Thésaurus Artefacts : présentation et retour d'expérience

Elise Vigier

THESAURUS ARTEFACTS : PRESENTATION ET RETOUR D'EXPERIENCE

Elise Vigier Chercheur associée UMR 5138 ArAr

Texte prononcé lors d'une communication invitée lors de l'atelier thématique « Thésaurus appliqués » organisé par Bibracte EPCC à Glux-en-Glenne les 24 et 25 octobre 2022.

Bonjour à tous,
Je vais vous présenter le thésaurus élaboré pour la base de données Artefacts et partager un retour d'expérience. La première version a été élaborée durant deux mois dans le cadre du programme ArteBib CollEx Persée 2019-2020.

1. Introduction : présentation d'Artefacts

Artefacts est un projet porté par l'UMR 5138 ArAr : Archéologie et Archéométrie depuis 2009. C'est une base de donnée collaborative en ligne qui traite des objets archéologiques de tous matériaux, sauf la vaisselle céramique, la monnaie et les restes non typologiquement identifiables, sur une période allant du Néolithique à la période contemporaine, et cela sur une surface qui couvre l'ensemble de l'Europe et le pourtour du bassin méditerranéen. Il faut bien comprendre qu'il ne s'agit pas de la démarche d'une base « catalogue » qui va répertorier un objet par fiche, mais d'un dictionnaire de formes, qui recense sur une fiche les occurrences d'une même forme archéologique pour comprendre sa diffusion dans le temps et dans l'espace. Les fiches sont alimentées et enrichies en permanence par les différents auteurs, qui sont aujourd'hui 235 sur 6076 membres inscrits. Ce n'est pas un produit fini, mais un outil de travail quotidien et un référentiel typologique partagé en perpétuelle évolution. Pour donner un ordre de grandeur, actuellement, la base répertorie plus de 197 000 objets enregistrés sur plus de 21 000 fiches typologiques.
Transition : on va maintenant voir rapidement ensemble la manière dont sont classées ces données et quelle conséquence cela a pour l'élaboration d'un thésaurus.

2. Structure de classement adopté

La base utilise un classement dit « fonctionnel » du mobilier hérité du classement Syslat, où une catégorie d'objet correspond à une fonction primaire qui s'insère elle-même dans un domaine. Donc si on reprend cette arborescence, on a 6 domaines, dans lesquels on trouve 58 fonctions dans lesquelles sont répartis 636 catégories d'objets, que l'on appelle « codes » dans la base. Pour chaque catégorie d'objet, on va avoir des fiches qui correspondent aux types archéologiques qui auront été renseignés. A l'intérieur de chaque fiche, on trouve une description du type et un commentaire, ainsi que la liste des attestations, soit des différents points où les objets de ce type ont été reconnus, la bibliographie associée, une carte et des images des objets. La liste des attestations dépend évidemment des données saisies en base.
Bien sûr, l'usage et la fonction des objets peut toujours être discuté au cas par cas, surtout en cas d'utilisation secondaire ou de détournement. Ici, il ne s'agit pas de clore les discussions sur les usages et les fonctions des objets, mais surtout de disposer d'un classement à des fins statistiques, où une fonction appartient à un domaine. Le but est de pouvoir comparer les données de différents sites entre elles et de pouvoir analyser l'évolution fonctionnelle de sites stratifiés et de comparer les faciès mobilier de sites avec une grille de lecture commune.
Ce qu'il faut retenir, c'est que ce classement en arborescence en codes, fonctions et domaines ne peut être traduit directement en un thésaurus, car il ne décrit pas la nature des objets mais leur fonction.

3. Contexte de création du thésaurus et choix de la granulométrie

La première mouture du thésaurus Artefacts a été élaborée pendant 2 mois dans le cadre du programme ArteBib, dont ce n'était pas l'objectif premier, mais disons que cela a été fait en « bonus ». Les liens avec le thésaurus pivot Hyperthéseau n'ont pas vraiment abouti.

Quand on a commencé à travailler sur le thésaurus, la première question qui s'est posée est celle de la granulométrie. Il a été décidé de travailler à l'échelle des catégories d'objets, soit à l'époque 618 codes et non pas à l'échelle des formes archéologiques (+21 000 fiches), objectif qui n'était clairement pas atteignable dans le temps imparti. Durant ce temps imparti, on s'est attaché tout d'abord à élaborer une définition pour chacune des catégories d'objet puis à les hiérarchiser selon leur « nature » et non selon leur fonction.

L'articulation entre les codes et les fiches se fait par leur nom : les codes sont un des trois éléments qui servent à nommer les fiches typologiques.

Rapidement, les fiches sont nommées selon un trigramme qui correspond au code de la catégorie d'objet, souvent un moyen mnémotechnique, séparé par un tiret d'une série de 4 à 5 chiffres. Le premier, de 0 à 9 désigne une période, -3 étant le 2^{e} âge du Fer et -4 la période romaine. Les trois derniers chiffres correspondent à un numéro d'ordre de la fiche.

Transition : je vous propose maintenant de voir un peu plus en détails ces fameux « codes ».

4. Codes

Les codes sont des trigrammes qui renvoient à des catégories d'objets, qui peuvent être aussi bien à des objets complets que des parties d'objets plus complexes. Ce sont des subdivisions pratiques et d'usage, qui tiennent aussi compte de leur fragmentation et de la fréquence relative des objets, dans le sens où certaines parties d'objets qui sont sujettes à de multiples variations morphologiques font l'objet de codes distincts.

On va voir par la suite comment cela impacte l'arborescence du thésaurus.

Les codes suivent un développement qu'on pourrait qualifier d'organique, au fil des besoins et des travaux effectués par les auteurs.

On tente de limiter la multiplication des codes pour éviter de perdre les utilisateurs, surtout extérieurs, car c'est un reproche qui nous a été fait : il y a beaucoup de codes et il est difficile de tous les retenir. Il est également très souhaitable d'avoir une définition claire de chacune de ces catégories. Actuellement, ils sont recensés dans une page baptisée « Codes », qui liste les abréviations et le nom complet.

La création d'un nouveau code se fait sur proposition des auteurs et par après discussion, par consensus de l'équipe des administrateurs, on essaie d'en limiter le nombre. Pour vous donner un ordre d'idée : en trois ans nous avons ajouté dix-huit codes.

Transition : Je vais maintenant rapidement présenter le déroulement du travail de création du thésaurus.

5. Déroulement du travail

Après avoir défini le périmètre et la granulométrie des termes à traiter,

a) Rédaction des définitions

La première étape a été, pour chacun des codes, de rédiger une définition, soit à partir d'une autorité, comme le TLfI ou d'un dictionnaire, soit originale.

Cet aspect pourrait être discuté, mais la terminologie étant parfois très spécifique, il n'y a pas nécessairement de définition clairement publiée, surtout pour des parties d'objets ou des objets dont la fonction est encore peu documentée. Pour la terminologie très spécifique : il n'y aura pas nécessairement de définition satisfaisantes dans les dictionnaires sur les « appliques en T », les « tubes porte-amulettes », les « enclumes à denteler les faucilles », ou les « terriers artificiels à loirs »... Les définitions ont donc été discutées rédigées et revues collectivement

dans un tableur partagé. Dans ce tableur, nous avons également commencé à stocker les ark et les identifiants des thésaurus avec lesquels nous souhaitions nous aligner par la suite.

b) Hiérarchisation des concepts

La seconde étape, qui n'était pas la plus intuitive, a été la hiérarchisation des concepts selon la logique propre aux thésaurus, et le passage d'une terminologie basée sur la fonction à une terminologie basée sur la nature. Cette étape de hiérarchisation et de création de concepts génériques aux codes a été faite dans un logiciel permettant de créer des cartes mentales, ce qui permet de travailler l'arborescence de manière plus ergonomique, en pliant, affichant ou déplaçant les branches selon les besoins. Cela nous a également permis ultérieurement de signaler de manière visuelle à l'aide de couleurs ou de marqueurs l'état d'avancée du travail de saisie.

Une fois qu'on a disposé d'une première version de notre arborescence, nous avons fait appel à Emmanuelle Perrin-Touche du projet Hyperthéseau, puis à Magali Lugnot, Chargée de ressources documentaires de la MOM pour contrôler avec nous le travail effectué à ce stade, pour vérifier que l'arborescence choisie n'était pas en contradiction avec les règles d'élaboration des thésaurus, tant dans la terminologie que dans l'arborescence.

c) Saisie et alignements

La 3^{e} étape a consisté en la saisie manuelle du thésaurus, des définitions, des notes d'application et l'alignement des différents concepts dans Opentheso, avec Pactols, l'AAT et Wikidata, avec les conseils et l'aide de Miled Rousset.

d) Synonymes et labels alternatifs

Nous avons ensuite travaillé sur les synonymes et alt label, dans l'optique d'une utilisation du thésaurus comme future aide à la recherche dans Artefacts, afin de prendre en compte différentes appellations du même objet et pouvoir rédiger à terme l'utilisateur vers le vocabulaire employé dans les codes.

e) Gestion des traductions : thésaurus multilingue

L'étape suivante a été et est toujours d'actualité : la gestion de la traduction de nos codes dans les différentes langues utilisées par le site : si nous pouvons récupérer certaines traductions via les alignements avec Pactols, certains termes trop spécifiques n'y figurent pas : il s'agit d'une saisie manuelle. Les traductions, quand elles existent, étaient jusqu'à présent stockées dans une table de la base de données Artefacts. Un des chantiers, toujours en cours, consiste à faire relire et réviser ces termes spécialisés par des archéologues natifs ou à les trouver dans des articles spécialisés, les traductions récupérées automatiquement par le biais des alignements n'étant pas toujours satisfaisantes. Le travail est presque finalisé pour l'italien, langue dans laquelle les traductions étaient lacunaires avec la contribution de Veronica Groppo, de l'Universita Ca' Foscari Venezia.

A terme, on pourra faire traduire également les définitions, mais il a été décidé dans un premier temps de se focaliser sur les noms des codes. Avec l'ingénieur informaticien du projet, Bertrand David, il a été décidé que le thésaurus serait la version maître pour le stockage et la gestion des traductions.

Transition : je vais maintenant vous présenter quelques chiffres

6. Quelques chiffres

918 concepts hiérarchisés, dont 636 correspondent à nos codes

Transition : maintenant qu'on a vu tout ça, on peut se demander, oui, mais à quoi ça sert ? J'en ai un peu évoqué au fil de mon intervention, mais je vais maintenant décrire les usages du thésaurus envisagés, en cours d'intégration et à venir.

7. Usages du thésaurus : en cours et à venir

a) Une première utilisation, dont la mise en œuvre serait assez simple, est de pouvoir afficher des définitions au survol des différents codes ainsi que des notes d'application : sur la page « Codes » du site et sur les fiches.
b) On a vu tout à l'heure également, l'affichage et la gestion des noms de codes dans les langues prises en charge par le site, à savoir l'anglais, l'allemand, l'espagnol et l'italien.
c) Une troisième utilisation va être l'utilisation de l'arborescence définie dans le thésaurus pour améliorer l'expérience utilisateur au niveau de la recherche. Nous comptons utiliser un moteur de recherche type ElasticSearch pour la refonte des dispositifs de recherche simple et avancée sur le site.

Donner un exemple : au lieu de chercher juste un terme, on devrait pouvoir définir des opérateurs de recherche, proposer des termes synonymes, élargir ou affiner les résultats à l'aide de l'arborescence.

Nous pourrions également proposer une nouvelle expérience de navigation et donner la possibilité de parcourir les fiches typologiques selon l'arborescence du thésaurus, un peu comme ce qui se fait sur certains sites, à l'image du MIMO.

d) Enfin, le thésaurus s'inscrit bien sûr dans la démarche de partage et d'ouverture et de dépôt des données et de vocabulaire contrôlé avec l'utilisation des ark dans les métadonnées des fichiers que l'on compte déposer en entrepôts de données (nakala).

Ces différentes utilisations et intégrations sont soit en cours de mise en œuvre (affichage des définitions, stockage des traductions), soit encore à l'état de projet. Cela s'explique principalement par deux causes :

- une diminution du temps alloué au développement de la base, qui est passé d'un poste à plein temps et plus occasionnellement un mi-temps à un seul poste à mi-temps réparti entre trois bases de données, soit concrètement à des périodes travail en roulement entre les projets de 3 à 4 semaines tous les 2 à 3 mois ;
- cette mise en œuvre lente s'explique également par un besoin de mise à niveau et de refonte de l'architecture et du code de la base, qui existe en ligne depuis 2009, et qui nécessite des mises à niveau assez régulière pour pouvoir continuer d'évoluer au gré des projets.

Transition : pour terminer et en guise de conclusion, je vais évoquer les enrichissements pressentis pour le thésaurus.

8. Quels enrichissements pressentis ?

a) Nous continuons bien sûr d'ajouter les nouveaux codes au fur et à mesure de leur création.
b) Nous envisageons d'affiner la granulométrie pour améliorer la recherche, sans forcément encore rentrer dans le cœur de la typologie, en utilisant les concepts utilisés dans les « titres » des fiches, à l'aide d'un parsing et d'un comptage des termes.
   Cela devra permettre de compléter l'ensemble des parties constitutives des objets, sans créer de nouveau codes :

- Par exemple : on pourra ajouter « plateau de balance » et « fléau de balance » en terme spécifique partitif sous notre concept « balance », aux côtés du concept « curseur de balance », cela permettrait de coller encore davantage à la réalité archéologique de fragmentation des objets.

Pour ce faire, nous prévoyons d'utiliser la fonction « candidats » d'Opentheso, que nous n'avons pas utilisé jusqu'à présent, les discussions et consensus sur les définitions s'étant fait par des discussions en direct en amont de la saisie.

c) Plus facile à mettre en œuvre, de nouvelles collections matériaux, périodes, épigraphie

d) Une autre piste d’enrichissement, à plus long terme, pourrait être de la même façon de définir les termes utilisés dans les descriptions typologiques afin de normaliser le vocabulaire utilisé dans les descriptions d’objets. On pourrait ainsi à l’avenir imaginer un outil de recherche performant et très précis ainsi qu’une aide à la saisie lors de la rédaction des fiches, par exemple sur le vocabulaire descriptif des formes, des sections, des techniques ou des décors.

Je vous remercie pour votre attention.