Guide d'annotation pour INCEpTION

Dernière mise à jour : 07/12/2022

> ***Note préliminaire*** *: pour en savoir plus sur l'utilisation d'INCEpTION, se référer à la documentation et aux tutoriels vidéo mis à disposition par le* Ubiquitous Knowledge Processing Lab*.*

Dans le cadre du projet européen ARIADNEplus[1] (Task 16.8), l'Inrap[2] a été amené à tester INCEpTION en l'appliquant à un échantillon de ses rapports d'opération archéologique.

INCEpTION[3] est un outil open source d'annotation textuelle développé par le *Ubiquitous Knowledge Processing Lab* de la *Technische Universität* Darmstadt (Allemagne). Ce type d'outil intervient en traitement automatique des langues[4] (TAL ou TALN pour traitement automatique du langage naturel, également nommé NLP pour *Natural Language Processing*) pour créer les données annotées nécessaires à l'entraînement et à l'évaluation de modèles.

Au croisement de « la linguistique, l'informatique, les mathématiques (sous la forme de l'algèbre, de la logique ainsi que des statistiques) et l'intelligence artificielle[5] », le TAL recouvre « l'ensemble des recherches et développements visant à modéliser et reproduire, à l'aide de machines, la capacité humaine à produire et à comprendre des énoncés linguistiques[6] ». Il permet notamment l'extraction d'informations (EI), permettant de recueillir des données structurées à partir de documents non-structurés.

Ce guide décrit l'annotation suivie pour extraire des rapports d'opération les entités nommées relatives aux références chronologiques, au contexte archéologique, aux intervalles chronologiques, aux matériaux, aux éléments de mobilier, ainsi qu'aux mentions de techniques et styles. Six rapports ont pour le moment pu être annotés selon cette méthode. A terme, ce processus pourrait permettre d'affiner la recherche de concepts au sein des rapports d'opération, leur enrichissement sémantique en reliant ces concepts à des bases de connaissances externes, ou bien proposer une aide à l'indexation.

Le corpus créé selon ce guide d'annotation est disponible sur un dépôt GitHub[7].

---

[1] https://ariadne-infrastructure.eu/ (consulté le 06/12/2022). ARIADNEplus est un projet Horizon 2020 financé par la Commission européenne dans le cadre de la convention de subvention n° 823914. Les points de vue et opinions exprimés dans cette publication relèvent de la seule responsabilité de l'auteur et ne reflètent pas nécessairement ceux de la Commission européenne.

[2] https://www.inrap.fr/ (consulté le 07/12/2022).

[3] Klie, J.-C., Bugert, M., Boullosa, B., Eckart de Castilho, R., & Gurevych, I. (2018). The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, 5–9. https://aclanthology.org/C18-2002.

[4] En matière d'apprentissage machine supervisé plus spécifiquement.

[5] Cori, M., & Léon, J. (2002). La constitution du TAL. *Revue TAL*, 43(3). halshs-00158854.

[6] Yvon, F. (2007). Une petite introduction au Traitement Automatique des Langues Naturelles. https://perso.limsi.fr/anne/coursM2R/intro.pdf.

[7] https://github.com/InrapFr/NLP_for_French_Archaeological_Reports_ARIADNEplus (consulté le 06/12/2022).

SOMMAIRE

1. Se connecter à INCEpTION

1. Se connecter au portail D4Science pour ARIADNEplus

2. Cliquer sur “*Go to*” (en haut à droite de la page d’accueil), puis cliquer sur “ARIADNEplus_Lab” (fig. 1)

Figure 1 : Localisation de l’onglet “Go to”.

3. Cliquer sur l’onglet “INCEpTION” (fig. 2)

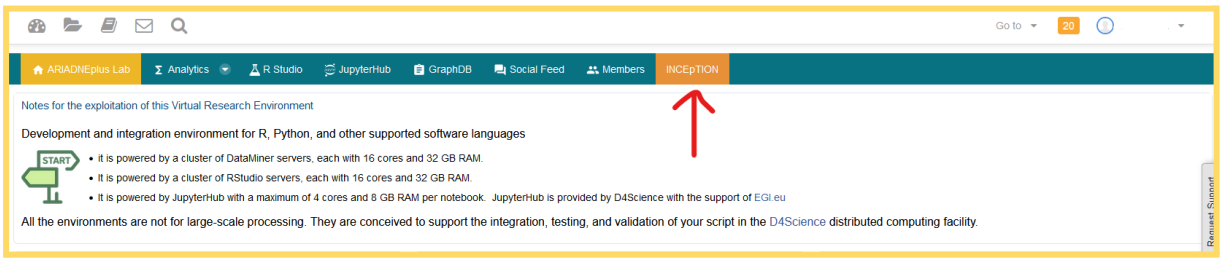

Figure 2 : Localisation de l’onglet “INCEpTION”.

4. Si INCEpTION ne s’ouvre pas, cliquer sur le lien fourni (Fig. 3)

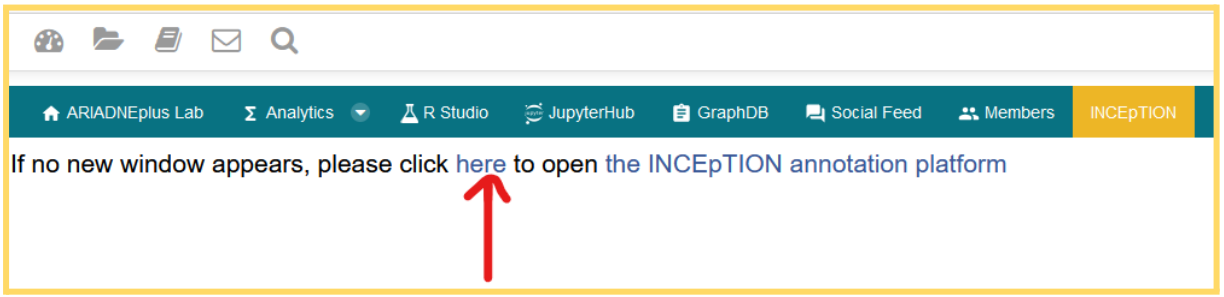

Figure 3 : Accès à INCEpTION.

L'utilisateur doit ensuite cliquer sur le nom du projet avant de sélectionner "Annotation" en haut du volet latéral gauche.

2. Annoter des documents

Aide à l'annotation : une suggestion automatique d'annotation existe sur INCEpTION, les *Recommenders*. Cliquer une fois sur une suggestion (étiquette grisée) pour l'accepter, cliquer deux fois pour la rejeter.

Cette aide ne vient cependant pas remplacer le travail de l'annotateur, qui doit s'assurer de signaler les annotations manquantes, ou de les modifier ou supprimer au besoin. Il ne s'agit pas uniquement de corriger les entités créées à l'aide des *Recommenders*, mais également de vérifier que des mentions n'ont pas été oubliées dans le reste du texte.

2.1 Déroulement du processus d'annotation

1. **Annoter** son document selon les recommandations d'annotation (voir ci-dessous).
2. Une fois que le document est prêt, cliquer sur "**Finish document**" (fig. 4) pour le verrouiller.

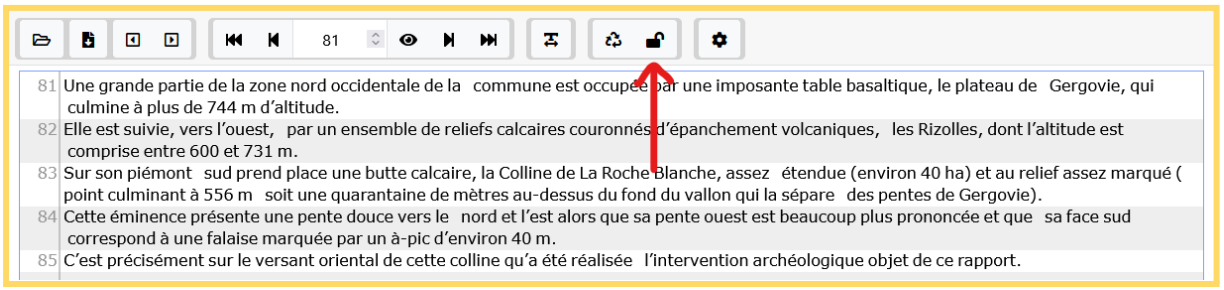

Figure 4 : Localisation du bouton "Finish document".

Le présent guide est disponible à tout moment lors de l'annotation en cliquant sur l'icône de livre localisée sur la barre d'outils (fig. 5).

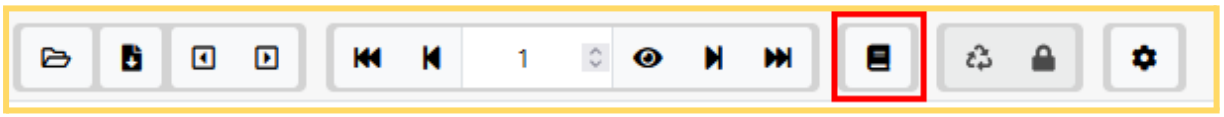

Figure 5 : Localisation du bouton "Guidelines".

2.2 Recommandations d'annotation

2.2.1 Entités à annoter

Les exemples d'entités à annoter sont ici surlignés en jaune.

- **Généralités**

- Prendre garde à ne pas annoter par erreur des caractères parasites : espaces entourant le mot, signes de ponctuation superflus, caractères appartenant à un mot que l'on ne souhaite pas annoter, etc. Cependant, cela est parfois inévitable en raison de la manière dont les mots ont été tokenisés. Par exemple, les chiffres indiquant une note de bas de page sont fréquemment accolés au mot auquel ils se rapportent (donnant un résultat de type "néolithique2").

- Annoter les fautes et variations orthographiques : "ep. medievale", "ép. médiévale", "époque médiévales".

- Les valeurs séparées par des "et", "ou" constituent deux annotations différentes : "Ier ou IIe siècle", "Néolithique ancien et moyen".

- Ne pas faire d'annotation si le contexte est à la forme **négative**. "Pas de traces de céramiques." "Aucune structure antique n'a été retrouvée."

- De manière générale, la logique suivie est celle d'une annotation "large" afin de recueillir le maximum d'informations et de vocabulaire : "période romaine", "amphore vinaire", "extrême fin de La Tène finale", "cruche à bec à pied" (au lieu de "période romaine", "amphore vinaire", "fin de La Tène finale", "cruche à lèvre en poulie et à pied annulaire"). Le but étant d'entraîner le modèle à reconnaître l'entité entière, et non seulement sa partie essentielle, afin d'obtenir les résultats les plus parlants possibles.

- Annoter les mentions pertinentes retrouvées en bibliographie (ceci exclut les dates de publication).

- Une annotation ne peut être à cheval sur deux phrases. Si c'est le cas d'une entité, renoncer à l'annoter, sauf si l'un des morceaux conserve du sens. Ainsi, "Moyen / Âge" n'est pas annoté ("/" représentant ici le retour à la ligne), mais "Haut / Moyen Âge" l'est partiellement.

- **CHRONO**

Cette catégorie recouvre les références chronologiques hors intervalles. Elle recouvre les concepts "dates absolues" et "périodisations" du concept parent "entités temporelles" de Pactols.

Données à annoter sous le tag “CHRONO” :

- Les périodes nommées : “Préhistoire”, “Haut Moyen Age”.
- Les références entre parenthèses de type “Paléolithique (supérieur)”, l’identifiant associé est alors “Paléolithique supérieur”.
- Les adjectifs se référant à une période : “occupation médiévale”, “camp romain”, “à la période médiévale”, “lors de l’époque romaine”..
- Les abréviations : “ep. médiévale”, “HMA”.
- Les siècles et millénaires et leurs fractions: “XIIe siècle”, “début du Ve s. av. J.-C.”, “Ier millénaire”, “premier quart du IIe siècle avant notre ère”, “2e moitié du IXe siècle”.
- Les dates absolues : “en 1223”, “700 avant notre ère” , “300 cal BC”.

Données à **ne pas** annoter sous le tag “CHRONO” :

- Les saisons, mois, jours du mois, jours de la semaine, heures et autres durées de temps : “Une heure”, “Douze jours”, “Quatre siècles”, “plusieurs millénaires”.
- Les ères géologiques (Quaternaire, Holocène, Cénozoïque, Boréal, Würm…).
- Les références chronologiques relatives aux opérations archéologiques : "une équipe est intervenue du 18 mars au 14 juin 2019".
- Ne pas annoter les informations trop imprécises : “aux environs du Xe siècle”, “au cours du Ve siècle”, "au siècle suivant", "un siècle plus tard".

- **CONTEXTE**

Il s’agit ici d’annoter les faits archéologiques, définis selon l'*Archaeological Institute of America*[8] comme étant "Toute structure ou élément physique, tel qu'un mur, un trou de poteau, une fosse ou un plancher, qui est fabriqué ou modifié par l'homme mais qui (contrairement à un artefact) n'est pas transportable et ne peut être retiré d'un site". Ils peuvent être positifs (ajout de matière, mur par exemple) ou négatifs (retrait de matière, fosse par exemple).

Cette définition correspond à la rubrique "environnement bâti" de Pactols et du BackBone Thesaurus, à savoir "toute structure, simple ou complexe, quelle que soit sa taille, sa durée de construction ou son utilisation, attachée au sol ou enfouie et qui ne peut être déplacée sans dommage irréversible" (DARIAH Backbone Thesaurus, v. 1.2.8[9]).

Cette catégorie recouvre plus ou moins le concept "environnements bâtis" de Pactols, ainsi que quelques entrées du concept "caractéristiques physiques" du concept parent "entités matérielles" (avec des exceptions, à l'instar de "maçonnerie").

Données à annoter sous le tag “CONTEXTE” :

[8] https://www.archaeological.org/programs/educators/introduction-to-archaeology/glossary/#f, terme *Feature,* (consulté le 28/11/2022).
[9] https://vocabs.dariah.eu/bbt/Concept/000018 (consulté le 28/11/2022).

- Les faits, vestiges et éléments de contexte : "une fosse", "le fossé", "un foyer", "un empierrement", un "mur", "deux trous de poteaux", "des ornières", "les traces d'une habitation", "une entrée".
- Les fossés parcellaires s'ils ont été fouillés.
- Les termes "bâtiment", "construction", et "maçonnerie".
- Les mentions de voies et chemins si elles constituent des entités archéologiques : "la voie romaine", mais "l'emprise est limitée au nord par la voie rapide".
- Les édifices mentionnés lors de la présentation de l'opération : "l'église Sainte-Anne". Ce dernier point peut être ouvert à débats.

Données à **ne pas** annoter sous le tag “CONTEXTE” :

- Les notions trop abstraites ou larges de type "habitat", "voirie", "ville", bourg". L'utilisation de certains termes peut cependant être débattue ("habitat" étant parfois utilisé pour "habitation", ou "voirie" pour "voie").
- Les mentions trop vagues de type "comblement" ou "creusement", mais penser à prendre "remblai".
- Ne pas annoter les sédiments, minéraux et autres éléments relatifs à l’environnement de l’opération : "du limon argileux".
- Les termes "vestiges", "structure" (sauf si caractérisé pour ce dernier terme, "une structure urbaine") et "aménagement".

- **INTERVALLE**

Il s’agit ici d’annoter les références chronologiques sous forme d'intervalle de temps. Ce type d'entité est le seul à ne pas être associé à des concepts Pactols.

Données à annoter sous le tag “INTERVALLE” :

- “du Néolithique à la fin de l'Antiquité”, “entre le Xe et le XIIe siècle”, "en 230-245 après notre ère", “2845 ± 30 BP”.

Données à **ne pas** annoter sous le tag “INTERVALLE” :

- Les restrictions de la catégorie CHRONO s'appliquent également à INTERVALLE (pas d'annotation d'ères géologiques, de saisons, etc.).
- Ne pas annoter les mentions trop imprécises : "pendant 20 ans".
- Les intervalles trop espacés au sein d'une phrase, typiquement lorsque séparés par d'autres entités : "entre la fin de l'époque médiévale, comme attesté par la présence d'un grès, et le début de l'époque moderne". Utiliser alors deux tags CHRONO pour les références chronologiques.

- **MAT**

Il s'agit ici d'annoter les matériaux relatifs aux éléments de mobilier ou aux vestiges. Cette catégorie recouvre le concept "matériaux" de Pactols.

Données à annoter sous le tag "MAT" :

- "Une coupe en bronze", "un pot en céramique".
- "Un sarcophage en briques".

Données à **ne pas** annoter sous le tag "MAT" :

- Prendre garde à la distinction entre mobilier et matériau : "une céramique", "un vase en céramique", ces derniers faisant l'objet du tag dédié MOB.
- Ne pas annoter les sédiments, minéraux et autres roches relatifs à l'environnement de l'opération : "du limon argileux".
- Ne pas annoter les matériaux récents : "plastique", "PVC".

- **MOB**

Il s'agit ici d'annoter les découvertes pouvant être déplacées ou retirées du site sans être altérées irréversiblement. Cette catégorie recouvre plus ou moins la rubrique "objets mobiles" de Pactols (avec des exceptions, à l'instar de "poteau").

Données à annoter sous le tag "MOB" :

- Les artéfacts : "une fibule", "la hache polie", "un pot"
- Les écofacts : "des graines d'orge", "des ossements".
- Leurs notions générales: "des éléments d'industrie lithique" "de la poterie", "le mobilier céramique", "un objet".
- Les mentions trop vagues sont annotées en bloc comme MOB : un tesson de céramique", "une esquille d'os", "une gouttelette en alliage cuivreux".
- Annoter également les éléments de construction pouvant être déplacés ou retirés sans être altérés : "des moellons", "un bloc sculpté", "un poteau" (mais pas "un trou de poteau" ou "mur").

Données à **ne pas** annoter sous le tag "MOB" :

- Les éléments morphologiques d'objets : "panse", "col", "talon", "épaule", etc.
- Les quantités : "trois amphores", "de nombreuses amphores".
- Les mentions de fragments et morceaux : "des fragments d'amphores", "des morceaux d'amphores". "Tesson", "éclat" ou encore "esquille" portant à l'inverse une quantité supérieure d'information.

- Les adjectifs de type "grand", "petit", "léger", "lourd" : "un petit pot". Une exception est faite quand cela relève d'une typologie, à l'instar de "os longs".
- Ne pas annoter le matériau associé s'il existe : "une cruche en bronze", "une assiette en céramique".
- Ne pas annoter des éléments se rapportant à la technique ou au style de l'objet : "une jatte tournée", une hache polie", ces derniers faisant l'objet du tag dédié TECH_STYLE.
- Prendre garde à la distinction entre mobilier et matériau : "une céramique", "un vase en céramique", ces derniers faisant l'objet du tag dédié MAT.
- Ne pas annoter les informations relatives à l'état de conservation l'artéfact : "amphore brisée", "fibule oxydée".

- **TECH_STYLE**

Il s'agit ici d'annoter les mentions de techniques et styles de fabrication ou construction. Cette catégorie recouvre plus ou moins les concepts "technique de construction", "technique de décor", et "technique de fabrication" de Pactols.

Données à annoter sous le tag "TECH_STYLE" :

- "une hache polie", "une céramique tournée", "un mur en petit appareil", "un bloc en remploi comme calage", "maçonné", "une céramique modelée", une pointe façonnée par troncature", "un silex taillé", "une céramique à figures noires".

Données à **ne pas** annoter sous le tag "TECH_STYLE" :

- Prendre garde à ne pas annoter l'élément de mobilier ou la structure à laquelle fait référence la technique ou le style : "un os gravé", ceux-ci faisant l'objet du tag dédié MOB.

2.2.2 Créer une annotation

Pour créer une annotation, vérifier en haut à droite de la fenêtre d'annotation que le calque (*layer*) sélectionné soit bien "*Named Entity*" (fig. 6). Ce choix se fait en créant ou sélectionnant une annotation (quitte à la supprimer juste après).

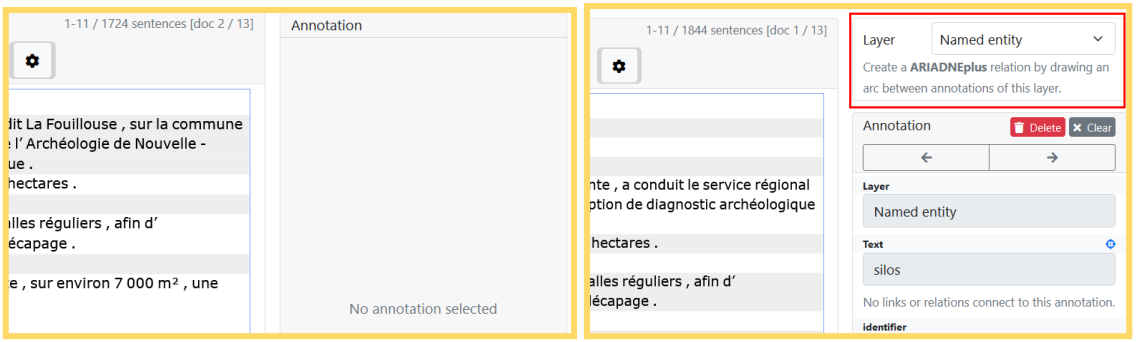

Figure 6 : Sélection du calque.

Il suffit après de sélectionner dans le texte l'entité que l'on souhaite annoter.

Le tag (ou étiquette) associé se sélectionne ensuite avec le menu déroulant de la rubrique "**value**" du menu de l'annotation, localisé en bas à droite de l'écran. Pour gagner du temps, il est possible de sélectionner le tag d'une entité directement avec des raccourcis clavier.

Tags et raccourcis claviers sont regroupés dans la table 1. Un code couleur a également été implémenté dans INCEpTION pour visualiser plus rapidement les différents types d'entités et détecter plus facilement les erreurs.

Tag	Portée	Raccourci clavier
CHRONO	Utilisé pour les références chronologiques.	ALT C
CONTEXTE	Utilisé pour les faits archéologiques.	ALT W
INTERVALLE	Utilisé pour les intervalles de temps.	ALT I
MAT	Utilisé pour les matériaux.	ALT X
MOB	Utilisé pour le mobilier.	ALT M
TECH_STYLE	Utilisé pour les techniques et styles de fabrication ou construction.	ALT T

Table 1 : Tags et raccourcis pour les entités.

Pensez à cliquer sur "***Show key bindings***" en bas à droite de la fenêtre d'annotation (fig. 7) pour afficher les raccourcis de sélection de tag.

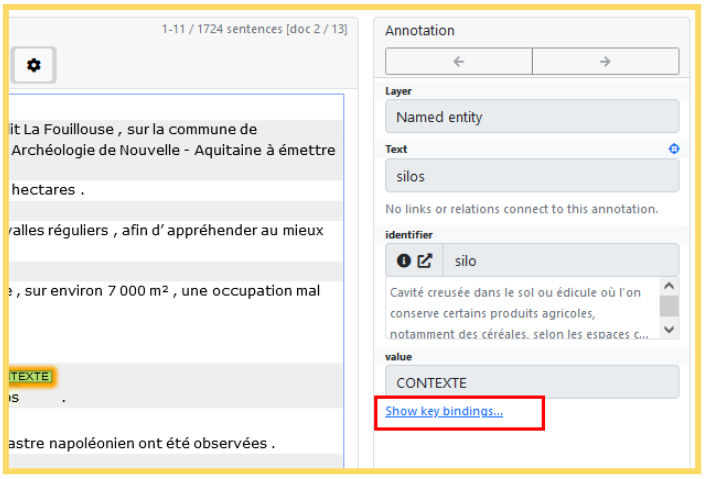

Figure 7 : Afficher les raccourcis de sélection de tag.

2.2.3 Relier une entité à la base de connaissance Pactols (annotation sémantique)

Toutes les entités (à l'exception de celles du type INTERVALLE) ont été reliées à des entrées Pactols 2. Il est recommandé d'avoir Pactols ouvert dans un autre onglet afin de pouvoir facilement explorer la base de connaissances.

L'identifiant d'une annotation se sélectionne dans la rubrique "***identifier***" du menu de l'annotation, localisé à droite de l'écran (au-dessus de "*value*"). L'utilisateur doit taper puis sélectionner dans la barre de recherche le nom de l'entrée du thésaurus Pactols qu'il souhaite associer à l'entité sélectionnée ("époque médiévale" pour "Moyen Âge", ou "époque moderne" pour "Temps Modernes" par exemple).

Le but est d'essayer de faire correspondre au maximum l'entrée Pactols avec l'entité en question. Ainsi, pour la séquence "une hache polie", composée de l'entité MOB "hache" et de l'entité TECH_STYLE "polie", "hache" est associée au concept Pactols "hache polie" (et non "hache") et "polie" au concept Pactols "polissage".

Cela peut être parfois complexe, comme illustré par ces exemples :

- Il n'existe pas dans Pactols d'entrée pour "**fossé**" (CONTEXTE) (même s'il existe des concepts plus spécifiques tels que "fossé parcellaire", ou "fossé défensif" par exemple). Il a donc été choisi de relier les entités de "fossé" au concept supérieur présent dans Pactols, à savoir "unité bâtie". Ce terme est cependant très générique.

- La mention de "**terres cuite**" ou "**céramique**" (MOB) en tant que mobilier a été reliée au concept "objet en terre cuite". "céramique (objet)" et "céramiques" étant des

variantes du libellée du concept. Il est cependant possible de débattre de l'utilisation du concept enfant "poterie" pour les entités "céramique".

- L'emploi du mot "**terre cuite**" ou "**céramique**" (MAT) en tant que matériau doit en revanche être associé aux concepts Pactols associés, à savoir "terre cuite" et "céramique (matériau)".

- De la même manière, prendre garde à ne pas confondre les concepts Pactols "**bronze**" (matériau) et "**objet en bronze**" (mobilier) (fig. 8).

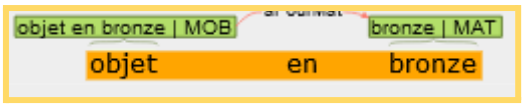

Figure 8 : Annotation de la séquence "objet en bronze"

- "**Zone de stockage**" (CONTEXTE) a été relié au concept Pactols "entrepôt", tandis que "**zone d'artisanat**" (CONTEXTE) a été relié au concept Pactols "atelier".

- "**Bâtiment**" ou "**construction**" (CONTEXTE) a été relié au concept "édifice".

- "**Pierres brûlées**" (MOB) a été relié au concept Pactols "Roche".

- "**Bloc**" (MOB) a été relié au concept Pactols "parties structurelles d'entités matérielles" (bien que trop vague à notre goût).

- Il a été décidé d'annoter les entités CHRONO relatives à l'**époque romaine** et **gallo-romaine** avec le concept Pactols "Empire romain". Il est cependant possible de débattre de l'utilisation du concept enfant "époque gallo-romaine" à la place.

- En cas de mention de "**céramique romaine**", "**céramique médiévale**" etc., l'entité "céramique" (MOB) est associée au concept Pactols équivalent (céramique romaine, céramique médiévale..). Cela ne vaut pas pour la mention de simple "terre cuite" ou "tesson" d'une certaine période, car ces termes ne sont pas synonymes de céramique (il peut s'agir d'un tesson de grès par exemple). On utilisera alors le concept Pactols générique approprié ("objet en terre cuite" par exemple). L'adjectif relatif à la chronologie sera quant à lui annoté comme une entité CHRONO, avec le concept Pactols approprié.

2.2.4 Créer des relations entre entités

Des relations entre entités ont été créées à titre expérimental (table 2). Pour créer une relation, il suffit de tracer un arc de cercle partant de l'entité de départ pour arriver à l'entité

d'arrivée ; un tag (label) doit ensuite lui être associé. Une relation ne peut pas être à cheval sur deux phrases différentes ; si c'est le cas, elle doit être abandonnée.

Tag de la relation	Sens de la relation	Exemples[10]
aPourChrono	MOB → CHRONO CONTEXTE→ CHRONO	« Une fibule protohistorique », « Une fosse du second âge du Fer »
aPourContexte	MOB → CONTEXTE	« Le fossé a livré des ossements »
aPourIntervalle	MOB → INTERVALLE CONTEXTE → INTERVALLE	« Un tesson de céramique daté des XVIe-XVIIe siècles », « Un puits utilisé du premier âge du Fer au Haut-Empire »
aPourMat	MOB → MAT CONTEXTE→ MAT	« Une tuile en terre cuite », "Un empierrement en meulières"
aPourTechOuStyle	MOB → TECH_STYLE CONTEXTE→ TECH_STYLE	« une céramique tournée », « un mur en petit appareil »

Table 2 : Typologie des relations créées.

Dans le futur, il pourrait être intéressant de créer des relations entre entités CONTEXTE et entre entités MOB pour indiquer des associations de type : "l'enclos comprend trois fosses" (CONTEXTE), "des os ont été retrouvés dans un pot" (MOB).

2.3 L'outil rechercher

Tant qu'un document n'a pas été cadenassé (fig. 4), il est possible pour l'utilisateur de rectifier ses annotations sur celui-ci. L'outil rechercher ("*search*") permet à cet égard de vérifier rapidement les mots et annotations d'un même projet. L'outil est accessible à partir du panneau latéral gauche de la fenêtre d'annotation (icône de loupe).

> ***N. B.*** : l'outil rechercher ne permet pas à l'heure actuelle d'ignorer les diacritiques et la casse. Rechercher "céramique", "ceramique", "Céramique" ou "Ceramique" retournera quatre types de résultats différents. L'outil récupère en effet les correspondances exactes ; "amphore" et "amphores" constituent donc deux recherches bien distinctes.

Il est possible d'accéder aux options de l'outil rechercher en cliquant sur le rouage du menu "*search*" :

[10] L'entité de départ de la relation est surlignée en jaune, celle d'arrivée en bleu.

- Cocher la case "*current document only*" permet de ne requêter que le document actuellement ouvert.

- L'option "*Grouping by*" permet de regrouper les résultats par calque (*layer*) et par rubrique (*feature*).
  Chercher le mot "céramique" en regroupant les résultats selon le calque *Named Entity* et la rubrique *value* permet par exemple d'obtenir les mentions de céramique, groupées par tag MOB et tag MAT.

- "*Rebuild index*" permet de résoudre d'éventuels problèmes rencontrés lors d'une recherche (résultats dupliqués par exemple).

L'outil rechercher permet des recherches fines sur les documents d'un même projet. Ainsi :

- Rechercher "**<Named_entity/>**" retourne l'intégralité des entités nommées annotées.
- Rechercher "**<Named_entity.value="MOB"/>**" retourne les entités nommées de type MOB.
- Rechercher "**<Named_entity/>{2}**" retourne les occurrences où deux entités nommées sont à la suite l'une de l'autre.
- Rechercher "**<Named_entity/> []{1,3} <Named_entity/>**" retourne les occurrences où une première entité nommée est séparée d'une seconde entité nommée par 1 à 3 tokens.
- Rechercher "**<Named_entity.identifier="cruche"/>**" retourne l'intégralité des entités nommées ayant été reliées au concept Pactols "cruche".

D'autres types de recherches fines sont décrits dans la documentation.

Enfin, l'outil rechercher permet de créer ou de supprimer en masse des annotations (voir le guide utilisateur de la documentation). L'impossibilité d'annuler ces actions impose une grande prudence quant à leur utilisation.

3. Remarques

Plusieurs points méritent d'être mieux définis et tranchés au sein du corpus annoté :

- **Les pâtes de céramiques**

Si ces mentions, de type "une céramique en pâte commune claire" ou tout simplement "des pâtes rouges", n'ont globalement pas été annotées, quelques entités de type MAT ont toutefois été créées.

- **Les adjectifs et compléments relatifs aux matériaux**

L'annotation des matériaux manque de régularité en ce qu'ils sont parfois annotés de manière large : "quatre pièces en silex secondaire" "en verre fin légèrement verdâtre", "un récipient en céramique commune à pâte sableuse" (en lien également avec la remarque précédente), et parfois au contraire de manière restrictive : "une hache polie en silex brûlé", "débité dans un silex santonien".

- **Les abréviations**

De même, toutes les abréviations n'ont pas été prises en compte de la même façon. Si des abréviations de type "TP" (trou de poteau), "GR" (gallo-romain), "HMA" (Haut Moyen Âge) ou encore "TCA" (terre cuite architecturale) ont été annotées, d'autres comme "LT" (La Tène), "MR" (mur), "FS" (fosse), "FO" (fossé) ou "TC" (terre cuite) ont été laissées de côté.

Ces trois questions devront donc être tranchées et le corpus annoté, rectifié en conséquence.

Remerciements

Ce guide d'annotation s'est inspiré des guidelines rédigées par Lucas Terriel pour le projet NER4Archives[11] (Inria ALMAnaCH et Archives nationales) ainsi que de celles mises au point par Alex Brandsen dans le cadre de sa thèse[12].

[11] https://gitlab.inria.fr/almanach/ner4archives/-/blob/master/inriaAlmanach/Inception-config/guidelines/guidelines_ner4Archives.pdf (consulté le 14 octobre 2022).

[12] Brandsen, A. (2022). *Digging in documents: Using text mining to access the hidden knowledge in Dutch archaeological excavation reports* [Thèse de doctorat, Université de Leiden]. https://www.researchgate.net/publication/358746796_Digging_in_documents_using_text_mining_to_access_the_hidden_knowledge_in_Dutch_archaeological_excavation_reports