humanités numériques

Humanités numériques

2 | 2020
Varia

Une chaîne d'extraction pour l'enrichissement de bases de données archéologiques

An Information Extraction Framework to Enrich Archeological Databases

Frédérique Mélanie-Becquet, Johan Ferguth, Michel Cartereau, Katherine Gruel et Thierry Poibeau

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/revuehn/471
DOI : 10.4000/revuehn.471
ISSN : 2736-2337

Éditeur
Humanistica

Référence électronique
Frédérique Mélanie-Becquet, Johan Ferguth, Michel Cartereau, Katherine Gruel et Thierry Poibeau, « Une chaîne d'extraction pour l'enrichissement de bases de données archéologiques », *Humanités numériques* [En ligne], 2 | 2020, mis en ligne le 01 juin 2020, consulté le 08 septembre 2020. URL : http://journals.openedition.org/revuehn/471 ; DOI : https://doi.org/10.4000/revuehn.471

humanités
numériques

2 | 2020
Varia

Une chaîne d'extraction pour l'enrichissement de bases de données archéologiques

An Information Extraction Framework to Enrich Archeological Databases

Frédérique Mélanie-Becquet, Johan Ferguth, Michel Cartereau, Katherine Gruel et Thierry Poibeau

Résumés

Cet article décrit une expérience visant à extraire des informations textuelles pour alimenter automatiquement des bases de données dans le domaine de l'archéologie. Les premières expériences ont porté sur les *Cartes archéologiques de la Gaule* (*CAG*). Elles ont permis d'observer des problèmes de transfert d'information et d'évolution des thésaurus, un même terme ne désignant pas toujours la même notion au cours du temps, ou un site archéologique pouvant avoir été catégorisé de différentes manières au cours du temps.

This article focuses on an experiment aimed at extracting information from text in order to automatically feed databases in the field of archaeology. The first experiments concerned a set of books: the *Cartes archéologiques de la Gaule* (*CAG*). Knowledge transfer and meaning evolution phenomena were observed when thesauri were examined, since the same term can refer to different notions, and the same archaeological site may be categorised differently, at different points in time.

Entrées d'index

MOTS-CLÉS : informatique, archéologie, base de données, chaîne de traitement, extraction d'informations, numérisation, ontologie, terminologie, thésaurus

KEYWORDS: computer science, archaeology, database, processing chain, information extraction, digitisation, ontology, terminology, thesaurus

Introduction

1 Les unités de recherche en archéologie gèrent souvent des archives volumineuses, constituées de documents hétéroclites : rapports de terrain, thèses, catalogues et ouvrages de référence. Ces archives représentent des décennies de travail, mais il est généralement difficile de les consulter et d'y accéder dans la mesure où il s'agit de documents fragiles, parfois uniques et stockés dans des laboratoires variés. Dans ce contexte, leur numérisation en texte intégral n'est pas suffisante ou, en tout cas, la numérisation est une bonne opportunité pour enrichir les documents d'index et de thésaurus structurés : pour cela, il est nécessaire de formaliser et structurer ces documents ainsi que les informations qu'ils contiennent, puis d'en faciliter l'accès grâce à des interfaces conviviales.

2 Cet article est une présentation du projet *EITAB*[1], qui vise à mettre en place une chaîne de traitement dont la finalité est l'enrichissement de bases de données à partir de documents numérisés. Pour ce faire, il a été nécessaire, dans un premier temps, de procéder à de nombreux échanges entre informaticiens, linguistes et archéologues. Cette étape a permis d'appréhender les données, d'une part les textes à scanner dont il faut extraire le lexique pertinent pour la tâche, et d'autre part les bases de données informatisées existantes, que le projet a pour finalité d'enrichir. Il en résulte la constitution d'un thésaurus qui servira de référence dans la chaîne de traitement. Les différents échanges entre acteurs du projet ont aussi montré la nécessité de créer une interface graphique permettant une utilisation optimale de l'application par l'expert, ainsi que la nécessité de standardiser les formats pour garantir une meilleure portabilité.

3 Il faut noter que cet article, pour l'essentiel, ne présente pas une nouvelle application finalisée, mais plutôt le travail préparatoire pour y parvenir. Il serait évidemment plus facile de valoriser une application finie et conviviale, avec différentes fonctionnalités répondant à des besoins précis, mais nous pensons que le travail préparatoire, plus ingrat, sur les données, les termes, leur signification et leurs problèmes inhérents (ambiguïté, polysémie, changement de sens au cours du temps, etc.) sont aussi importants et assez peu mis en avant dans le domaine.

4 Nous décrivons dans un premier temps les sources du projet, à savoir les systèmes d'information et les corpus utilisés afin d'enrichir les bases de données visées. Dans un second temps, nous abordons la constitution du lexique : comment nous avons sélectionné automatiquement les termes dans le corpus d'enrichissement, comment nous avons procédé pour structurer manuellement ces termes, quel format standard nous avons décidé d'utiliser afin de structurer notre thésaurus. Enfin, nous décrivons la plateforme mise en place, les extractions et manipulations qu'elle permet, les modifications et améliorations qui restent à développer.

À l'origine, le PEPS CNRS-PSL *EITAB*

5 La coopération entre archéologues, linguistes et informaticiens vise à concevoir et à valider un traitement automatisé des corpus, de façon à réduire le temps d'intervention, à accroître la fiabilité des résultats et à faciliter le partage des données dans un contexte interdisciplinaire (traitements statistiques, historiques, thématiques) : autrement dit, améliorer l'environnement d'étude et de recherche de l'archéologue. Il s'agit de mettre en correspondance semi-automatiquement des informations extraites de textes avec des « champs » d'une base de données.

Les sources brutes

6 Les documents dont disposent les archéologues sont nombreux. L'étude pilote dont il est question dans cet article a été menée à partir de deux types de documents, fortement structurés ou non. Les documents structurés nécessitent peu de traitement : il est aisé d'en extraire les informations essentielles par un jeu de simples « scripts » informatiques. Ce sont des fiches, des notices, des documents où l'information est classée, facile à repérer. Inversement, les documents peu structurés sont plus complexes à analyser. Notre propos est ici de montrer qu'il est possible d'analyser ce type de documents, d'utiliser des documents bruts en vue de l'enrichissement de données structurées. Si dans cet article nous parlons principalement du second, les deux types de documents ont cependant servi à tester la robustesse de la méthode.

7 Lors de notre projet, nous avons utilisé une collection de textes couvrant des fouilles liées à la période gauloise, les *Cartes archéologiques de la Gaule* (*CAG*) (Coulon *et al.* 1992 ; Provost *et al.* 1992). La zone englobe une grande partie de la France actuelle, de l'âge du fer à la période médiévale. 128 volumes ont été publiés jusqu'à présent. Chaque volume correspond à un département français, et certains départements sont couverts par plusieurs volumes. L'étude pilote concernait trois de ces volumes, correspondant aux départements de l'Indre et du Cher. Ces textes retracent la plupart des découvertes archéologiques recensées dans les zones et les époques concernées.

8 L'idée est bien sûr de numériser ces documents et de permettre leur mise en ligne, mais aussi (et surtout) d'en extraire les informations clés afin d'alimenter les différents « champs » des bases de données. L'exemple de la figure 1 est un texte restituant les découvertes faites à Moulins-sur-Yèvre, commune du Cher. Il est fait mention de trois sites archéologiques : le *camp de Chou*, le *camp de Maubranches* et le *camp de Vercingétorix*.

FIGURE 1. STRUCTURATION DES *CARTES ARCHÉOLOGIQUES DE LA GAULE* (*CAG*)

030 - Moulins-sur-Yèvre (I.N.S.E.E. n° 158)

La commune est traversée par la voie antique : M.-E. Fonvielle, J. Déguéret, dans *Cahiers du Berry*, 60, 1980, p. 21-26.

À *Maubranches*, en 1621, Étienne de Clavière dit qu'on a trouvé un autel élevé à Mars et à Auguste : *Flavia Cuba/ Firmani filia/ Cososo deo Marti suo/ hoc signum donavit/ Augusto : Commission Historique du Cher*, 1854, p.15 ; - A. Buhot de Kersers, *Statistique monumentale*, I, 1875, p. 249 ; - G. Thaumas de la Thaumassière, *Histoire de Berry et du diocèse de Bourges*, Bourges, 1689, p. 567 et p. 574 ; - N. Catherinot, *Antiquités romaines du Berry*, 1682, p. 7 ; - *C.I.L.*, XIII, 1899, n° 1353 ; - É. Chénon, dans *B.S.A.F.*, 21.07.1915, p. 230-238 (avec bibliographie complète) ; - C.-Ch. Pierquin de Gembloux, *Notices*, 1840, p. 142.

À *Maubranches*, au-dessus de la chapelle, dans les travaux du chemin de fer de Bourges à Sancerre, mise au jour d'une portion d'aqueduc romain qui mesure 1,70 m de large au fond de la cuvette. Le radier est à 18 ou 20 cm au-dessus du niveau de l'*Ouatier*. Ce pourrait être le même qui fut constaté en 1870, à la Pyrotechnie. A. Buhot de Kersers pense qu'il a pu capter les eaux de la fontaine de *Valentigny* : L. de Raynal, *Histoire du Berry*, I, 1845-1847, p. 81; - A. Buhot de Kersers, dans *Mém. Soc. Ant. Centre*, I, 1867, p. 21 et *Statistique monumentale*, I, 1875, p. 249 ; - *Bull. Soc. Ant. Centre*, 1.07.1891, Arch. Dép. Cher, 2F 586, p. 97-98 ; - A. Lefort, A. Buhot de Kersers, dans *Mém. Soc. Ant. Centre*, XX, 1893-1894, p. 23-28 ; - É. Chénon, dans *B.S.A.F.*, 1915, p. 230-238.

Au lieu-dit *Camp de Chou* autrement *Maubranches*, *Camp des Monts*, *Camp de César*, à 600 m au sud de la R.N. 151, au confluent de l'*Ouatier* et de la *Tripaude*, un éperon barré (surface : 14 à 17 ha) avec remparts, fossés de 11 m de large et *vallum*. Ce site n'est pas précisément daté : A. Frémont, *Le département du Cher*, 1862, p. 211 ; - De Rouvre, dans *Revue du Berry*, 1864, p. 83-84 ; - A. Buhot de Kersers, dans *Mém. Soc. Ant. Centre*, I, 1867, p. 13-57 et II, 1868, p. 21-25 ; A. Buhot de Kersers, *Statistique monumentale*, I, 1875, p. 248-249, pl. VII, fig. 4 et pl. VI, fig. 4 ; - De Mortillet, dans *L'homme préhistorique*, 1906, p. 193-206 ; - P. Dubois de la Sablonnière, dans *Mém. S.H. Cher*, XL, 1933, p. 1-27 ; - O. Buchsenschutz, *Les oppida*, 1968, p. 53 ; - J. Allain, dans *Cahiers du Berry*, 16, 1969, p. 43-44 ("*vallum* doublé d'un large fossé en auge") ; - J. Holmgren, site 298 (photo du 10.04.1977). "Tout autour, sont disséminés de petits mondrains dans lesquels on a trouvé des armes en fer et des (monnaies) romaines" : L. Martinet, *Le Berry préhistorique*, 1878, p. 98.

Sur le site de *Maubranches*, Ferrand de Saligny a fouillé, en 1783, un tumulus dans lequel il a trouvé un "fer de lance" et "une poignée de sabre" sous un "amas de pierres". Il signale une légende de "monnaies d'or" : J.-G. Ferrand de Saligny, dans *Mém. Soc. Ant. Centre*, III, 1869, p. 38. Pierquin de Gembloux dit, qu'en janvier 1687, on aurait trouvé l'inscription : *Solimarak sacrum aedem cum suis ornamentis firmana cobrici mater d.s.d. (sic)* : C.-Ch. Pierquin de Gembloux, *Notices*, 1840, p. 452-456.

Dans le champ de *Maubranches*, sur la route de Sainte-Solange à la R.N. 151, en avril 1884, P. de Goy a fouillé un cimetière de La Tène, composé de 4 fosses orientées est-ouest, contenant des sépultures avec des armes de fer. 1ère sépulture : inhumation avec pointe et talon de lance, des chaînes de suspension d'épée, une épée en fer, un umbo, des fragments de bouclier, une fibule en fer dans les côtes, un bracelet de bronze au bras, un anneau à bossettes à l'épaule, un tesson de céramique noire ; 2e sépulture : squelette avec une épée et son fourreau, à droite du corps, une lance, 3 anneaux de suspension de l'épée, un bracelet en lignite ; 3e sépulture: 2 squelettes avec bracelets ; (4e sépulture ?) : P. de Goy, dans *Mém. Soc. Ant. Centre*, XIII, 1885, p. 97-108, pl. II ; - A. Buhot de Kersers, dans *B.A.C.T.H.*, 1885, p. 222 ; - *Bull. Soc. Ant. Centre*, 2.07.1919, XIII, p. 97 (don à la société des fouilles de Maubranches) ; - *Bull. Soc. Ant. Centre*, 5.11.1919, (dépôt au Musée de Bourges) ; - J. Déchelette, *Manuel*, II, 1910, appendice V, p. 22-23 - app. VI, n° 153-154, p. 72-73 et n° 76-77, p. 136 ; - M. Willaume, *Le Berry*, 1985, p. 99-103 ; - Musée du Berry : 950.1.423 (fibule) et 424 (céramique) - 950.1.427 (bracelet en lignite) et 428 (anneau en fer, LaTène I) - 950.1.426 (24 fragments de fer) - 950.1.425 (2 pièces en forme d'étrier, La Tène) - 950.1.422 (bracelet en bronze, La Tène) - 950.1.421 (anneaux en fer, La Tène II) - 950.1.205 (bracelet ovale en fer) - 950.1.201 à 204 (anneaux en bronze, La Tène II) - 950.1.200 et 199 (bracelets en bronze, La Tène) - 950.1.196 et 197 (fourreaux d'épées en fer) - 950.1.196 (épée en fer) - 950.1.195 (garniture d'épée ou de bouclier) - 950.1.194 (fer de lance) - 950.1.193 (10 fragments de bouclier) - 950.1.191 (fer et talon de lance) - 950.1.190 (umbo de bouclier) - 950.1.189 et 187 (épées à soie avec leur fourreau) - 950.1.186 (chaîne pour suspendre une épée) - 966.16.1 (vase ovoïde en terre).

Non précisé, un as d'Alexandre Sévère (R.I.C. 554) : A. Cothenet, dans *Cahiers du Berry*, 24-25, 1971, p. 122.

Extrait d'une version numérisée de *Carte archéologique de la Gaule. 18. Le Cher*, par Michel Provost, Jean-François Chevrot et Jacques Troadec

AOROC

9 L'ensemble des *CAG* est structuré de façon identique. Une *CAG* liste les communes d'un département. Chaque commune est un bloc textuel qui recense les découvertes archéologiques. Chaque bloc commence par un chiffre suivi d'un tiret puis du nom de la commune et enfin, entre parenthèses, son numéro INSEE. Cette information peut être modélisée au moyen d'une expression régulière simple, qui permet d'établir une relation entre un lieu et les découvertes archéologiques réalisées.

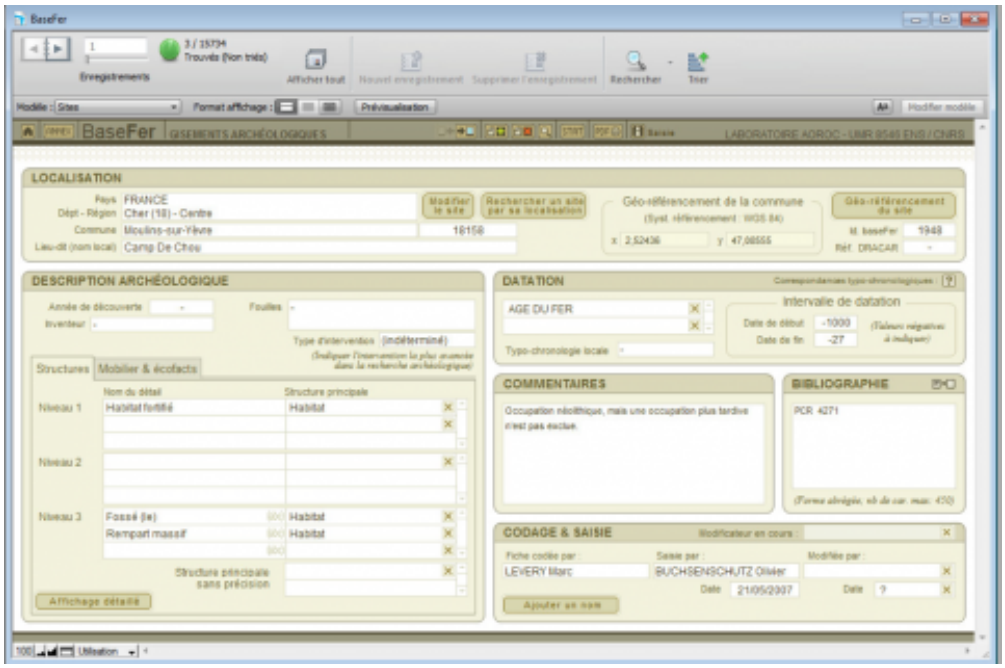
Les sources à enrichir

10 La BaseFer (Buchenschutz *et al.* 2015) répertorie les gisements de l'âge du Fer, par commune, en indiquant les types de structures et les principaux mobiliers associés. Elle donne une image globale des connaissances sur l'âge du Fer. Elle n'est pas exhaustive. Le projet a été pensé pour enrichir cette base de données.

11 Les tables de la BaseFer permettent de décrire les éléments ou objets découverts lors de fouille, site par site. Sur la commune de Moulins-sur-Yèvres (figure 2) sont recensés trois sites archéologiques, ce qui correspond à trois fiches dans la base (comme l'indique le compteur *3/15734*).

L'une d'elles recense les découvertes archéologiques du site *Champ de Chou*. Les champs permettent de lister les *structures* ou *mobiliers et ecofacts* de chacun des sites. Ainsi le *fossé* est un type de *structure*, parmi les *habitats*. En cliquant sur l'onglet *mobiliers et ecofacts* de ce formulaire, l'utilisateur accède au recensement de la découverte du *lignite* qui correspond à un type de *mobilier*, parmi les mobiliers de type *parure*.

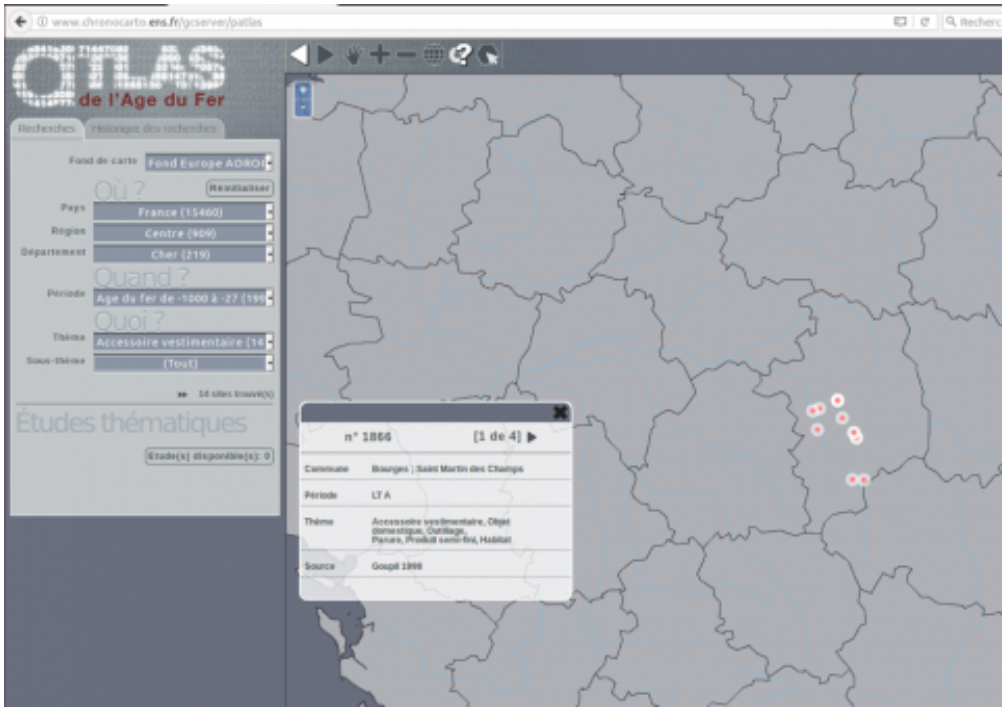
FIGURE 2. LA BASEFER



Base Fer, AOROC, https://www.chronocarto.eu/spip.php?article8&lang=fr

12 La BaseFer alimente un atlas disponible en ligne : l'atlas de l'âge du Fer. L'atlas permet de répondre aux questions « où ? quoi ? quand ? » (figure 3) [2].

FIGURE 3. ATLAS DE LA BASEFER



Base Fer, AOROC, https://www.chronocarto.eu/spip.php?article8&lang=fr

13 L'ensemble des participants du projet a étudié la BaseFer afin d'appréhender son modèle de données. Les experts du domaine – les archéologues – ont expliqué les éléments présents dans la base. Les linguistes informaticiens ont appréhendé *a minima* le lexique afin de comprendre le domaine d'application. Jamais cependant ces derniers n'ont prétendu en saisir toute la finesse et n'ont pu acquérir la sensibilité linguistique optimale nécessaire à la structuration du lexique. En aucun cas, ils ne se sont substitués à l'expert. Sur toute la durée du projet, un perpétuel échange entre archéologues et linguistes informaticiens a été nécessaire.

14 En effet, il est difficile pour une néophyte de savoir ce qu'est une *fibule* (sorte d'agrafe), une *armille* (bracelet de bronze pour le bras), ou encore un *artefact* (objet résultant de l'activité humaine). À cela s'ajoute le problème de la polysémie lexicale, un même terme peut référer à différents objets. Par exemple, la *fosse* peut être un *dépôt*, un *cimetière* ou encore une *structure en creux*. Un même terme peut évoluer au fil des années, des siècles et ne plus désigner la même chose. Ainsi quand un archéologue rédige un article, un carnet de terrain, il emploie toutes les richesses de la langue. C'est pourquoi, lors de la constitution d'un thésaurus, il revient à l'expert de déterminer la signification exacte d'un terme en fonction du contexte.

15 À partir de cette première analyse des sources, l'idée était d'extraire automatiquement l'ensemble des termes des *CAG*, de créer une liste à plat de termes, c'est-à-dire une simple liste de termes, non hiérarchisée, non structurée. Il est important que cette liste soit large, ne soit pas trop restreinte. L'outil ne doit pas écarter des termes potentiellement importants, significatifs pour le domaine. Mieux vaut une présélection large que l'utilisateur peut à souhait consulter, retravailler, restreindre.

Extraire et hiérarchiser les termes

16 Le projet n'avait pas pour ambition de développer de nouveaux outils. Le travail a essentiellement consisté à choisir parmi les outils existants ceux qui étaient le mieux adaptés à la situation, puis à les paramétrer et les lier entre eux pour constituer une chaîne de traitement adaptée (Lamani 2013).

Extraire les termes

17 Le premier objectif était d'extraire les termes des *CAG*[3]. Nous avons utilisé YaTeA, outil d'extraction permettant de récupérer l'ensemble des termes du domaine. YaTeA a été développé au LIPN par Thierry Hamon et Sophie Aubin pour aider le processus d'identification des termes en corpus.

> YaTeA (*Yet Another Term ExtrActor*) est un extracteur de termes. Il identifie et extrait des groupes nominaux pouvant être des termes, c'est-à-dire des termes candidats. Chaque terme candidat est analysé syntaxiquement pour faire apparaître sa structure sous la forme de têtes et modifieurs [4].

18 YaTeA offre la possibilité de redéfinir ou de modifier le processus d'extraction à travers plusieurs fichiers de configuration. On peut ainsi préciser la langue, les portions de texte qu'il faut ou non inclure (inclu-

sion ou non des titres pour les traitements, par exemple), le type de termes à considérer (termes simples – monogramme – ou complexes), si le logiciel doit aussi extraire les termes simples inclus dans des termes complexes, etc. (figure 4).

FIGURE 4. FICHIER DE CONFIGURATION DE YATEA

```
<DefaultConfig>
CONFIG_DIR = /usr/share/YaTeA/config
RESULT_DIR = \$PWD/RESULTS
LOCALE_DIR = /usr/share/YaTeA/locale
</DefaultConfig>
<OPTIONS>
language = FR
MESSAGE_DISPLAY = FR
TC-for-BioLG = 0
TT-for-BioLG = 0
TTG-style-term-candidates = multi
XML-corpus-for-BioLG = 1
annotate-only = 0
debug = 0
match-type =
monolexical-all = 0
monolexical-included = 0
printChunking = 1
suffix = default
XML-corpus-raw = 1
termList =
xmlout = 1
</OPTIONS>
```

YaTeA, Thierry Hamon et Sophie Aubin, https://perso.limsi.fr/hamon/YaTeA

19 YaTeA est un module Perl. C'est un logiciel libre disponible en ligne. Son utilisation s'effectue en ligne de commande. Il prend en entrée un texte étiqueté morphosyntaxiquement à l'aide de TreeTagger (Schmid 1997). Il n'a pas d'interface graphique, et est de ce point de vue difficile à utiliser pour un non-informaticien.

20 Les informations sont extraites du corpus sous forme de listes de termes, appelés *termes candidats*. Chaque *terme candidat* de la liste est constitué d'un mot ou groupe de mots.

21 En sortie, la liste des termes est importante, elle contient beaucoup de bruit, c'est-à-dire beaucoup de *termes candidats* qui ne sont pas pertinents pour la tâche. À titre d'exemple, la *CAG* du Cher est un document de 132 462 mots. YaTeA extrait de ce texte 14 000 *termes candidats*. La liste des *termes candidats* contient des monogrammes et des multigrammes, elle est ainsi composée de plus de 3 500 mots (chaque mot pouvant apparaître plusieurs fois dans la liste des termes extraits, d'où le fait qu'il y ait beaucoup moins de mots différents que de termes extraits).

22 C'est principalement pour effectuer des sélections dans cette liste de termes candidats que la plateforme développée est d'une aide précieuse pour l'expert, comme nous le verrons ultérieurement.

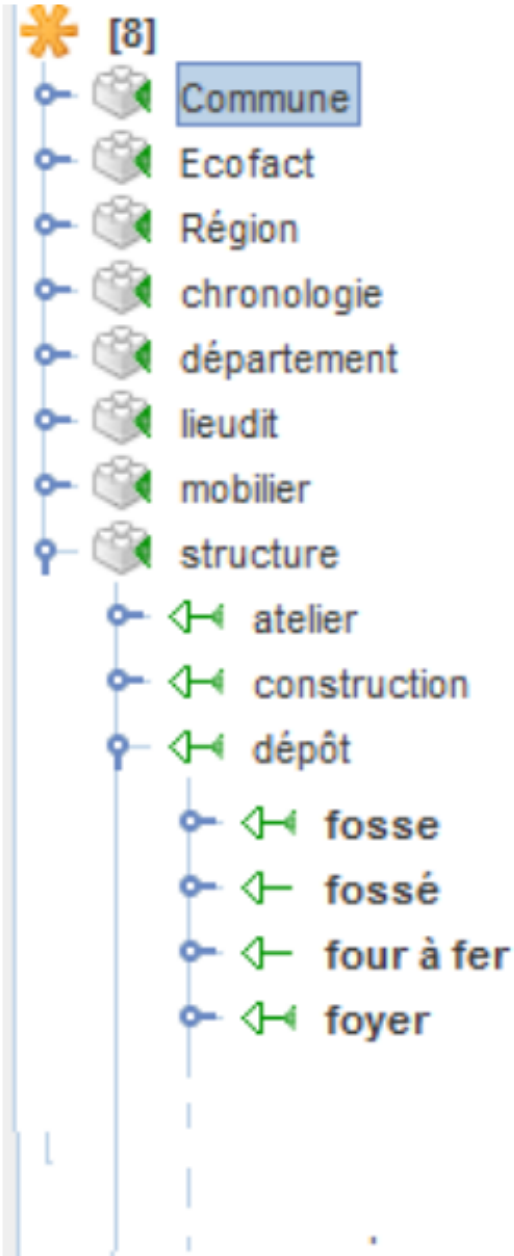
Structurer les termes

23 La liste obtenue en sortie de YaTeA est une liste plate, un fichier texte. Il s'agit maintenant non seulement de la nettoyer – enlever le bruit – mais aussi et surtout de la structurer. C'est sur ce second point, complexe, qu'un premier travail est mené.

24 Dans la liste obtenue, il est plus ou moins facile de repérer différentes variantes d'un même terme. Il peut s'agir d'une simple variante graphique (*âge du fer, âge du Fer, Age du Fer*), d'une différence de nombre (*amphore républicaine, amphores républicaines*), d'écriture en chiffres ou lettres (*premier âge du Fer, 1er âge du Fer*), d'écriture alternative (*clef, clé*) ou encore de différence terminologique (*cuillère, cuilleron*). Mais comment structurer ces variations ?

25 Afin de structurer le vocabulaire, l'ensemble des termes extraits, nous avons utilisé un éditeur initialement développé par l'INRA. TyDI (*Terminology Design Interface*) est un outil collaboratif pour la validation et la structuration de termes en ontologie (figure 5). Il est intégré à AlvisAE (*Alvis Annotation Editor* [Papazian *et al.* 2012]), plateforme d'annotation en ligne. Cette plateforme permet de visualiser et d'annoter les entités, de lier et de typer les entités. TyDI permet la gestion de campagnes d'annotation. C'est un éditeur complexe, spécifique au traitement de données biologiques.

FIGURE 5. TYDI : UN ÉDITEUR D'ONTOLOGIE



TyDI, INRA, http://faculty.washington.edu/fxia/LAWVI/proceedings/cdrom/pdf/LAWVI21.pdf

26 Comme tout éditeur, TyDI permet de faciliter la classification des nombreuses données fournies en entrée. Ce travail aurait pu être fait avec un autre éditeur d'ontologie, mais TyDI (contrairement à d'autres éditeurs) permet de voir le terme en contexte, ce qui aide au travail de classification et au nettoyage de la liste. En observant parallèlement le texte et l'ontologie en devenir (c'est-à-dire en cours de construction), l'expert parvient plus aisément à structurer sa terminologie et son domaine. La contextualisation est un avantage non négligeable pour ce type de tâches.

27 La structuration de la BaseFer (les champs et les tables de la figure 2) est reprise pour constituer une première ontologie des termes extraits, dégager des concepts. Comme l'illustre la figure 2, les concepts tels que *ecofact*, *mobilier* ou encore *structure* sont les parents des termes extraits du corpus tels que *fosse*, *dépôt*.

28 OBO (*Open Biomedical Ontologies*) est un format de la famille XML à l'origine utilisé en biomédecine, comme son nom l'indique, mais dont les propriétés ont *a priori* semblé intéressantes au-delà de ce domaine scientifique (figure 6). Il permet d'organiser les concepts biologiques entre eux à la fois de manière hiérarchique et par synonymie, et de typer les relations. Les concepts sont organisés selon qu'ils sont père ou fils. Ainsi le terme *fosse* est le fils du concept *structure*. Cette information est portée par la balise <is_a>. Du côté du père, la valeur de la balise est *father*. Du côté du fils, la balise reprend l'identifiant du père ainsi que le concept lui-même [5].

FIGURE 6. LE FORMAT OBO

```
<?xml version="1.0" encoding="UTF-8"?>
<racine>
...
<name>structure</name>
<id>0000743</id>
<is_a>father</is_a>
<xref>xref n°1 TyDI:1363507 xref n°2 TyDI_semClass:1613727 </xref>
<exact_synonym>exact_synonym n°1 "structure" [TyDI:1405823]
exact_synonym n°2 "structure" [TyDI:1279182] </exact_synonym>
....
<name>dépôt</name>
<id>0000243</id>
<is_a>ID:0000743 ! structure</is_a>
<xref>xref n°1 TyDI:1265993 xref n°2 TyDI_semClass:1613178 </xref>
<exact_synonym>exact_synonym n°1 "dépôt" [TyDI:1500006]
exact_synonym n°2 "dépôt" [TyDI:1357747] </exact_synonym>
...
<name>fosse</name>
<id>0000392</id>
<is_a>ID:0000243 ! dépôt</is_a>
<xref>xref n°1 TyDI:1357294 xref n°2 TyDI_semClass:1613338 </xref>
<exact_synonym>exact_synonym n°1 "fosse" [TyDI:1443548]
exact_synonym n°2 "fosse" [TyDI:1307331] </exact_synonym>
...
</racine>
```

TyDI, INRA, http://faculty.washington.edu/fxia/LAWVI/proceedings/cdrom/pdf/LAWVI21.pdf

29 Le format OBO permet de gérer les phénomènes de synonymie et variation lexicale. Ainsi dans notre ontologie, *âge du Fer* a pour *exact synonyme* l'ensemble des variations possibles de ce terme : *âge du Fer* ou *Âge du Fer* (figure 7). De même, les mentions en texte de la période protohistorique *Âge du bronze* ou *Âge du bronze moyen* sont rattachées à un seul et même concept *Âge du bronze*, fils du concept *période protohistorique*.

FIGURE 7. LE FORMAT OBO : LA SYNONYMIE

```
[Term]
id: ID:0000003
name: Âge du Bronze
xref: TyDI:1394715
xref: TyDI_semClass:1612918
exact_synonym: "âge du bronze" [TyDI:1264890]
is_a: ID:0000348 ! époque protohistorique

[Term]
id: ID:0000004
name: Âge du Bronze final
xref: TyDI:1393647
xref: TyDI_semClass:1612919
is_a: ID:0000003 ! Âge du Bronze

[Term]
id: ID:0000005
name: Âge du Bronze moyen
xref: TyDI:1562589
xref: TyDI_semClass:1612920
is_a: ID:0000003 ! Âge du Bronze

[Term]
id: ID:0000006
name: âge du Fer
xref: TyDI:1403050
xref: TyDI_semClass:1612921
exact_synonym: "âge du Fer" [TyDI:1362574]
exact_synonym: "âge du fer" [TyDI:1309595]
is_a: ID:0000348 ! époque protohistorique
```

TyDI, INRA, http://faculty.washington.edu/fxia/LAWVI/proceedings/cdrom/pdf/LAWVI21.pdf

30 Quel que soit l'éditeur d'ontologie utilisé, les variations lexicales restent les mêmes et sont au cœur de la structuration de l'ontologie. Selon le type de texte à analyser et la richesse de la structure souhaitée, le format de l'ontologie est différent. Le classement des termes est fait par le terminologue, en l'occurrence ici l'expert du domaine. Seul l'archéologue peut établir un modèle de classification en vertu de ses connaissances. L'expert doit établir un modèle, définir les critères de normalisation.

31 Le format OBO a permis aux membres du projet de créer un premier thésaurus, depuis l'extraction lexicale faite par YaTeA. Si ce format s'est avéré trop spécifique au domaine de la biologie, il a cependant été facile

de transformer ce format standard en un autre format davantage en adéquation avec la tâche d'application. Nous avons, par le biais de scripts, transformé le format OBO en un format SKOS [6]. Nous aurions pu, dans ce compte rendu d'expérience, évoquer directement SKOS et laisser de côté OBO, mais il nous semble intéressant malgré tout de montrer les tâtonnements qui ont été les nôtres. Il existe de multiples outils et de multiples standards. Il était sans doute prévisible que SKOS était plus approprié que OBO (car SKOS est plus standard, compatible avec RDF, etc.) mais la présence d'outils disponibles et efficaces autour du standard OBO nous a d'abord incités à aller voir de ce côté.

Des lignes de commande à un environnement graphique

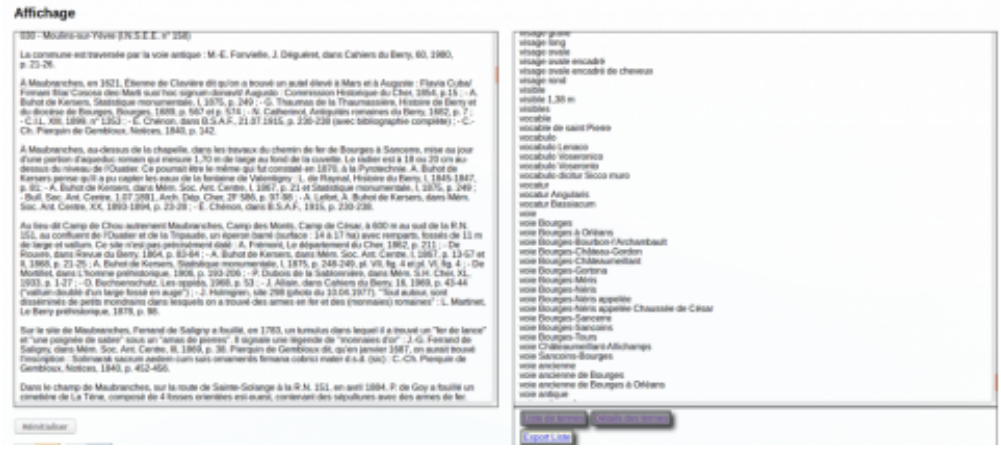
32 La seconde étape de l'étude pilote était la mise en place d'une plateforme pour l'analyse des termes par les experts (extraire les termes, comparer les termes extraits à un thésaurus existant, et permettre un travail itératif pour l'enrichissement de bases de données existantes).

Extraire, visualiser, exporter le lexique

33 Une fois l'analyse terminologique terminée, l'interface permet de voir le texte support et les termes extraits. La figure 8 illustre la liste obtenue en sortie de traitement de YaTeA. YaTeA est utilisé avec ses paramétrages par défaut (comme expliqué ultérieurement), la liste des termes obtenue contient les monogrammes même quand ceux-ci appartiennent à une chaîne plus longue, autrement dit la liste comprend des termes enchâssés, comme *voie* par rapport à *voie antique*.

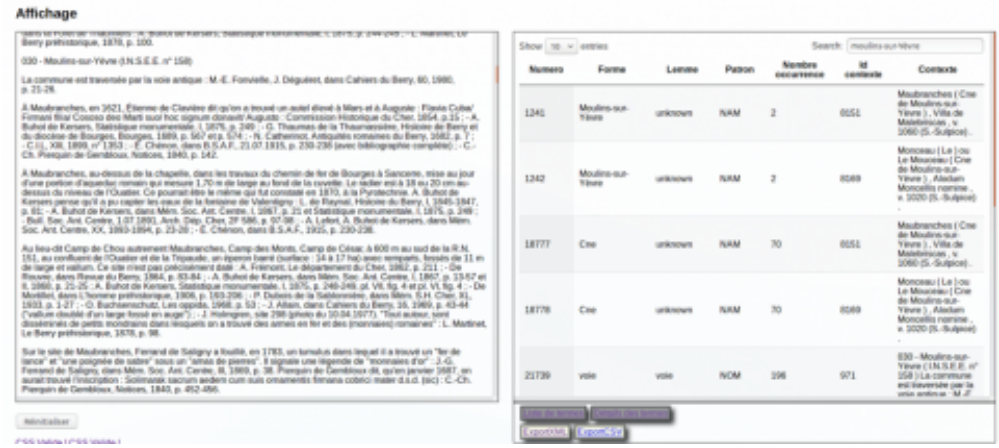
34 Il est aussi possible de produire une liste enrichie. Cette liste regroupe l'ensemble des informations mises à disposition par YaTeA : non seulement les *termes candidats* repérés et extraits par YaTeA, mais aussi la forme lemmatisée du terme [7], les patrons syntaxiques [8], le nombre d'occurrences de chacun des termes ainsi que leur contexte. Autant d'informations utiles à l'expert pour observer, manipuler le résultat. Ces informations sont présentées sous forme d'un tableau (figure 9) qui peut être filtré directement sur la plateforme ou exporté.

FIGURE 8. LE CHER : LISTE TERMINOLOGIQUE



Outil d'extraction automatique d'informations textuelles
AOROC, http://www.archeo.ens.fr

FIGURE 9. LE CHER : LISTE TERMINOLOGIQUE ENRICHIE



Outil d'extraction automatique d'informations textuelles
AOROC, http://www.archeo.ens.fr

35 L'ensemble de ces informations peut aider l'expert à traiter ses données. Tout d'abord, travailler les termes selon leur nombre d'occurrences permet de les classer du plus fréquent au moins fréquent (ou inversement). Après export au format CSV, il est alors facile de sélectionner une fréquence précise, de rechercher et isoler les hapax. Ensuite travailler la liste selon les patrons extraits permet de regrouper les termes selon leur schéma syntaxique. L'utilisateur peut lister, quantifier les patrons récurrents, et sélectionner si besoin un type de patron ou un patron contenant une étiquette donnée (par exemple ceux ne comptant que des noms propres). Les étiquettes syntaxiques sont celles de TreeTagger (voir plus haut).

36 L'expérience montre qu'il est souvent intéressant de travailler motif par motif : certains motifs sont globalement pertinents et d'autres non, mais il est souvent difficile de le prédire sans regarder les exemples associés. Par exemple, sont souvent pertinentes les séquences composées de deux noms propres (étiquetage *Npropre du Npropre* pour *Domaine du Colombier*), nom commun déterminé par un autre nom commun (étiquetage *Nom de Nom* : *enceinte de terre, cachette de fondeur…*), etc.

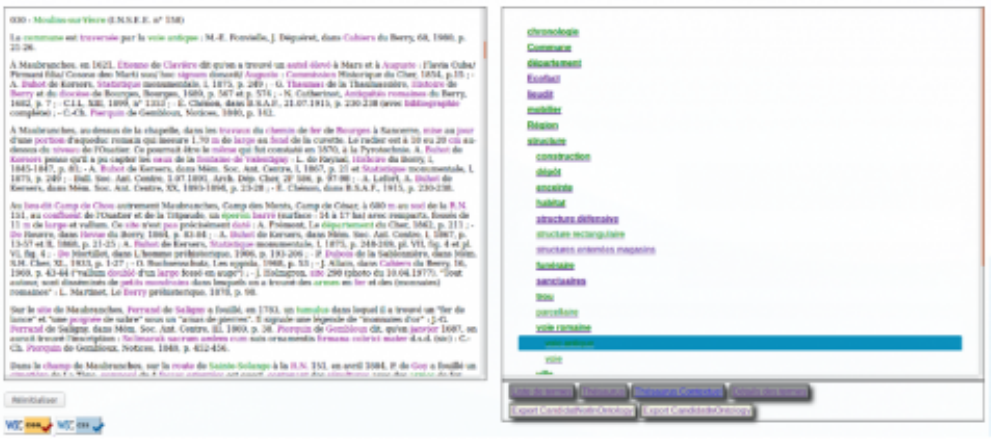
37 L'export de ce tableau au format CSV permet à l'expert d'avoir accès à l'ensemble des informations disponibles hors de l'outil, d'observer et de retravailler la liste à souhait.

Projeter une ontologie

38 L'utilisateur a la possibilité d'associer à son extraction terminologique une ontologie, à condition qu'elle soit au format SKOS.

39 Comme l'illustre la figure 10, lorsqu'une ontologie est associée à l'extraction, dans le texte analysé affiché dans la partie droite de la plateforme certains termes sont mis en évidence soit en vert, soit en pourpre. Les termes extraits présents dans l'ontologie sont en vert, les termes extraits absents de l'ontologie sont en pourpre.

FIGURE 10. PROJECTION DE L'ONTOLOGIE



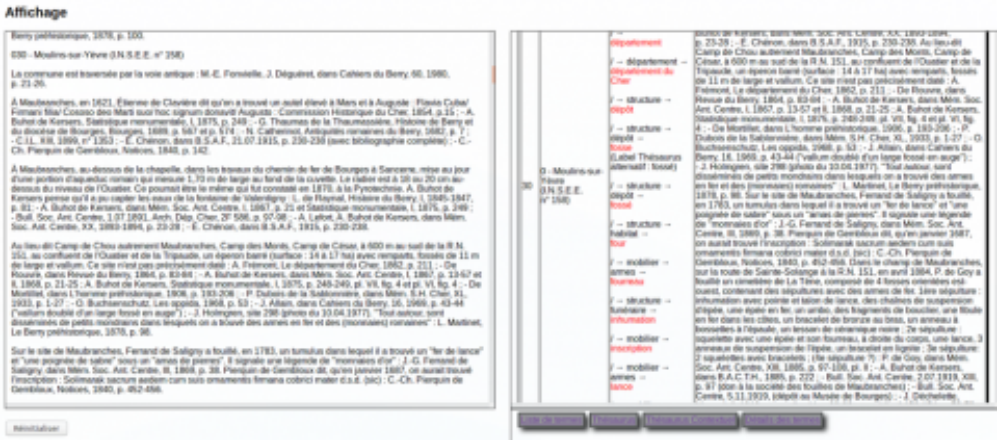
Outil d'extraction automatique d'informations textuelles
AOROC, http://www.archeo.ens.fr

40 Les mêmes couleurs sont utilisées à droite dans l'ontologie pour symboliser en vert les termes présents dans l'ontologie et retrouvés par YaTeA dans le texte, en pourpre les termes ou concepts présents dans l'ontologie mais absents du texte affiché.

41 L'utilisateur peut proposer s'il le souhaite un découpage du texte. Ce découpage permet une projection de l'ontologie sur le texte, c'est-à-dire une contextualisation de l'ontologie, sans que cette projection concerne tout le texte, ce qui rendrait l'application inutilisable (les temps de traitement seraient trop longs). L'interface propose deux découpages possibles, sur la base des paragraphes ou des communes, suivant en cela la structure du document. On rappelle que toutes les *CAG* sont structurées par commune, identifiées par un numéro INSEE.

42 Comme l'illustre la figure 11, il s'agit pour chacune des parties du texte analysé de mettre en évidence dans l'ontologie le terme extrait par YaTeA et de lui associer l'arborescence correspondante. L'utilisateur a alors un accès direct à l'information, il sait quel chemin parcourir dans l'ontologie pour atteindre le terme en question. Il est donc obligatoire qu'il y ait découpage du texte pour qu'il y ait ontologie contextuelle.

FIGURE 11. PROJECTION DE L'ONTOLOGIE



Outil d'extraction automatique d'informations textuelles
AOROC, http://www.archeo.ens.fr

Bilan et perspectives

43 Cette collaboration a permis de développer une solution pratique au problème de la gestion des connaissances dans le domaine de l'archéologie. Bien qu'aucune des technologies utilisées ne soit très novatrice en elle-même, on voit que la mise au point d'une application particulière nécessite de rassembler des personnes d'horizons divers, maîtrisant d'un côté le domaine à traiter et de l'autres les technologies d'extraction et de structuration des connaissances. Les outils actuels sont issus de travaux de recherche, encore délicats à maîtriser en l'absence de véritable documentation et d'assistance, et ils nécessitent des compétences informatiques parfois pointues.

44 L'application visée répondant à un besoin relativement classique, il serait souhaitable que des outils et des solutions plus génériques apparaissent. Il s'agit d'un effort indispensable pour valoriser les connaissances engrangées dans les laboratoires depuis des décennies, qui ne demandent qu'à être utilisées plus largement. Nous n'avons pas montré ici l'application finale, qui reste à déployer. Il s'agit *in fine*, rappelons-le, d'offrir une plateforme d'accès à une littérature difficilement disponible à l'heure actuelle. C'est un enjeu particulièrement important en archéologie, mais aussi dans de nombreux autres domaines des lettres et sciences humaines.

Bibliographie

Buchenschutz, Olivier, Christophe Batardy, Michel Cartereau, Katherine Gruel et Marc Levéry. 2015. « Une base pour l'élaboration de modèles de peuplement de l'Âge du fer en France ». *ArcheoSciences, revue d'archéométrie* 39 : 157-175. https://doi.org/10.4000/archeosciences.4457.

Coulon, Gérard, Jean Holmberg et Michel Provost, 1992. *Carte archéologique de la Gaule. 36. Indre.* Paris : Académie des inscriptions et belles-lettres, ministère de l'Éducation nationale et de la Culture.

Lamani, Sihem. 2013. « Extraction d'information à partir de documents en archéologie ». Rapport de stage.

Papazian, Frédéric, Robert Bossy et Claire Nédellec. 2012. « AlvisAE : a Collaborative Web Text Annotation Editor for Knowledge Acquisition ». Dans *Proceedings of the 6th Linguistic Annotation Workshop*. Édité par l'Association for Computational Linguistics, 149-152, Jeju Island, South Korea, https://www.aclweb.org/anthology/W12-3621.

Provost, Michel, Jean-François Chevrot et Jacques Troadec, 1992. *Carte archéologique de la Gaule. 18. Le Cher*. Paris : Académie des inscriptions et belles-lettres, ministère de l'Éducation nationale et de la Culture.

Schmid, Helmut. 1997. « Probabilistic Part-of-speech Tagging Using Decision Trees ». Dans *New Methods in Language Processing, Studies in Computational Linguistics*. Édité par Daniel Jones et Harold Somers, 154-164. Londres : UCL Press.

Notes

1 *EITAB* est l'acronyme d'un projet PEPS CNRS-PSL et signifie « extraction automatique d'informations textuelles pour alimenter des bases de données » : http://www.archeo.ens.fr/spip.php?article586.

2 L'atlas propose, en complément de ces réponses ponctuelles, des cartes thématiques enrichies et commentées par les chercheurs (www.chronocarto.eu).

3 Nous n'aborderons pas ici la numérisation par reconnaissance de caractères (OCR). Nous nous plaçons au moment où il est possible d'exploiter informatiquement les données. Les documents que nous avons traités dans cette expérience nous ont été fournis au format « texte brut » et nettoyés.

4 https://perso.limsi.fr/hamon/YaTeA/.

5 La balise *part of* est utilisée dans la structure OBO pour indiquer qu'un concept est une partie d'un autre concept. Il n'y a pas d'exemple d'utilisation de cette balise dans notre ontologie.

6 Nous avons testé divers formats d'import et d'export de notre ontologie avec le logiciel libre *Protégé* : http://protege.stanford.edu/. En effet, cet éditeur d'ontologie a l'avantage d'être un logiciel libre et d'avoir de nombreux formats d'import et d'export (RDF, OWL, OBO, etc.) disponibles.

7 Le lemme est l'entrée dictionnairique du terme, sa forme canonique.

8 Un patron syntaxique est un motif, une séquence d'une ou plusieurs étiquettes syntaxiques.

Auteurs

Frédérique Mélanie-Becquet

UMR 8094 Lattice, CNRS, Montrouge, France
Frédérique Mélanie-Becquet est ingénieure d'études au Lattice (UMR 8094 Langues, textes, traitements informatiques, cognition, CNRS, ENS, PSL, université Sorbonne nou-

velle et USPC). Elle travaille au traitement et à l'analyse de données. Elle a plus particulièrement en charge l'élaboration de bases de données linguistiques.
frederique.melanie@ens.fr

Johan Ferguth

Konverso, Ville-d'Avray, France
Johan Ferguth est développeur informatique. Il a été ingénieur d'études au Lattice (UMR 8094 Langues, textes, traitements informatiques, cognition, CNRS, ENS, PSL, université Sorbonne nouvelle et USPC) et au LLF (Laboratoire de linguistique formelle, CNRS) de 2014 à 2017. Il est actuellement ingénieur *devops*/NLP chez Konverso, société d'édition de chatbots et assistants virtuels.
jferguth@gmail.com

Michel Cartereau

UMR 8546 AOROC, AgroParisTech, Paris, France
Michel Cartereau est maître de conférences en informatique à AgroParisTech et au laboratoire AOROC (Archéologie et philologie d'Orient et d'Occident, UMR 8546, CNRS, ENS-PSL), docteur de l'université Pierre et Marie Curie (Paris).
michel.cartereau@agroparistech.fr

Katherine Gruel

AOROC, CNRS, Paris, France
Katherine Gruel est directrice de recherche au CNRS (AOROC, Archéologie et philologie d'Orient et d'Occident, UMR 8546, CNRS, ENS-PSL). Archéologue spécialiste en protohistoire celtique et en numismatique, elle participe au développement d'outils d'exploitation numérique des données. Elle est responsable de la BaseFer et du portail Chronocarto.
katherine.gruel@ens.fr

Thierry Poibeau

UMR 8094 Lattice, CNRS, Montrouge, France
Thierry Poibeau est directeur de recherche au CNRS et directeur adjoint du laboratoire Lattice (UMR 8094 Langues, textes, traitements informatiques, cognition, CNRS, ENS, PSL, université Sorbonne nouvelle et USPC). Il est spécialiste de traitement automatique des langues, un domaine de recherche à la frontière de la linguistique et de l'informatique.
thierry.poibeau@ens.fr

Droits d'auteur