Du thésaurus au graphe : un nouveau dispositif numérique basé sur le vocabulaire pour la création de corpus de données archéologiques

Guillaume Reich, Sébastien Durost

DU THÉSAURUS AU GRAPHE : UN NOUVEAU DISPOSITIF NUMÉRIQUE BASÉ SUR LE VOCABULAIRE POUR LA CRÉATION DE CORPUS DE DONNÉES ARCHÉOLOGIQUES

Guillaume REICH, Sébastien DUROST
(Bibracte EPCC)

Les archéologues collectent, identifient, décrivent, quantifient, puis comprennent et expliquent des réalités matérielles et leurs connexions pour écrire une histoire rationnelle des sociétés passées. L'actualisation de leurs discours s'inscrit dans un processus itératif permanent, qui nécessite de publier régulièrement. Bibracte souffre de plusieurs lacunes pour produire des données compatibles avec les standards d'interopérabilité attendus pour un partage en ligne : faiblesse de la contextualisation administrative (propriété intellectuelle, instanciation et conditions de partage), hétérogénéité des formats et des protocoles d'enregistrement, et utilisation d'un vocabulaire de description qui n'est, ni totalement contrôlé lors de la saisie, ni normalisé par référence à un vocabulaire partagé. En outre, une majorité de la documentation est produite hors de Bibracte, ce qui oblige à un effort permanent de rétro-documentation, source potentielle d'erreurs et de disparité dans la qualité générale des données.

Le projet *Bibracte Ville Ouverte* (2021-2022), porté par la Maison des Sciences de l'Homme et de l'Environnement Claude Nicolas Ledoux (Besançon) et financé par le dispositif *CollEx Persée* de soutien à la valorisation des données des sciences humaines par les technologies numériques, a permis d'intégrer le catalogue de la bibliothèque de Bibracte (76676 notices) au *catalogue collectif indexé* (CCI) Frantiq, qui agrège les fichiers d'une quarantaine de bibliothèques françaises spécialisées en archéologie et sciences de l'Antiquité. Cette importation enrichit considérablement le catalogue Frantiq, dont Bibracte devient le plus important contributeur avec un fonds de protohistoire largement européen et 50988 nouvelles notices bibliographiques correspondant à des dépouillements d'ouvrages (actes de colloques et recueils d'articles).

Pour organiser les mots-clés utilisés dans l'indexation des notices du CCI, l'équipe de Frantiq a créé et utilise le thésaurus PACTOLS, géré avec le logiciel *Opentheso*, développé par Miled Rousset (Maison de l'Orient et de la Méditerranée, Lyon) et respectant la norme ISO 25964. Bibracte participe à l'enrichissement sémantique de PACTOLS, mais dans beaucoup de cas, la finesse terminologique des partenaires du programme de recherche de Bibracte est beaucoup trop importante pour être exposée directement dans l'arborescence de PACTOLS. Ainsi, pour prendre l'exemple de la céramique à Bibracte, un céramologue doit pouvoir retrouver facilement une catégorie de pâte, par exemple «PGFINLF», une forme, comme «assiette» ou un type, tel que «A15». Ce vocabulaire dépasse généralement les ambitions d'un thésaurus d'indexation bibliographique (mots-clés). S'il est parfois commun, sa définition est alors plus circonscrite. Le souci, c'est que sans permettre de rendre compte de cette spécificité du vocabulaire, la pertinence des requêtes dans le CCI est plus réduite pour des spécialistes (beaucoup trop de «bruit» dans les résultats, donc énormément de tri).

Le choix a donc été fait d'externaliser un thésaurus spécifique à Bibracte (*Bibracte_Thesaurus*), respectant la même structuration que PACTOLS sur lequel il est aligné, géré sur *Opentheso*, reprenant ses normes, mais enrichi des besoins spécifiques des partenaires du programme de recherche (Durost, Reich, Girard, à paraître). La branche qui concerne la céramique est déjà disponible, et publiée simultanément avec le volume *La vaisselle céramique de Bibracte : de l'identification à l'analyse* (Barrier, Luginbühl 2021). Près de 700 concepts ont été formalisés en reprenant les vocabulaires et les définitions des auteurs. Pour prendre l'exemple du type «assiette A15», chaque concept est exposé suivant une même logique (Fig. 1).

Un terme préféré, libellé ou *prefLabel* : «assiette A15», auquel a été ajouté un millésime, c'est-à-dire un référentiel bibliographique permettant de connaître le cadre de production de ce vocabulaire ; en l'occurrence la publication «BARRIER, LUGINBÜHL 2021». Ce procédé, inhabituel pour les collègues documentalistes, mais plus en conformité avec les usages des archéologues,

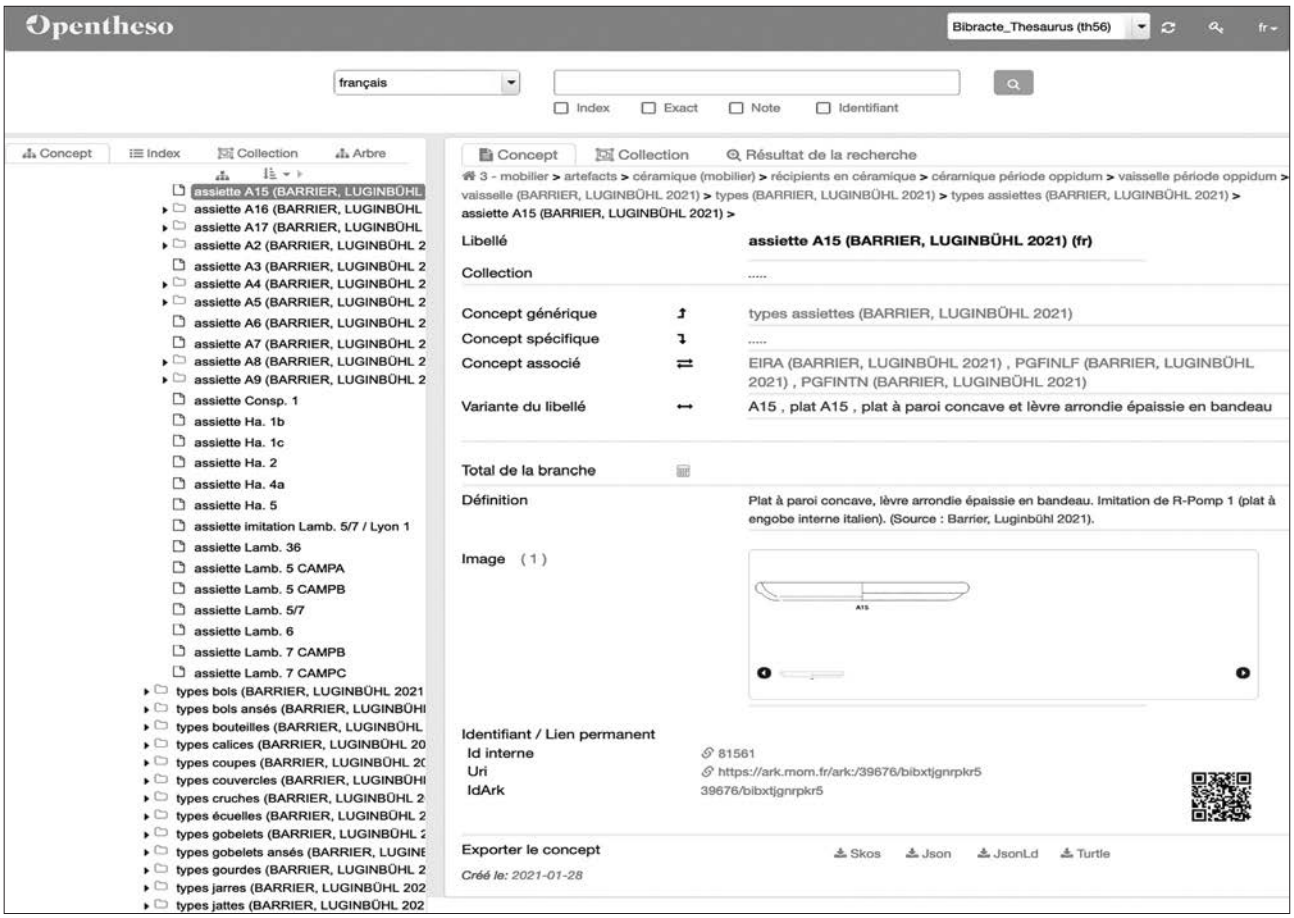

Fig. 1 : Un exemple de vocabulaire spécialisé dans le thésaurus : "assiette A15 (BARRIER, LUGINBÜHL 2021)"

ouvre des voies importantes en matière d'exposition du vocabulaire, tout en en respectant la paternité.

Le concept se situe dans une arborescence, dans laquelle il est possible de naviguer : un mot est plus générique ou plus spécifique qu'un autre ; il peut aussi être associé à d'autres vocabulaires. Ici, le type «assiette A15» est associé à la catégorie de pâte «PGFINLF».

«PGFINLF» n'étant pas forcément explicite, un synonyme ou variante du libellé est proposé : «céramique à pâte grise fine et surface lissée fumigée».

Suit une définition sourcée, essentielle pour les archéologues, reprenant les termes exacts de la publication.

La bibliographie mobilisée par les auteurs est éditorialisée : «Barral 1994» pointe vers la notice correspondante du CCI ; «Barral (Ph.)» renvoie sur la notice *IdRef* de l'auteur et le titre «Céramique indigène…» permet de fournir un lien vers la ressource documentaire en ligne, si elle s'y trouve.

Participant de la définition, mais en ressource externe directement téléchargeable depuis un entrepôt *Nakala*, il est possible d'atteindre des documents iconographiques. Les conditions de partage et de réutilisation sont précisées dans les métadonnées du document. Ici, les images sont enregistrées en *.svg*, un format lisible par des logiciels propriétaires type *Illustrator*, mais aussi par son équivalent libre *Inkscape*.

Le concept bénéficie d'un URI, c'est-à-dire une adresse pérenne, permettant de toujours l'atteindre.

Le référentiel « vaisselle céramique (BARRIER, LUGINBÜHL 2021) (fr) », transposition numérique d'une ressource bibliographique publiée, permet l'accès libre et permanent à la définition et à la datation millésimées des 53 catégories, des 25 formes et des 423 types de céramiques répertoriés à ce jour sur le site pour la période d'occupation de l'*oppidum* de Bibracte. Il témoigne d'une réalité archéologique matérielle riche (plus de 300000 éléments enregistrés dans la base de données du site) et génère des concepts de granularité fine, équilibrés selon une logique construite par l'usage. Il reflète la problématique initiale des auteurs : fournir les clés de l'identification d'un tesson grâce à l'observation et à l'expérience du céramologue, du plus spécifique (le tesson) au plus générique (un groupe de tessons partageant les mêmes propriétés), dans le cadre de l'élaboration d'une classification morphotypologique. Par la suite, dans un contexte informationnel, les concepts seront présentés du plus générique au plus spécifique : « céramique tournée (BARRIER,

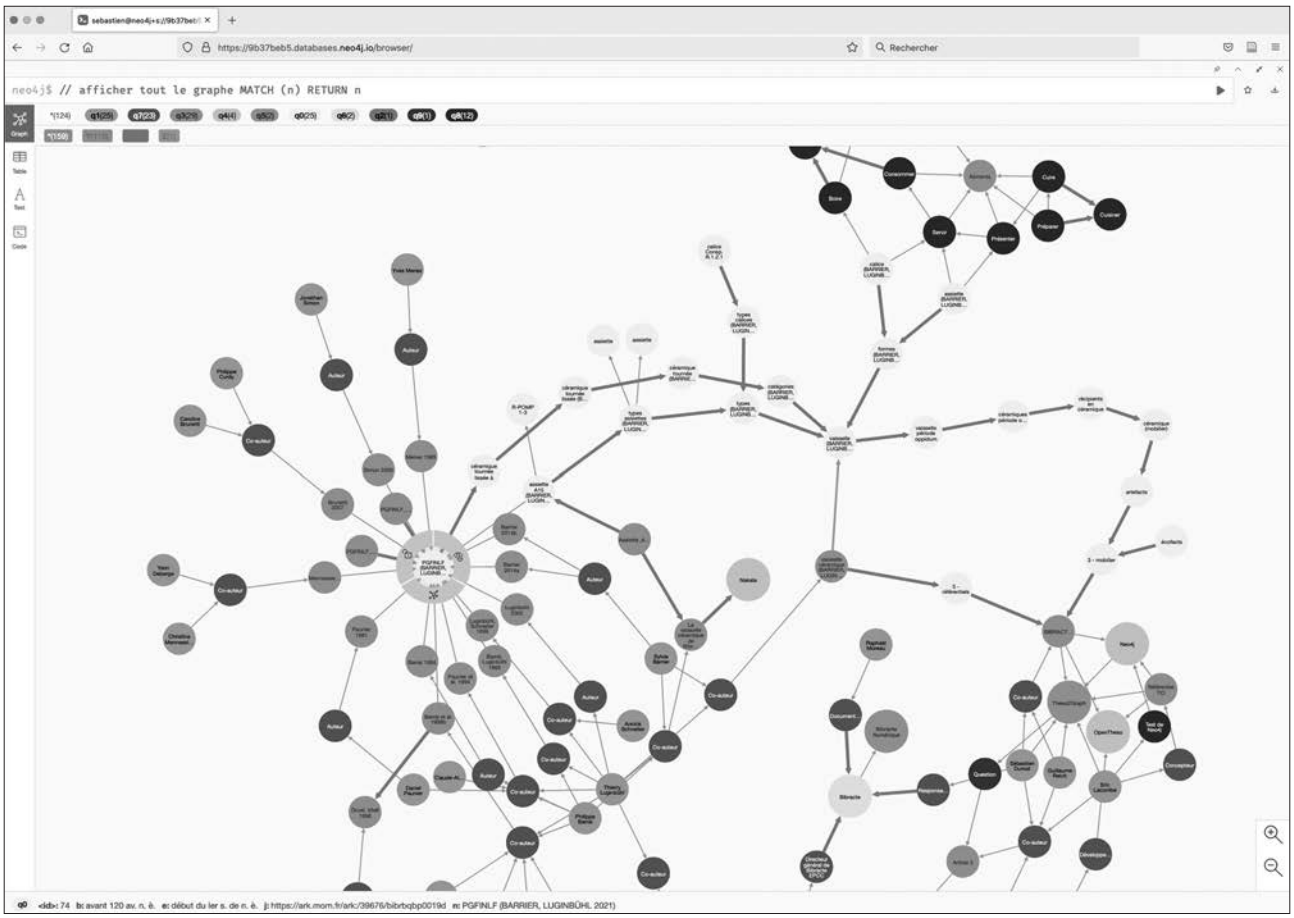

Fig. 2 : Contextualisation en graphe de quelques concepts céramologiques modélisés en thésaurus

LUGINBÜHL 2021) (fr) » > « céramique tournée lissée (BARRIER, LUGINBÜHL 2021) (fr) » > « céramique lissée tournée à pâte sombre/grise (BARRIER, LUGINBÜHL 2021) (fr) » > « PGFINLF (BARRIER, LUGINBÜHL 2021) (fr) » ; les définitions, construites sur une réalité matérielle tangible, garantissent la compréhension du libellé (*prefLabel* et *altLabel*) de chacun des concepts.

La norme ISO 25964 fournit un cadre structurel utile à l'archéologue pour créer son vocabulaire, organiser une approche pédagogique de sa nomenclature et l'outiller pour la comparaison de sa terminologie avec les champs lexicaux développés par d'autres chercheurs. Les règles et les contraintes formelles de la norme, comme l'impossibilité de produire deux termes strictement identiques ou la hiérarchisation depuis une nomenclature générique vers un vocabulaire de fine granularité, deviennent des outils efficaces pour traquer et décrire les critères implicites du raisonnement et progresser rigoureusement vers une approche plus systématique et plus explicite. Par ailleurs, la vue arborescente du thésaurus, des termes les plus génériques vers les concepts les plus spécifiques, favorise l'exploration du sujet, notamment lorsque l'archéologue cherche à s'imprégner des univers sémantiques et des paradigmes de ses collègues. Cependant, la spécificité de l'intentionnalité de ce vocabulaire (dans ce cas précis l'identification d'un tesson) a conduit à s'écarter de l'application usuelle (documentaire) de la norme ISO 25964, en privilégiant la définition plutôt que les libellés (*prefLabel* et *altLabel*) d'un concept. Cette approche, dans laquelle le terme isolé (décontextualisé) n'a pas sa place, a pour conséquence logique l'évolutivité du thésaurus par l'enrichissement itératif de ses définitions, induisant potentiellement le déplacement du concept dans l'arborescence. L'outil thésaurus rend ici compte des connaissances produites par la recherche et nécessaires à l'activité scientifique.

Grâce à son articulation avec l'entrepôt de données de Bibracte (sur *Nakala*), le thésaurus a été pensé comme un portail d'accès aux ressources produites par les partenaires du programme de recherche. Ce lien a été rendu possible, d'une part, par les métadonnées associées aux documents déposés dans *Nakala*, qui reprennent des termes du thésaurus ; d'autre part par les adresses Internet pérennes (*Ark*, *Handle* et/ou *DOI*) associées aux libellés du thésaurus et aux fichiers partagés sur l'entrepôt, qui garantissent de part et d'autre le maintien de l'interconnexion entre eux. Ce processus fondé sur *Opentheso* et *Nakala* organise une chaîne de partage des données, liant matérialités archéologiques, données numériques brutes et publications imprimées. Mais la réutilisation des données pose la question de leur interopérabilité, c'est-à-dire la possibilité qu'elles soient exploitées par le plus grand nombre et parta-

gées dans un format libre et standardisé. Une preuve de concept a porté sur une sélection de vaisselle céramique issue de la fouille de la *domus* PC1, dont l'inventaire déposé sur *Nakala* est organisé dans un tableur structuré en *.xml* pour s'affranchir des formats propriétaires, comme *FileMaker Pro* ou *Excel*, et garantir leur lisibilité par des logiciels libres, en renvoyant grâce à des hyperliens à des concepts du volume 31 de la collection *Bibracte* définis dans le thésaurus.

Ces recherches montrent que le thésaurus possède intrinsèquement les qualités d'un outil heuristique capable d'harmoniser les pratiques et les jeux de données sans les soumettre à un consensus rigide, tout en offrant les conditions d'une analyse et d'une critique de la modélisation du raisonnement archéologique sous forme de graphes (**Fig. 2**). Cette approche testée à Bibracte semble pertinente pour améliorer la qualité des recherches archéologiques en identifiant les éléments implicites dans les raisonnements exposés dans les référentiels publiés. Elle apparaît aussi comme un moyen efficace de rechercher l'explicite et de clarifier les points de vue des chercheurs. Au final, le thésaurus ouvre la voie à un changement de paradigme.

Bibliographie

Barrier S., Luginbühl Th., 2021. *La vaisselle céramique de Bibracte. De l'identification à l'analyse*. Glux-en-Glenne, Bibracte, 318 p. (*Bibracte*, 31).

Durost S., Reich G., Girard J.-P., à paraître. Terminologies, modèles de données archéologiques et thésaurus documentaires : réflexions à partir d'une typologie de céramique. 14 p.
https://hal.archives-ouvertes.fr/hal-03278684