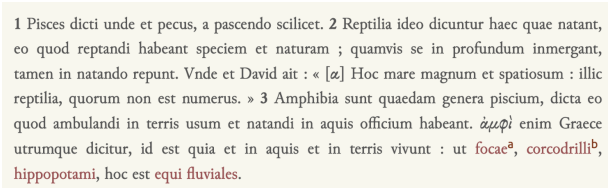

1 Pisces dicti unde et pecus, a pascendo scilicet. **2** Reptilia ideo dicuntur haec quae natant, eo quod reptandi habeant speciem et naturam ; quamvis se in profundum inmergant, tamen in natando repunt. Vnde et David ait : « [α] Hoc mare magnum et spatiosum : illic reptilia, quorum non est numerus. » **3** Amphibia sunt quaedam genera piscium, dicta eo quod ambulandi in terris usum et natandi in aquis officium habeant. ἀμφὶ enim Graece utrumque dicitur, id est quia et in aquis et in terris vivunt : ut focae[a], corcodrilli[b], hippopotami, hoc est equi fluviales.

Fig. 2 : Affichage en ligne de l'extrait d'Isid. orig. XII, 6 1-3

Mais conjointement, dans la plus grande partie des textes, on s'est aussi attaché à identifier les citations ou emprunts à des œuvres antérieures, à contextualiser ces fragments et à expliciter, dans la mesure du possible, la manière dont ils ont été transmis. Le suivi d'un même fragment à travers les siècles permet de comprendre où se sont produites les erreurs et pour quelle raison certaines aberrations apparaissent, parfois au milieu d'un ensemble très cohérent. Dans ce but, on a donc veillé à respecter la même granularité de description des textes, du traité jusqu'au segment de citation, pour permettre des traitements automatiques et semi-automatiques du corpus par la

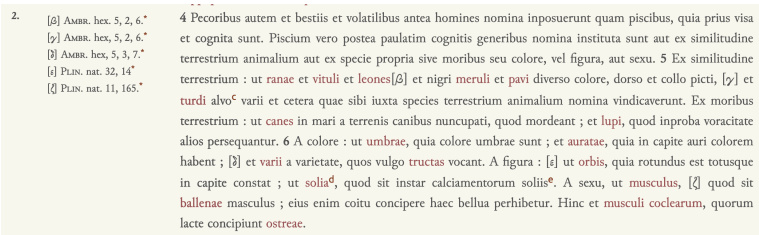
[4]Sur l'auteur Carolus Figulus, voir notamment Gauvin, 2019 : p. 33-36.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

suite. La structure du texte est reproduite et encodée (, <head>, <p>, <lg> et <l>) et chaque citation ou emprunt identifié est encodé au moyen de l'élément <seg> et l'identification est précisée au moyen du template :

```
<bibl>
    <author>Abréviation normalisée de l'auteur</author>
    <title>titre de l'œuvre</title>
    <biblScope>référence absolue à la source</biblScope>
    <note>texte et commentaire du texte cité ou emprunté</note>
</bibl>
```

Le lecteur qui consulte un texte dans la bibliothèque Ichtya voit ainsi, dans la marge, quelles sont les sources du passage qu'il est en train de lire. Un simple clic sur l'astérisque qui accompagne la référence lui permet d'afficher la citation source (élément) et de comparer les deux textes.


2.
[β] Ambr. hex. 5, 2, 6.*
[γ] Ambr. hex, 5, 2, 6.*
[δ] Ambr. hex, 5, 3, 7.*
[ε] Plin. nat. 32, 14*
[ζ] Plin. nat. 11, 165.*

4 Pecoribus autem et bestiis et volatilibus antea homines nomina inposuerunt quam piscibus, quia prius visa et cognita sunt. Piscium vero postea paulatim cognitis generibus nomina instituta sunt aut ex similitudine terrestrium animalium aut ex specie propria sive moribus seu colore, vel figura, aut sexu. 5 Ex similitudine terrestrium : ut ranae et vituli et leones[β] et nigri meruli et pavi diverso colore, dorso et collo picti, [γ] et turdi alvo[c] varii et cetera quae sibi iuxta species terrestrium animalium nomina vindicaverunt. Ex moribus terrestrium : ut canes in mari a terrenis canibus nuncupati, quod mordeant ; et lupi, quod inproba voracitate alios persequantur. 6 A colore : ut umbrae, quia colore umbrae sunt ; et auratae, quia in capite auri colorem habent ; [δ] et varii a varietate, quos vulgo tructas vocant. A figura : [ε] ut orbis, quia rotundus est totusque in capite constat ; ut solia[d], quod sit instar calciamentorum soliis[e]. A sexu, ut musculus, [ζ] quod sit ballenae masculus ; eius enim coitu concipere haec bellua perhibetur. Hinc et musculi coclearum, quorum lacte concipiunt ostreae.

Fig. 3 : Affichage marginal des sources de l'extrait d'Isid. orig. XII, 6, 4-6

Voici un extrait du codage correspondant en XML-TEI d'Isid. orig. XII, 6, 4-6 :

```
[…]
<p xml:id="isid.12.6.p2">
    <seg n="4" next="5" prev="3" type="segmentation_classique"
    xml:id="isid.12.6.4">
        <num type="segmentation">4</num> Pecoribus autem et bestiis et
        volatilibus antea homines nomina inposuerunt quam piscibus, quia prius
        visa et cognita sunt. Piscium vero postea paulatim cognitis generibus
        nomina instituta sunt aut ex similitudine terrestrium animalium aut ex specie
        propria sive moribus seu colore, vel figura, aut sexu.
    </seg>
    <seg n="5" next="6" prev="4" type="segmentation_classique"
    xml:id="isid.12.6.5">
        <num type="segmentation">5</num>Ex similitudine terrestrium :
        ut <term ref="wpddn/indexAnimaux.xml#pddn_pisc.1495530567489"
        type="animal" xml:lang="la">ranae</term>
         et <term ref="wpddn/indexAnimaux.xml#pddn_pisc.1459242964377"
        type="animal" xml:lang="la">vituli</term>
         et <term ref="wpddn/indexAnimaux.xml#pddn_pisc.1512838073251"
        type="animal" xml:lang="la">leones</term>
        <seg type="citation">
```

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

```
        <bibl>
            <author>Ambr.</author>
            <title type="oeuvre">hex.</title>
            <biblScope>5, 2, 6.</biblScope>
            <note type="sources">ut nigri merulae, pavi diverso colore dorsa et
            colla depicti sint.</note>
        </bibl>
        et nigri <term
        ref="wpddn/indexAnimaux.xml#pddn_pisc.1542120588704"
        type="animal" xml:lang="la">meruli</term>
        et <term ref="wpddn/indexAnimaux.xml#pddn_pisc.1512728748175"
        type="animal" xml:lang="la">pavi</term>
        diverso colore, dorso et collo picti
    </seg>,
    <seg type="citation">
        <bibl>
            <author>Ambr.</author>
            <title type="oeuvre">hex</title>
            <biblScope>5, 2, 6.</biblScope>
            <note type="sources">turdi alvo varii</note>
        </bibl>
        et <term ref="wpddn/indexAnimaux.xml#pddn_pisc.1531924961200"
        type="animal" xml:lang="la">turdi</term> alvo
            <note n="C" place="foot" resp="MB" type="apparat"
            xml:id="ftn_apparat_4">albo <hi
            rend="italic">Oxford</hi>.</note> varii
    </seg>
    et cetera quae sibi iuxta species terrestrium animalium nomina
    vindicaverunt. Ex moribus terrestrium : ut
    <term ref="wpddn/indexAnimaux.xml#pddn_pisc.1459263456147"
    type="animal" xml:lang="la">canes</term> in mari a terrenis canibus
    nuncupati, quod mordeant ; et
    <term ref="wpddn/indexAnimaux.xml#pddn_pisc.1459246129966"
    type="animal" xml:lang="la">lupi</term>, quod inproba voracitate alios
    persequantur.
</seg>
<seg n="6" next="7" prev="5" type="segmentation_classique"
xml:id="isid.12.6.6">
    <num type="segmentation">6</num> A colore : ut
    <term ref="wpddn/indexAnimaux.xml#pddn_pisc.1508321801926"
    type="animal" xml:lang="la">umbrae</term>,
    quia colore umbrae sunt ; et
    <term ref="wpddn/indexAnimaux.xml#pddn_pisc.1459346965767"
    type="animal" xml:lang="la">auratae</term>,
    quia in capite auri colorem habent ;
    <seg type="citation">
        <bibl>
            <author>Ambr.</author>
            <title type="oeuvre">hex</title>
            <biblScope>5, 3, 7.</biblScope>
```

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

```
                <note type="sources">ut varii maiores quos <term
                ref="pddn_pisc.1554367887845.xml#pddn_pisc.1554367887845"
                type="animal" xml:lang="la">troctas</term> vocant.</note>
            </bibl>
            et <term ref="pddn_pisc.1554905940400.xml#pddn_pisc.1554905940400"
            type="animal" xml:lang="la">varii</term>
            a varietate, quos vulgo
            <term ref="pddn_pisc.1505229101958.xml#pddn_pisc.1505229101958"
            type="animal" xml:lang="la">tructas</term> vocant
        </seg>.
        A figura :
        <seg type="citation">
            <bibl>
                <author>Plin.</author>
                <title type="oeuvre">nat.</title>
                <biblScope>32, 14</biblScope>
                <note type="sources">Rotundus est sine squamis totusque capite
                constat.</note>
            </bibl>
            ut <term ref="wpddn/indexAnimaux.xml#pddn_pisc.1512730298337"
            type="animal" xml:lang="la">orbis</term>, quia rotundus est totusque in
            capite constat
        </seg> ;
        ut <term ref="wpddn/indexAnimaux.xml#pddn_pisc.1542121203859"
        type="animal" xml:lang="la">solia</term>
            <note n="D" place="foot" resp="MB" type="apparat"
            xml:id="ftn_apparat_5">solea <hi rend="italic">Oxford</hi>.</note>,
        quod sit instar calciamentorum soliis
            <note n="E" place="foot" resp="MB" type="apparat"
            xml:id="ftn_apparat_6">soleis <hi rend="italic">Oxford</hi>.</note>.
        A sexu, ut <term ref="wpddn/indexAnimaux.xml#pddn_pisc.1459243491740"
        type="animal" xml:lang="la">musculus</term>,
        <seg type="citation">
            <bibl>
                <author>Plin.</author>
                <title type="oeuvre">nat.</title>
                <biblScope>11, 165.</biblScope>
                <note type="sources">Musculus marinus qui ballaenam antecedit.</note>
            </bibl>
            quod sit <term ref="wpddn/indexAnimaux.xml#pddn_pisc.1458560354760"
            type="animal" xml:lang="la">ballenae</term> masculus
        </seg> ;
        eius enim coitu concipere haec bellua perhibetur. Hinc et
        <term ref="wpddn/indexAnimaux.xml#pddn_pisc.1459243491740"
        type="animal" xml:lang="la">musculi</term>
        <term ref="wpddn/indexAnimaux.xml#pddn_pisc.1459262069910"
        type="animal" xml:lang="la">coclearum</term>,
        quorum lacte concipiunt
        <term ref="wpddn/indexAnimaux.xml#pddn_pisc.1459244504759"
        type="animal" xml:lang="la">ostreae</term>.
    </seg>
</p>
```

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

Le fait d'avoir utilisé le langage XML en suivant les recommandations de la Text Encoding Initiative (TEI) à la fois pour les éditions, la bibliothèque Ichtya et le thesaurus de noms de poissons et créatures aquatiques permet une interopérabilité totale. Grâce à la complémentarité synchronisée des trois volets, 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 : un nom de poisson indexé dans un texte de la bibliothèque se trouve immédiatement relié à une fiche dans le thesaurus ; une citation balisée dans un texte permet d'accéder directement à la source de celle-ci ou à la bibliothèque des citations du même auteur dans le corpus. Cette utilisation du langage XML permet de créer des connexions avec d'autres projets, comme SourcEncyMe, consacré à la constitution d'un corpus annoté en ligne des encyclopédies médiévales latines de l'IRHT[5] ; les textes indexés et balisés à Caen peuvent ainsi être versés dans une autre base, ici, celle de SourcEncyMe[6].

L'utilisateur de la bibliothèque Ichtya a de plus à sa disposition deux outils précieux accessibles depuis la page d'accueil. D'une part, un index des noms de poissons qui lui permet d'accéder directement aux passages où se trouve le terme qui l'intéresse :

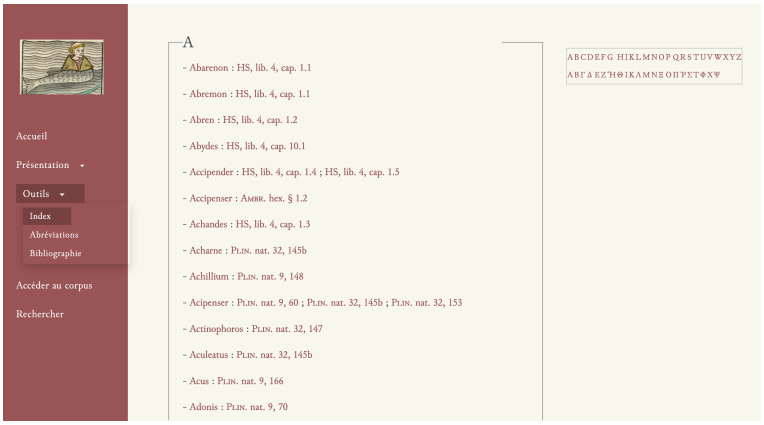


Fig. 4 : Index des noms de poissons

D'autre part, une fonction recherche plein texte dans laquelle il peut sélectionner tout ou partie du corpus : la recherche s'effectue à la fois dans les textes et dans les notes.

Notons pour finir que le corpus de la bibliothèque Ichtya est un corpus d'œuvres latines. Mais certains textes sont accompagnés d'une traduction française, effectuée par les membres du groupe Ichtya. Celles-ci sont également signalées sur le site, qu'elles soient disponibles ou à venir.

Tous les textes n'ont pas encore été traités dans leur totalité. Ils sont ouverts au public au fur et à mesure de l'avancement des travaux de l'équipe.

[5]Le programme SourcEncyMe est dirigé par Isabelle Draelants, voir [http ://sourcencyme.irht.cnrs.fr/], consulté le 20 août 2021.

[6]Le texte latin du *De piscibus* est disponible dans SourcEncyMe : [http://sourcencyme.irht.cnrs.fr/encyclopedie/voir/137], consulté le 27 juillet 2021.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

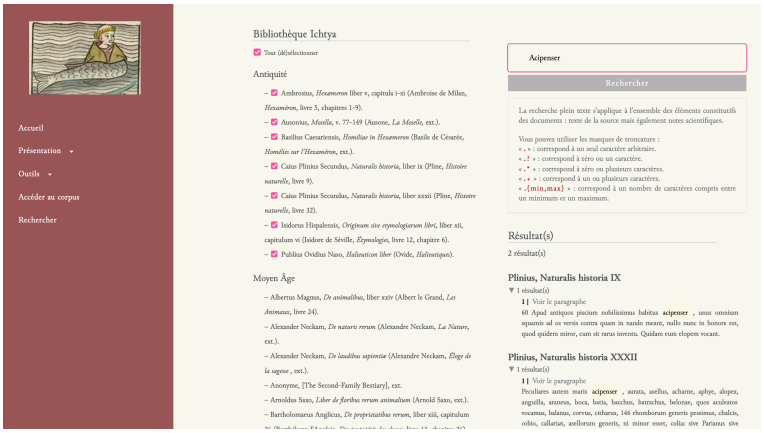


Fig. 5 : Moteur de recherche plein texte

2 La création du thesaurus et son articulation à la bibliothèque

2.1 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, la création d'un index semblait indispensable pour permettre aux utilisateurs de faire une recherche par nom de poisson ou de créature aquatique et accéder immédiatement à tous les passages du corpus Ichtya dans lesquels le nom apparaît.

Pour la première édition de texte (*De piscibus* de l'*Hortus sanitatis* ; Jacquemard, Gauvin et Avenel (éd.), 2013), la méthodologie d'indexation retenue était traditionnelle : un marqueur normalisé précédait le terme rencontré. C'est la méthode par ancre qui est également utilisée par les traitements de texte actuels ou les éditeurs de l'édition papier. En XML-TEI, l'indexation par marqueur est réalisée au moyen de l'élément <index> contenant un élément <term> dont le nœud texte donne la forme normalisée du concept. Dans l'édition du *De piscibus*, une typologie – au moyen de l'attribut @indexName – a été élaborée pour permettre la création de cinq index : les noms de poissons en latin (@indexName="piscis") et les noms des autres animaux cités (@indexName="animales") pour la version latine du texte ; les noms de poissons en français (@indexName="poissons"), les noms des autres animaux (@indexName="animaux") pour la traduction française du texte et les noms scientifiques des poissons (@indexName=”poissons_sc”) cités en note de commentaire dans le flux de la traduction française. L'objectif étant d'étudier le lexique ichtyologique et sa transmission au fil des siècles, toutes les formes latines au nominatif ont été retenues, y compris les variantes graphiques alors que le reste du texte a fait l'objet d'une normalisation (ponctuation, lettres ramistes, géminées, etc. ; voir Jacquemard, Gauvin et Avenel (éd.) 2013 : p. 67-69). Dans l'édition en ligne, le lecteur peut retrouver deux index : l'index des formes latines et l'index des formes françaises (les poissons et les autres animaux ayant été rassemblés). Dans l'édition papier, le lecteur trouvera quatre index : poissons latins ; poissons français ; autres animaux latins ; autres

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

animaux français. L'index des noms scientifiques n'a pas été produit dans les deux versions publiées, mais a été exploité par les éditrices pour l'analyse.

Si cette méthodologie a donné de bons résultats pour l'édition multi-support du *De piscibus* – des index papier traditionnels et des index sur l'édition en ligne ramenant à la page du chapitre où se trouvait l'occurrence –, on peut regretter l'absence de mise en surbrillance de la forme indexée lors du retour du texte.

Par ailleurs, cette méthode demande une rigueur très importante pour la constitution des formes normalisées, faisable à l'échelle d'une édition, mais plus fastidieuse à réaliser à l'échelle d'un corpus « mouvant » dans le cadre d'une recherche en cours, celui de la bibliothèque Ichtya, puisque la création des entrées d'index va s'appuyer sur l'exacte comparaison des chaînes de caractères.

Après la publication du *De piscibus,* l'indexation du corpus Ichtya a amené des questions qui s'étaient déjà posées lors de l'indexation de la première édition, mais de façon plus importante encore : indexait-on le terme rencontré ? l'identification du poisson était-elle réalisée par le chercheur ? Inspiré de la méthodologie qui avait été mise en place pour l'indexation des personnes, lieux et œuvres rencontrés dans les sources anciennes éditées au CRAHAM et au pôle Document numérique de Caen (notamment Bisson, 2015 et Combalbert, 2019), nous avons décidé de mettre en place un réservoir de notices d'autorité pour toutes les créatures aquatiques que nous rencontrions dans le corpus Ichtya, ce qui permettait de garder une indexation des formes présentes dans le corpus tout en proposant, au sein de ces notices externes, des identifications anciennes et modernes.

La méthodologie d'indexation de la bibliothèque Ichtya n'a pas été pensée à plat, texte après texte, en ajoutant les chaînes de caractères normalisées terme après terme, mais conçue au moyen d'un thesaurus indépendant[7] comme un réservoir de notices d'autorité permettant d'établir un lien entre des formes rencontrées dans les textes (de la bibliothèque Ichtya) et des notices d'autorité (correspondant à une forme normalisée et un certain nombre de connaissances scientifiques, bibliographiques et de relation).

Cette remise en question de la méthodologie d'indexation a correspondu à la création, dans le cadre du GDRI Zoomathia, d'un thesaurus[8] dont deux branches rejoignent nos préoccupations : une branche consacrée aux zoonymes et une autre consacrée aux archéotaxons. Nous avons dans un premier temps envisagé d'utiliser ce thesaurus pour notre travail. La classification des zoonymes a été créée en s'appuyant sur la classification des animaux et une équivalence français/latin a été faite en s'appuyant sur les identifications faites dans l'*Histoire naturelle* de Pline. Cependant les « poissons », ou plutôt créatures aquatiques présents dans notre corpus, ne correspondent pas toujours à une réalité bien précise et nous ne pouvions pas nous approprier la branche des zoonymes, sinon notre indexation aurait porté sur l'identification et non sur la forme graphique latine du nom. *Aranea*, *Aranea marina* et *Araneus* auraient été rassemblés dans l'entrée *Araneus2.* Comment indexer le terme *echinus* qui est l'oursin

[7] 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.

[8] [http://web.cepam.cnrs.fr/opentheso/{]}, consulté le 27 juillet 2021. Ce thesaurus est réalisé avec Opentheso [https://github.com/miledrousset/opentheso{]} ; les données sont encodées en SKOS.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

de mer chez Pline, mais le rémora (*echeneis* dans le thesaurus Zoomathia) dans le *De piscibus*, chez Thomas de Cantimpré et Albert le Grand[9] ?

Parmi les archéotaxons existent plusieurs branches rassemblant des entrées proches de nos préoccupations : aquatique, marin, poisson, etc. Mais là encore, la correspondance latin/français reproduit les identifications de Pline, tandis qu'à l'intérieur de notre corpus un même nom (graphie) de poisson peut correspondre à plusieurs identifications, à plusieurs réalités. Or, garder ces identifications multiples était nécessaire dans le cadre de notre recherche sur la transmission du savoir ichtyologique.

2.2 Le thesaurus

Nous avons donc opté pour un thesaurus propre au projet Ichtya, tout en choisissant de garder un lien étroit avec le thesaurus Zoomathia, c'est-à-dire en renvoyant pour les notices concernées à l'URL du thesaurus Opentheso Zoomathia[10].

À l'instar de ce qui a été fait pour les notices « Personnes », « Lieux » et « Œuvres » (voir à ce sujet Bisson et Goloubkoff, 2020), le thesaurus des créatures aquatiques est construit en XML-TEI. Il est composé d'autant de fichiers XML (notices) que de formes latines (au nominatif), grecques ou vernaculaires différentes rencontrées dans le corpus de la bibliothèque Ichtya.

Chaque notice possède un identifiant propre (@xml :id sur l'élément racine <TEI>) et quand on rencontre un ichtyonyme dans le corpus Ichtya, on peut établir un lien entre cette forme et la notice. Chaque forme graphique rencontrée fait l'objet d'une notice séparée. On va ainsi distinguer l'*araneus, i*[11] m. (Plin. *nat.* 9, 155), l'*aranea, ae*[12] f. (Isid. orig. 12, 6, 18) et l'*aranea marina*[13] (Avic. *Canon*, 4, 6, 5, 26 cité par *HS* 4, 4, 1) et créer une notice pour chacune de ces formes. On va pouvoir établir des liens entre ces notices. Dans le premier cas, s'il ne s'agit que de variantes graphiques : *accipenser/accipender/acipenser,* et nous allons distinguer les notices mères et des notices filles. La notice mère porte l'ensemble des informations : énumération des variantes graphiques et des synonymies (équivalence dans une autre langue) ; identification ; sources et études ; note de commentaire. La notice fille se contente alors de renvoyer à la notice mère (sans répéter l'ensemble des informations susdites) et en ne précisant que la source d'où est tirée cette graphie et les informations propres à cette forme. Pour déterminer la notice mère, on s'appuie sur deux critères dans cet ordre :

- la présence du terme dans un dictionnaire de latin classique ; par exemple la forme *acipenser*[14] est la forme couramment usitée (utilisée dans les éditions de Cicéron, Pline, etc.), notamment retenue par le Gaffiot. *Acipenser* est donc la

[9] Voir la notice [https://ichtya.unicaen.fr/lab/thesaurus/doc/pddn_pisc.1459245667629.html], consulté le 27 juillet 2021.

[10] Voir par exemple la notice *Acipenser*, [https://ichtya.unicaen.fr/lab/thesaurus/doc/pddn_pisc.1497560613208.html], consulté le 27 juillet 2021.

[11] [https://ichtya.unicaen.fr/lab/thesaurus/doc/pddn_pisc.1458555357648.html], consulté le 27 juillet 2021.

[12] [https://ichtya.unicaen.fr/lab/thesaurus/doc/pddn_pisc.2.html], consulté le 27 juillet 2021.

[13] [https://ichtya.unicaen.fr/lab/thesaurus/doc/pddn_pisc.1496740194336.html], consulté le 27 juillet 2021.

[14] [https://ichtya.unicaen.fr/lab/thesaurus/doc/pddn_pisc.1497560613208.html], consulté le 27 juillet 2021.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

notice mère, même si on rencontre également les formes *accipenser* et *accipender* dans notre corpus ;

- la première occurrence par ordre alphabétique ; par exemple on va retenir comme « notice mère » *aranea* (forme utilisée par Isidore) plutôt qu'*araneus* (forme utilisée par Pline).

La définition de ces deux critères – qui peuvent sembler subjectifs – nous a permis d'établir une cohérence à l'échelle de notre corpus et nous a évité de répéter des informations pour des variantes purement graphiques tout en gardant la même logique que pour l'indexation du *De piscibus*.

Deuxième cas de figure de relation, les termes sont « en relation » mais le lien établi entre les notices est différent d'une variante graphique : il peut s'agir du même poisson ou bien la forme est proche mais la littérature ne permet pas de déterminer s'il s'agit du même poisson. C'est le cas entre *aranea* et *aranea marina* (voir Jacquemard, Gauvin et Lucas-Avenel, 2013, 93 *s. v. Araneus)*. Les deux notices ont un statut de notice « mère » mais on va pouvoir établir un lien entre les deux au sein d'une note de relation. Par exemple, pour la notice *Aranea* :

<entry>

<form type="ppal"xml :lang="la">Aranea</ form>
<form type="variante"xml :lang="la">Araneus</form>
[…]

</entry>

Chaque notice est comprise au sein de l'élément <entry> et comprend au moins un élément <form> qualifié de l'attribut @type="ppal"(pour principal) et d'un attribut qualifiant la langue du terme. Dans le cas d'une notice mère (notice par défaut), on peut ajouter des formes variantes (@type="variante"), une forme d'origine (notamment pour les termes directement issus du grec ; @type="origine") et des synonymies (@type="traduction").

Cette indication de source s'accompagne, autant que possible, d'un ou plusieurs blocs d'identification (élément <sense>) comprenant :

- une proposition d'identification en français et éventuellement la nomenclature scientifique binomiale (notice mère) ou d'un renvoi vers une notice mère (notice fille) : élément <def>.
- au moins une référence précise à une source dans laquelle le terme apparaît (élément obligatoire) : élément <note @subtype="source"@type="bibliographie">;
- pour les appellations latines et grecques, de la référence à la littérature scientifique qui valide ces identifications : <note @subtype="etude"@type="bibliographie">;
- ces identifications peuvent être accompagnées d'une note de commentaire : <note @subtype="autres"@type="commentaire">.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

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 (cf. supra), indication qui figure en tête de la fiche, à la place de l'identification (élément <def>) ; d'autre part aux autres termes désignant le même animal sous un autre nom : <note @subtype="relation"@type="commentaire">. L'indexation par le biais du format XML permet de faire des liens directement d'une forme à l'autre au moyen de l'élément <term ref="[lien vers l'identifiant de la notice liée]"type="animal">. Par exemple, pour la notice d'*Acipenser* :

```
<TEI xmlns:tei="http://www.tei-c.org/ns/1.0"
xsi:schemaLocation="http://www.tei-c.org/ns/1.0
http://www.unicaen.fr/mrsh/pddn/schemas/tei_index.xsd"
xml:id="pddn_pisc.1497560613208">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>Acipenser</title>
        <author xml:id="CJacquemard">Catherine Jacquemard</author>
      </titleStmt>
      <publicationStmt><p>Notice Poisson</p></publicationStmt>
      <sourceDesc><bibl>XML natif</bibl></sourceDesc>
    </fileDesc>
    <encodingDesc>
      <appInfo>
        <application ident="TEI_indexation.zip" version="1.0">
          <label>Environnement pour l'indexation</label>
          <desc>Environnement pour encoder des notices de thesauri</desc>
          <ref
          target="http://www.unicaen.fr/recherche/mrsh/document_numerique/
          seme/TEI_indexation.zip">Lien vers l'environnement dans SEM</ref>
        </application>
      </appInfo>
    </encodingDesc>
  </teiHeader>
  <text>
    <body>
      <div>
        <entry
        sameAs="http://web.cepam.cnrs.fr/opentheso/?idc=107079&amp;idt=1"
        type="animal">
            <form type="ppal" xml:lang="la">Acipenser</form>
            <form type="variante" xml:lang="la">Accipenser</form>
            <form type="variante" xml:lang="la">Accipender</form>
            <form type="variante" xml:lang="la">Aquipenser</form>
            <form type="origine" xml:lang="grc">Ἀκιπήσιος</form>
            <sense xml:id="pddn_ident.06162017-103456">
```

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

```
            <def>l'esturgeon <hi rend="italic">(Acipenser sturio</hi> Linné,
            1758).</def>
            <note subtype="etude" type="bibliographie">
              <ref
              target="https://api.zotero.org/groups/356871/items/NX73ZBFU
              " type="bibref">De Saint-Denis 1947</ref>, 1-3.
            </note>
            <note subtype="source" type="bibliographie">
               <hi rend="small-caps">Cic.</hi> fin. 2, 25 ; 2, 91 ; Tusc. 3, 43 ;
               <hi rend="small-caps">Hor.</hi> sat. 2, 246 ;
               <hi rend="small-caps">Mart.</hi> 13, 91 ;
               <hi rend="small-caps">Ov.</hi> hal. 134 ;
               <hi rend="small-caps">Plin.</hi> nat. 9, 60 ; 32, 145 et 153;
               <hi rend="small-caps">Ambr.</hi> hex. 5, 1, 2 ;
               <hi rend="small-caps">Hier.</hi> epist. 45, 5 ; et surtout
               <hi rend="small-caps">Macr.</hi> sat. 3, 16, 1.</note>
            <note resp="CJ" subtype="autres" type="commentaire">
               De Saint-Denis suggère de voir dans le poisson nommé <hi
               rend="italic">acipenser</hi> l'esturgeon commun alors que l'<hi
               rend="italic">helops</hi>, <hi rend="italic">elops</hi>, avec
               lequel il était confondu selon le témoignage de Pline (<hi
               rend="small-caps">Plin.</hi> nat. 9, 60 et <hi rend="small-
               caps">Plin.</hi> nat. 32, 153), pourrait être le petit estugeon ou
               sterlet (<hi rend="italic">Acipenser ruthenus </hi>Linné, 1758).
            </note>
            <note resp="CJ" subtype="relation" type="commentaire">
               Voir <term ref="
               pddn_pisc.1458915982871.xml#pddn_pisc.1458915982871"
               type="animal">(H)elops</term>.
            </note>
            <note resp="CJ" subtype="autres" type="commentaire">
               Dans quelques rares cas, le nom peut figurer sous les graphies <hi
               rend="italic">aquipenser</hi>, <hi
               rend="italic">acupenser</hi>.
            </note>
         </sense>
      </entry>
    </div>
  </body>
 </text>
</TEI>
```

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

Ce thesaurus fournit un outil de première utilité pour l'étude des synonymies et polyonomies entre les noms de poissons dans les traités ichtyologiques.

Chaque forme de nom de poisson ou créature aquatique rencontrée 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[15] au moyen du template <term ref="[lien vers l'identifiant de la notice liée]"type="animal"xml :lang="iso-639">[forme dans le texte]</term>. L'attribut @ref n'est pas renseigné « à la main », ce lien entre le texte source et la notice de thesaurus étant ajouté grâce à l'outil PluCo[16], construit par le CERTIC de l'université de Caen. Il permet de rapatrier des données d'une base de données XML distante ; dans ce cas, il s'agit de récupérer l'identifiant @xml :id et de l'ajouter en valeur de l'attribut @ref.

Un environnement XMLmind XML Editor a été créé pour l'encodage des notices d'autorité et une méthodologie détaillée expliquant son installation, son utilisation et le schéma XML choisi est accessible en ligne[17].

2.3 Résultats textuels obtenus

L'établissement d'un lien entre une occurrence dans un texte et une notice d'autorité répond aux objectifs de recherche d'étude et de transmission de la connaissance ichtyologique par le biais des ichtyonymes et nous permet de générer l'index de la bibliothèque Ichtya. On applique une XSL qui parcourt le corpus de la bibliothèque Ichtya disponible et qui récupère, pour chaque terme lié à une notice (<term ref="lien vers l'identifiant de la notice liée"type="animal"xml :lang="iso-639">forme dans le texte</term>).

- l'identifiant de la notice citée : @ref des éléments <term> ;
- le nom du fichier dans lequel se trouve l'occurrence ;
- l'identifiant de la section dans lequel le terme se trouve : @xml :id du premier élément ancêtre ;
- l'identifiant du paragraphe dans lequel se trouve le terme : @xml :id de l'élément <p> parent qui permettra de faire le focus lors du retour au texte

À partir de l'identifiant cité (@ref), on peut récupérer – via la XSL – la forme principale de la notice concernée (<form @type="ppal">), puis trier la liste par ordre alphabétique avec les renvois aux textes[18]. Avec une seule XSL, une page HTML est créée qui permet la consultation transversale de la bibliothèque Ichtya. L'index peut être régulièrement mis à jour, au fur et à mesure de l'alimentation publique du corpus.

[15] 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).

[16] [https://www.unicaen.fr/recherche/mrsh/document_numerique/outils/pluco], consulté le 27 juillet 2021.

[17] [https://www.unicaen.fr/recherche/mrsh/document_numerique/outils/thesauri], consulté le 27 juillet 2021.

[18] Voir ainsi l'index généré pour l'état actuel public de la bibliothèque Ichtya : [https://ichtya.unicaen.fr/lab/bibliotheque/index_poissons.html], consulté le 27 juillet 2021.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

Cette méthode permet également de donner accès, de façon dynamique, à chaque notice de thesaurus lors de la lecture d'un texte : chaque terme indexé au sein du corpus est signalé par la couleur et la notice sous forme de fenêtre modale est accessible au clic.

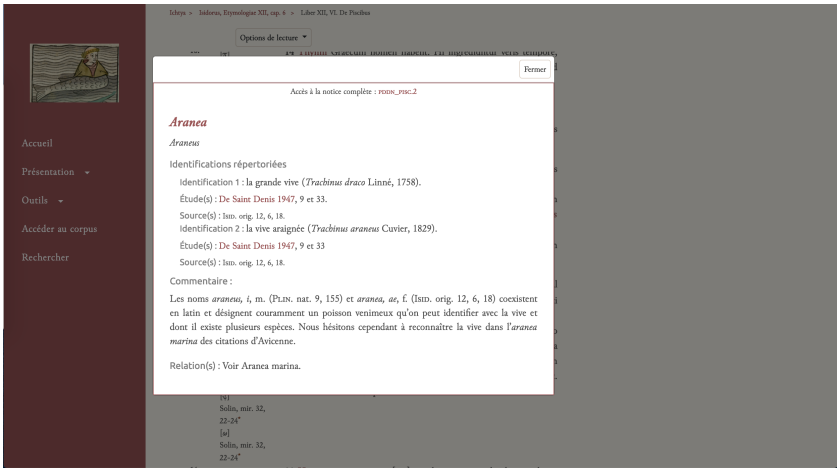


Fig. 6 : Accès à la notice *Aranea* depuis le texte d'Isidore, Isid. orig., XII, cap. 6, 18

Enfin, cette méthode nous a amenés à constituer un réservoir de notices indépendantes et conduits à construire un site différent[19] : le *Thesaurus de noms de poissons et de créatures aquatiques*[20]. Celui-ci permet d'accéder à l'ensemble des notices créées dans le cadre du programme de recherche Ichtya par différents accès.

Le thesaurus est accessible de manière linéaire au moyen de sommaires alphabétiques qu'on a distingués dans deux onglets :

- un accès aux notices latines et grecques[21], qui est le cœur de notre étude ;
- un accès aux notices en langues vernaculaires[22], qu'on rencontre dans quelques textes médiévaux, mais surtout dans les textes de l'époque moderne pour lesquels les auteurs ont donné un certain nombre de synonymies et de traductions. Si nous souhaitions créer et posséder ces notices, car elles enrichissent notre

[19] Tout comme celui de la bibliothèque Ichtya, ce site est construit à partir du moteur d'affichage XML MaX, une réalisation du CERTIC de l'université de Caen Normandie, grâce au soutien de l'Equipex Biblissima. Cette application est basée sur le système de base de données XML native *BaseX* (http://basex.org/) et sur les technologies RESTXQ et XSLT. Les sources sont librement téléchargeables : [https://git.unicaen.fr/pdn-certic/MaX], consulté le 27 juillet 2021.

[20] [https://ichtya.unicaen.fr/lab/thesaurus/], consulté le 27 juillet 2021.

[21] [https://ichtya.unicaen.fr/lab/thesaurus/sommaire.html], consulté le 27 juillet 2021.

[22] [https://ichtya.unicaen.fr/lab/thesaurus/sommaireVernaculaire.html], consulté le 27 juillet 2021.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

travail, leur très grand nombre dans les œuvres de l'époque moderne et le fait qu'elles ne soient pas directement au centre de nos préoccupations, qui visent avant tout à définir un corpus latin, fait que nous avons distingué en deux ensembles séparés les notices de noms de poissons latins et grecs et les notices de noms vernaculaires. Il faut préciser par ailleurs qu'avec les formes vernaculaires, on se heurte au problème des identifications multiples et parfois fautives, ce qui pouvait faire de nos fiches latines des entités à la fois illisibles et inexploitables, à cause d'un trop grand nombre d'identifications fallacieuses.

Ces deux sommaires se retrouvent sous forme de menu déroulant lors de la lecture d'une notice et permettent de naviguer d'une notice à l'autre sans repasser par la page sommaire. Ces menus déroulants offrent une forme de lecture linéaire du corpus.

Les notices du thesaurus sont également accessibles de manière transversale au moyen d'un moteur de recherche plein texte, qui pourrait être affiné en permettant de rechercher dans des champs précis : formes nominales ; notes bibliographiques – études et sources – ; notes de commentaires ; notes de relations.

Enfin, depuis les notices, on peut consulter celles qui sont en relation : aller d'une notice fille à une notice mère et consulter les notices citées en note de relation.

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.

Le choix qui a été fait de mettre en ligne, en accès libre, le travail en cours sur les textes permet de suivre le thesaurus en cours de constitution. En quatre ans, 2400 notices ont été créées et les ajouts et les corrections sont visibles au fur et à mesure par les chercheurs qui encodent, mais aussi par les lecteurs.

3 Des arbres XML aux graphes

3.1 Principe

Nous avons, en suivant une logique que nous avons présentée dans des travaux antérieurs portant soit sur les sources anciennes (Buard, 2015 et 2019), soit sur la représentation des raisonnements en archéologie[23], exploité la richesse de l'annotation et de la structuration produites pendant les phases d'analyse par les chercheurs pour produire de nouvelles solutions d'accès aux informations scientifiques. L'objectif n'était donc pas, dans un premier temps, de fournir de nouvelles annotations, mais de tirer profit de l'effort déjà réalisé dans ce domaine dans le but de permettre la production

[23] Buard et Zadora-Rio, 2019 ; Marlet *et al.*, 2019. Voir également Elisabeth Zadora-Rio et Henri Galinié « L'église de Rigny et ses abords. De la *colonia* de Saint-Martin de Tours au transfert du centre paroissial (600-1865) », Caen, Presses universitaires de Caen, 2020 ([https://www.unicaen.fr/puc/rigny], consulté le 27 juillet 2021). La revue *Arkeotek Journal* ([http://www.thearkeotekjournal.org/], consulté le 27 juillet 2021), propose également une implémentation de ces principes appliquée à l'archéologie des techniques.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

d'un graphe par poisson, constitué de l'ensemble des informations nécessaires à la compréhension des processus de circulation des savoirs.

Une fois les graphes produits, ils permettent également le contrôle des données enregistrées. En effet, dans la mesure où ils regroupent, sous une forme graphique, un ensemble d'informations dispersées, ils apportent aux chercheurs un puissant moyen de vérification et de relecture. Ces vues peuvent être considérées comme une première exploitation de l'annotation scientifique. Le résultat de cette exploitation permet de critiquer les étapes précédentes sur lesquelles elle s'appuie.

L'arbre XML-TEI, produit au terme du travail de structuration, contient une grande quantité d'informations : formes principale et secondaires, variantes, identification, liens entre poissons, traduction, etc. Chacune de ces informations est spécifiquement annotée par une séquence XML-TEI et, pour certaines d'entre elles, donnent une relation entre une notice et une ou plusieurs autres.

Nous avons ainsi exploité les arbres résultant des opérations d'annotation, et en particulier les liens vers d'autres notices, pour réorganiser l'information scientifique sous forme d'un graphe RDF (*Resource Description Framework*) explicitant les connexions entre les éléments qui le composent en les organisant sous la forme de triplets (sujet, prédicat, objet).

3.2 Processus

Dans un premier temps, il est nécessaire d'identifier les éléments qui, dans les fiches de chaque poisson, doivent être retenus pour alimenter les graphes. Nous avons choisi d'appuyer la génération des graphes sur les éléments suivants :

- la forme principale : <form @type="ppal">;
- les synonymies : <form @type="traduction">;
- les variantes : <form @type="variante">;
- l'identification scientifique : <def>;
- les relations (les renvois d'une notice à une ou plusieurs autres) : <note @subtype="relation">.

Chacun des items de cette liste correspond à un nœud XML-TEI spécifique. La figure 7 donne une représentation graphique du principe que nous avons mis en place en nous appuyant sur l'exemple de l'*Abarenon*[24].

[24] Voir la notice en ligne dans le thesaurus du laboratoire du projet Ichtya : [https://ichtya.unicaen.fr/lab/thesaurus/doc/pddn_pisc.1458914124839.html], consulté le 27 juillet 2021.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

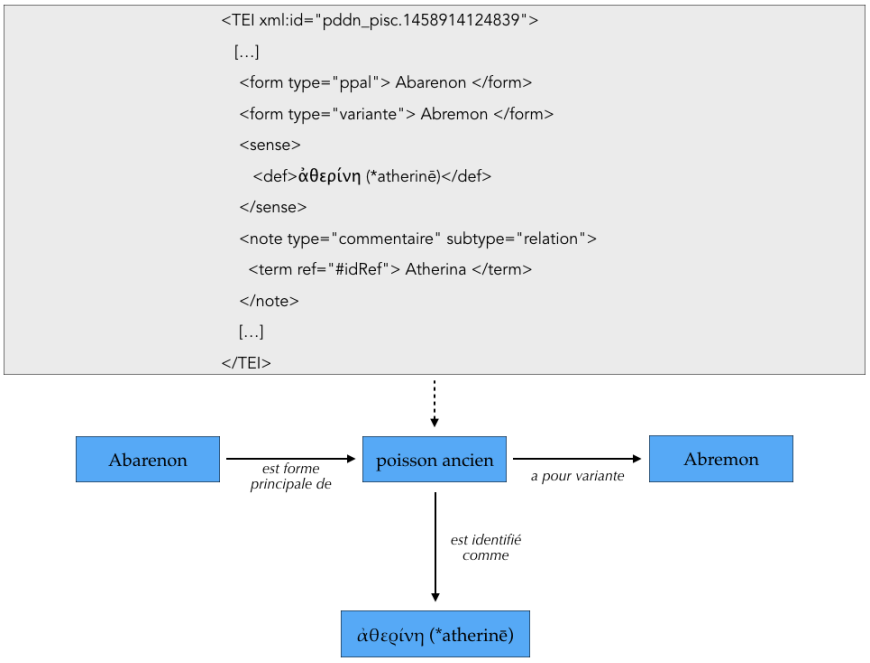


Fig. 7 : Principes du modèle de graphe à partir de l'*Abarenon*

Les notices encodées en XML-TEI contiennent donc l'ensemble des informations concernant les relations, mais c'est l'interprétation de l'arborescence, en incluant bien entendu les couples attributs/valeurs, qui permet d'accéder au sens. Pour expliciter ces relations, graphiquement en l'occurrence, nous avons développé une série de traitements qui interprètent les éléments mobilisés dans les arbres XML et les transforment en triplets explicites. Par exemple la séquence XML suivante, volontairement tronquée ici, donne un exemple de l'encodage d'une note de mise en relation de la notice de l'*Abrenon* avec celle de l'*Atherina* :

```
<sense xml:id="pddn_ident.06152017-093525">
    <def>remonte au nom grec de l'athérine, ἀθερίνη (*atherinē).</def>
    […]
    <note resp="BG" subtype="relation" type="commentaire">Voir<term
    ref="pddn_pisc.1569318239815.xml#pddn_pisc.1569318239815"
    type="animal">Atherina</term>.</note>
</sense>
```

Cette note est interprétée et transformée dans le RDF ; l'individu est ici donné en totalité et la relation « lié à » est en caractères gras :

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

```
<owl:NamedIndividual
rdf:about="https://www.unicaen.fr/ichtya.xml#pddn_pisc.1458914124839">
   <pddn:Position>2332</pddn:Position>
   <rdf:type rdf:resource="https://www.unicaen.fr/ichtya.xml#PoissonAncien"/>
   <pddn:Nom>Abarenon</pddn:Nom>
   <pddn:est_identifie_par
   rdf:resource="https://www.unicaen.fr/ichtya.xml#pm3293"/>
   <pddn:lie_a
   rdf:resource="https://www.unicaen.fr/ichtya.xml#pddn_pisc.1569318239815"/>
   <pddn:lie_a
   rdf:resource="https://www.unicaen.fr/ichtya.xml#pddn_pisc.1459347998358"/>
   <pddn:a_pour_variante
   rdf:resource="https://www.unicaen.fr/ichtya.xml#pddn_pisc.1544690586484"/>
   <pddn:a_pour_variante
   rdf:resource="https://www.unicaen.fr/ichtya.xml#pddn_pisc.1458912766905"/>
   <pddn:a_pour_variante
   rdf:resource="https://www.unicaen.fr/ichtya.xml#pddn_pisc.1542103197225"/>
</owl:NamedIndividual>
```

Dans cet exemple, c'est l'apparition explicite de la qualification qui fait la différence. En définitive, nos outils suivent un principe similaire à celui d'un humain décodant l'arbre : ils interprètent l'encodage de la *note* de *type* « commentaire » et de *subtype* « relation » et la transforment en une relation « lié à » entre l'*Abarenon* et l'*Atherina*. Ce type de relation, formulé en tant que tel, va nous permettre de produire des vues synoptiques des réseaux de notices, et de procéder à des interrogations comme, par exemple, l'extraction de toutes les notices dans lesquelles est exprimée une relation de type « est identifié comme grande vive ».

Du point de vue technique, concernant la production des graphes, nous utilisons les technologies XML, XPATH, XSL, XQUERY, ainsi que le langage DOT[25] pour la production des fichiers SVG intégrés au moteur d'affichage en ligne.

Du point de vue de la procédure, la première étape du traitement consiste à produire un graphe RDF unique à partir de l'ensemble des notices encodées en XML-TEI en utilisant une feuille de transformation XSLT. C'est pendant cette phase qu'une ontologie dédiée est produite. Cette dernière explicite de quoi les graphes sont composés et quels types de relations les individus pourront entretenir. L'exemple RDF donné plus haut présente l'ontologie actuellement utilisée ; elle pourra bien entendu faire l'objet d'évolutions et d'enrichissements en fonction des pistes de recherche à venir. Les deux premières lignes donnent les types d'individus, ici *PoissonAncien* et *PoissonModerne* ; les lignes suivantes renseignent les propriétés (comme le nom du poisson) et les relations associées à ces types d'individus. Certaines de ces relations sont symétriques, comme « lié à ».exprimant la réciprocité : si A est lié à B, alors B est lié à A. D'autres précisent le sens des relations, par exemple : *a pour variante* est la relation inverse de *est variante de*.

[25] DOT est un langage de description de graphe et fait partie de l'ensemble d'outils *open source* graphiz (http://www.graphviz.org).

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

```
<owl:ObjectProperty rdf:about="&pddn;a_pour_variante">
   <rdfs:domain rdf:resource="&pddn;PoissonAncien"/>
   <rdfs:range rdf:resource="&pddn;PoissonAncien"/>
   <owl:inverseOf rdf:resource="&pddn;est_variante_de"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:about="&pddn;est_variante_de">
   <rdfs:domain rdf:resource="&pddn;PoissonAncien"/>
   <rdfs:range rdf:resource="&pddn;PoissonAncien"/>
   <owl:inverseOf rdf:resource="&pddn;a_pour_variante"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:about="&pddn;a_pour_traduction">
   <rdfs:domain rdf:resource="&pddn;PoissonAncien"/>
   <rdfs:range rdf:resource="&pddn;PoissonAncien"/>
   <owl:inverseOf rdf:resource="&pddn;est_traduction_de"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:about="&pddn;est_traduction_de">
   <rdfs:domain rdf:resource="&pddn;PoissonAncien"/>
   <rdfs:range rdf:resource="&pddn;PoissonAncien"/>
   <owl:inverseOf rdf:resource="&pddn;a_pour_traduction"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:about="&pddn;identifie">
   <rdfs:domain rdf:resource="&pddn;PoissonModerne"/>
   <rdfs:range rdf:resource="&pddn;PoissonAncien"/>
   <owl:inverseOf rdf:resource="&pddn;est_identifie_par"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:about="&pddn;est_identifie_par">
   <rdfs:domain rdf:resource="&pddn;PoissonAncien"/>
   <rdfs:range rdf:resource="&pddn;PoissonModerne"/>
   <owl:inverseOf rdf:resource="&pddn;identifie"/>
</owl:ObjectProperty>
<owl:SymmetricProperty rdf:about="&pddn;lie_a">
   <rdfs:domain rdf:resource="&pddn;PoissonAncien"/>
   <rdfs:range rdf:resource="&pddn;PoissonAncien"/>
</owl:SymmetricProperty>
<owl:DatatypeProperty rdf:about="&pddn;Nom">
   <rdfs:domain rdf:resource="&pddn;PoissonAncien"/>
</owl:DatatypeProperty>
<owl:DatatypeProperty rdf:about="&pddn;Nom">
   <rdfs:domain rdf:resource="&pddn;PoissonModerne"/>
</owl:DatatypeProperty>
<owl:DatatypeProperty rdf:about="&pddn;Position">
   <rdfs:domain rdf:resource="&pddn;PoissonAncien"/>
</owl:DatatypeProperty>
<owl:DatatypeProperty rdf:about="&pddn;Position">
   <rdfs:domain rdf:resource="&pddn;PoissonModerne"/>
</owl:DatatypeProperty>
```

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

Une fois l'ontologie produite, chaque poisson faisant l'objet d'une notice est ajouté au graphe en intégrant les éléments retenus (formes principale et variantes, identifications scientifiques, relations, etc.). La constitution de ce réseau est rendue possible par l'application d'une logique rigoureuse d'identification dès les phases d'annotation et de structuration en XML – avec une exception sur laquelle nous devons nous arrêter.

La création des poissons anciens ne pose pas problème : il s'agit en quelque sorte de l'objet principal des notices, il suffit donc de s'appuyer sur leur identifiant pour les projeter dans les graphes. En revanche, dans l'état actuel de la structuration XML mis en œuvre, l'identification scientifique est un peu plus délicate. En effet, les identifications scientifiques sont structurées à l'intérieur de l'élément <sense> dans un élément <def> sans plus de caractérisation ou autre système de normalisation. Autrement dit, les identifications ne font l'objet d'aucune contrainte de saisie et c'est uniquement en s'appuyant sur la rigueur des chercheurs que nous pouvons créer des poissons modernes sur la base de regroupements des éléments <def>. En réalité, c'est en s'appuyant sur des chaînes de caractères identiques dans les éléments <def> que notre système fonctionne. Si les résultats obtenus sont satisfaisants, on ne peut considérer que cette méthode le soit. Une solution plus pertinente et rigoureuse consisterait à mettre en place un système exploitant des nomenclatures binomiales des identifiants uniques à l'échelle de l'ensemble du thesaurus. En plus de gagner en organisation documentaire, les coûts de calculs de générations des graphes seraient améliorés.

Notons enfin que c'est au moment de l'exploitation des données pour la fabrication des graphes que la solution de structuration des identifications scientifiques nous est apparue comme un problème : la consultation des fiches et la compréhension des informations n'étaient en rien handicapées. On peut donc considérer, comme nous l'avons évoqué plus haut, que c'est la production des graphes qui nous permet de mettre nos solutions de structuration et notre modèle de données à l'épreuve et, en définitive, de l'améliorer.

Une fois le graphe général produit, le réseau de chaque poisson ancien est extrait en appliquant une requête XQUERY dans le logiciel de base de données native XML BaseX[26]. Il est important de noter que, pendant la fabrication des graphes, on intègre toutes les relations, d'un poisson à un autre. Ainsi, les graphes fournissent l'ensemble des informations sur un réseau de notices de poissons ; ils permettent de mesurer la densité des relations entretenues par les fiches composant le thesaurus à travers une vue synoptique complète. Autrement dit, la consultation ne se fait plus seulement à l'échelle de la notice d'origine, même si elle est graphiquement signalée dans le graphe par une couleur dédiée, mais bien au niveau du réseau incluant cette notice de base.

La figure 8 donne l'exemple du graphe associé à la notice de l'*Abarenon*. Celui-ci apparaît en jaune dans le graphe, mais c'est bien l'ensemble du réseau auquel il appartient qui est représenté. Le lecteur peut ainsi circuler dans le corpus en suivant l'ensemble des relations, d'un poisson à un autre, au sein d'un réseau donné.

[26] [https://basex.org/], consulté le 27 juillet 2021.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

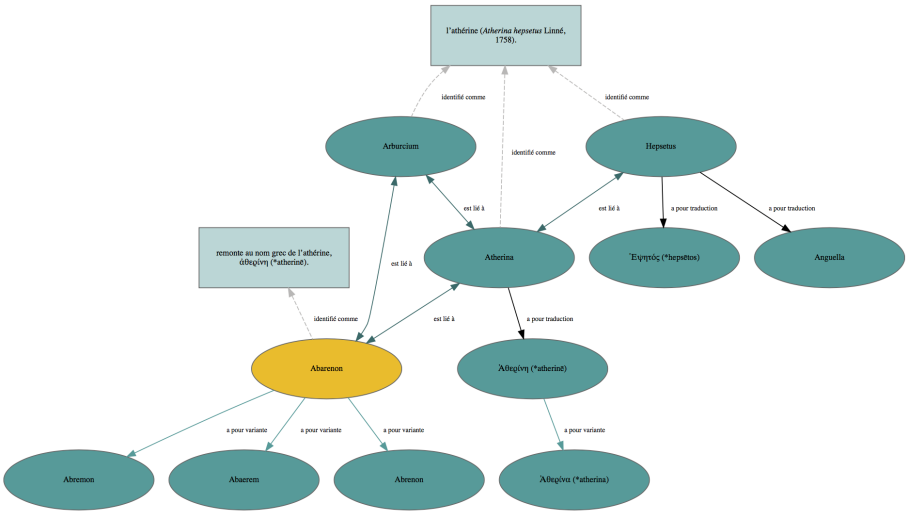


Fig. 8 : Graphe de l'*Abarenon*

Nos outils ont permis d'enrichir le thesaurus en ligne avec les réseaux dans lesquels les poissons s'intègrent. Cette fonctionnalité offre au lecteur de nouvelles possibilités de circulation dans le corpus en passant d'une notice à une autre au sein du même réseau. Cependant, il ne s'agit là que d'une première exploitation, que nous souhaitons compléter par d'autres.

En effet, les graphes sont, pour l'instant, accessibles uniquement via les notices. Or, en ajoutant une propriété basée sur l'apparition dans les œuvres, information dont nous disposons dans la bibliothèque de textes, nous pourrions produire un graphe par œuvre, regroupant ainsi l'ensemble des informations scientifiques sur les poissons qui sont, pour le moment, dispersées et sans solution d'interrogation unifiée. De plus, l'intégration des sources textuelles dans la chaîne de traitement permettrait également d'enrichir les notices de toutes les occurrences des poissons trouvées dans les œuvres. Cet enrichissement fournirait au lecteur une solution pour revenir aux sources directement depuis le thesaurus.

Par ailleurs, la relation « lié à » actuellement utilisée est, reconnaissons-le, relativement générique. Une caractérisation plus précise de cette relation offrirait de riches possibilités d'exploitation tant du point de vue ergonomique (les graphes seraient plus précis), que du point de vue de la recherche dans le corpus, en permettant l'interrogation des réseaux sur des types de relations spécifiques. L'exemple du traitement des synonymies donne ici une bonne idée de ce qui pourrait être mis en place pour la qualification des relations. Les synonymies sont encodées dans les notices XML-TEI avec des éléments <form>, contenant la traduction proposée par les encyclopédistes et caractérisés par l'attribut @type="traduction", la langue étant renseignée dans l'attribut @xml :lang. On obtient donc des séquences de ce type :

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

<form type="traduction"xml :lang="it">Pesce columbo</form>

Cet encodage est ensuite exploité pour la production des graphes selon les principes présentés plus haut pour construire un triplet de la forme « Orbis, a pour traduction, Pesce columbo ».

L'utilisation d'une annotation TEI similaire dans sa construction, avec par exemple l'élément <note subtype="relation">, permettrait de gagner en précision dans l'encodage des notices et, à terme, dans l'expressivité des graphes.

Cependant, aussi prometteuse que soit cette piste, elle implique un très gros travail d'enrichissement sur l'ensemble des notices ; il s'agirait donc d'outiller ce travail pour éviter un important travail « manuel » qui se ferait alors au détriment de la création de nouvelles notices.

En conclusion, grâce à l'équipe pluridisciplinaire réunie dans le programme de recherche Ichtya – historiens des textes, spécialistes de sources anciennes, informaticiens – et à une organisation collective du travail, nous avons mis en place des collaborations particulièrement fructueuses combinant des connaissances et compétences qui, isolées, ne permettaient à aucun des membres de parvenir seul au même résultat. C'est la mise en place de l'objet d'étude numérique commun qui a permis d'enrichir le projet. En effet, l'effort d'objectivation nécessaire à la numérisation a constitué un socle partagé auquel chacun était en mesure d'apporter une pierre. Les interventions des uns étant immédiatement observables par les autres, nous pouvions, au fur et à mesure des avancées du projet, critiquer la construction commune. Nous avons mis en place un « laboratoire de textes », c'est-à-dire un écosystème permettant de travailler collectivement sur la bibliothèque Ichtya et son thesaurus : les fichiers XML sont hébergés dans une base de données XML distante (serveur de l'université de Caen) ; ils sont modifiés par les chercheurs au moyen du plugin collaboratif PluCo dans l'interface de travail habituelle XMLmind XML Editor. Les données XML-TEI élaborées par les chercheurs sont directement visibles sur le portail Ichtyalab qui rassemble le site de la bibliothèque Ichtya et le thesaurus des poissons et créatures aquatiques. Dans cette logique, le développement du laboratoire de textes est, de notre point de vue, un atout considérable, car c'est à travers le portail que nous pouvons expérimenter nos idées, valider nos hypothèses de travail et que chacun peut mesurer l'apport des autres.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352