Thésaurus et interopérabilité des données archéologiques : le projet HyperThesau

Emmanuelle Perrin

▶ **To cite this version:**

Emmanuelle Perrin. Thésaurus et interopérabilité des données archéologiques : le projet HyperThesau. Humanités numériques, 2021, Varia, 4, 10.4000/revuehn.2384 . hal-03500465

HAL Id: hal-03500465
https://hal.science/hal-03500465

Submitted on 30 Mar 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Humanités numériques

4 | 2021
Varia

Thésaurus et interopérabilité des données archéologiques : le projet *HyperThesau*

Thesaurus and Archaeological Data Interoperability: the HyperThesau *Project*

Emmanuelle Perrin

Édition électronique
URL : https://journals.openedition.org/revuehn/2384
DOI : 10.4000/revuehn.2384
ISSN : 2736-2337

Éditeur
Humanistica

Référence électronique
Emmanuelle Perrin, « Thésaurus et interopérabilité des données archéologiques : le projet *HyperThesau* », *Humanités numériques* [En ligne], 4 | 2021, mis en ligne le 01 décembre 2021, consulté le 23 décembre 2021. URL : http://journals.openedition.org/revuehn/2384 ; DOI : https://doi.org/10.4000/revuehn.2384

Les contenus de la revue *Humanités numériques* sont mis à disposition selon les termes de la Licence Creative Commons Attribution 4.0 International.

humanités
numériques

4 | 2021
Varia

Thésaurus et interopérabilité des données archéologiques : le projet *HyperThesau*

Thesaurus and Archaeological Data Interoperability: the HyperThesau Project

Emmanuelle Perrin

Résumés

Bâtir l'interopérabilité des données archéologiques avec un vocabulaire normalisé et partagé est la perspective adoptée par le projet *HyperThesau* (Labex IMU [Intelligences des mondes urbains], université de Lyon). Comme preuve de concept est amorcé un thésaurus pivot, qui permet d'aligner les concepts de l'archéologie sur les systèmes internationaux d'organisation des connaissances, avec les technologies du Web sémantique. Ce thésaurus doit servir à la description, à l'enregistrement, à l'indexation et à l'extraction de données au sein d'un vaste écosystème numérique : métadonnées, bases de données, catalogues de bibliothèques, bibliothèques numériques, signalement des dépôts sur différentes plateformes.

La problématique du traitement informatique et documentaire des données archéologiques implique que le langage devienne lui-même un objet d'étude. La méthode mise en œuvre pour construire ce thésaurus a consisté à recueillir, à traiter et à structurer le vocabulaire de cette discipline en délogeant l'implicite et en décloisonnant les classifications archéologiques. Un thésaurus peut être un outil de réconciliation entre recherche et sciences de l'information. Il est aussi porteur de réflexions méthodologiques, ouvrant sur des perspectives historiographiques intéressantes.

The *HyperThesau* project (IMU Labex, University of Lyon) aims at building archaeological data interoperability through a standardised, shared vocabulary. As proof of concept, it initiates a pivot thesaurus, which maps the concepts of archeology on to international knowledge

organisation systems, using semantic Web technologies. This thesaurus should be used for data processing, recording, crawling and retrieval within a vast digital ecosystem: metadata, databases, library catalogs, digital libraries and platforms.

The computer and documentary processing of archaeological data implies that language itself becomes a study topic. The method used to build this thesaurus consisted in collecting, processing and structuring archaeological vocabulary by trying to determine the underlying meaning of the terms and building bridges between the classifications used by archaeologists and current knowledge organisation systems. A thesaurus can be a tool for reconciling research and information science. It also feeds methodological reflections, opening interesting historiographical perspectives.

Entrées d'index

MOTS-CLÉS : archéologie, thésaurus, gestion des données de la recherche, interopérabilité, Web sémantique

KEYWORDS: archaeology, research data management, thesaurus, interoperability, semantic Web

« Ce n'est pas une question de choix. Il n'y a que ce mot. C'est comme dans les mots croisés, il ne faut pas choisir, il n'y a que le mot qui correspond à la définition. » (De Lucca 2020)

« La définition d'un mot est ce qui permet de ne pas s'en servir. » (Ginouvès et Guimier-Sorbets 1978)

1 La dimension heuristique du partage des données intéresse l'archéologie, science d'érudition dans laquelle les questions de documentation et la démarche comparative occupent une place centrale. Les possibilités de « découvrabilité » et d'agrégation dépendent de la manière dont les données sont décrites et de la façon dont données et descriptions correspondent à un système d'organisation des connaissances (Rabinowitz *et al.* 2016, 49). La construction de l'interopérabilité des données archéologiques au travers d'un vocabulaire normalisé et partagé est la perspective adoptée par le projet *HyperThesau* (Hyper thésaurus et lacs de données : fouiller la ville et ses archives archéologiques), financé par le Labex « Intelligences des mondes urbains » de l'université de Lyon de 2019 à 2020. Cette approche repose, d'une part, sur la création d'une architecture issue de l'informatique décisionnelle et des big data – le « lac de données » (Darmont *et al.* 2020) – et, d'autre part, sur la composition d'un thésaurus pivot, permettant d'aligner les concepts de l'archéologie sur les systèmes internationaux d'organisation des connaissances et les grands référentiels publiés sur le Web sémantique par la communauté des bibliothèques. Ce thésaurus doit servir à la description, à l'enregistrement, à l'indexation et à l'extraction de données au sein d'un vaste

écosystème numérique : métadonnées, bases de données, catalogues de bibliothèques, bibliothèques numériques, signalement des dépôts sur différentes plateformes.

2 Sous la coordination scientifique de Marie-Odile Rousset (archéologue, CNRS, UMR 5133 Archéorient), ce projet a réuni un large éventail d'acteurs de l'archéologie française et européenne, dans un consortium interdisciplinaire (archéologie, informatique, sciences de la documentation, écologie et science participative) : les laboratoires d'archéologie Archéorient (UMR 5133) et ArAr (UMR 5138), l'Équipe de recherche en ingénierie des connaissances (EA 3083 ERIC), le Centre d'écologie et des sciences de la conservation (UMR 7204 CESCO), la fédération de recherche de la Maison de l'Orient et de la Méditerranée – Jean Pouilloux (FR 3747), la plateforme de numérisation, d'enrichissement et de diffusion Persée (UMS 3602), l'université autonome de Barcelone, le Centre archéologique européen Bibracte EPCC, le service archéologique de la ville de Lyon, les musées d'archéologie de Catalogne (site d'Ullastret) et la société d'archéologie préventive Archéodunum.

3 Le projet *HyperThesau* s'inscrit tout à la fois dans le contexte de l'ouverture des données de la recherche, des humanités numériques et du Web sémantique, dont les aspects technologiques ne vont pas sans induire des questionnements méthodologiques et épistémologiques pour les sciences humaines et sociales. Dans l'objectif de construire une science « plus cumulative, plus fortement étayée par des données, plus transparente, plus rapide et d'accès plus universel » (MESRI 2018), l'ouverture et le partage des données de la recherche sont devenus des enjeux majeurs, dont se sont saisis les réglementations de l'Union européenne, des administrations et des établissements publics [1]. Au centre de ce dispositif, la mise en œuvre des principes FAIR [2] permet de rendre les données faciles à trouver, accessibles, interopérables et réutilisables. Plus précisément, la question de l'interopérabilité se définit comme la capacité de deux ou plusieurs systèmes à échanger des informations et à utiliser les informations qui ont été échangées avec une perte minimale de contenu [3]. On peut ainsi distinguer quatre niveaux d'interopérabilité : une interopérabilité système au niveau du matériel et du système d'exploitation ; une interopérabilité syntaxique liée au format des données et à leur encodage ; une interopérabilité structurelle fondée sur le modèle des données ; et enfin une interopérabilité sémantique, au niveau terminologique, que garantit l'emploi de vocabulaires contrôlés et partagés (Zeng 2019). Selon les principes FAIR, les données doivent être décrites à l'aide d'un vocabulaire contrôlé respectant lui-même les principes FAIR. Il doit être facile à trouver grâce à un identifiant pérenne et unique, documenté et lisible par les machines [4] pour être publié sur le Web des données.

4 Au travers des caractéristiques des données archéologiques, d'un état de l'art sur les langages documentaires et de la méthode mise en œuvre pour structurer le vocabulaire de l'archéologie, cette contribution montre comment la problématique du traitement documentaire des données archéologiques implique que le langage devienne lui-même un objet d'étude. La perspective adoptée révèle également qu'un thésaurus peut être un outil de réconciliation entre recherche et sciences de l'infor-

mation et qu'il alimente des réflexions méthodologiques, qui ouvrent d'intéressantes perspectives historiographiques sur la construction du vocabulaire d'une discipline.

Les données de l'archéologie entre observation et interprétation

5 L'archéologie est une « grande productrice de données, en volume et en variété typologique » (Rabot 2017). Des activités diverses caractérisent tout d'abord le métier de l'archéologie : la documentation ; la prospection des zones archéologiques ; les fouilles et les prélèvements sur un site archéologique ; les analyses physiques, chimiques ou biologiques ; l'étude, la restauration et la conservation des objets archéologiques ; l'archivage ; la publication et la diffusion des résultats ; la muséographie ; et enfin l'administration du patrimoine archéologique (Djindjian 2013, 71). La chaîne opératoire de l'archéologie, du terrain jusqu'à la publication des données, fait également intervenir de nombreuses disciplines, qui impliquent une grande variété dans les méthodes d'acquisition des données : sciences physiques et chimiques, sciences de la Terre et de la vie, mathématiques, géographie, ethnologie, épigraphie, histoire de l'art ou architecture entre autres. À ce caractère pluridisciplinaire, s'ajoutent différents types et formats de données tels que textes, bases de données et tableurs, images raster et vectorielles, données spatiales, enregistrements audiovisuels, etc.

6 Les fouilles opèrent une « destruction savante » des niveaux d'occupation archéologique et des structures à mesure qu'elles sont démontées (Kaeser 2015, 2). C'est pourquoi le « contexte archéologique » se trouve précisément documenté. Cette notion fondamentale désigne l'ensemble des informations associées à un site et à ses vestiges. La localisation des découvertes, leur niveau archéologique et géologique d'origine, les relations stratigraphiques entre les couches, les associations entre vestiges sont archivés grâce à différentes techniques de relevé : fiche d'enregistrement papier, plan, coupe, croquis, photographies, relevé 3D, etc. Avec les moulages, les empreintes et les prélèvements de matériaux naturels et de nature biologique, l'ensemble de ces documents forme les archives de fouille ou la documentation scientifique d'une opération archéologique. La réglementation de l'archéologie distingue cette documentation scientifique du mobilier archéologique qui, au sens strict, correspond à la notion d'artefacts, c'est-à-dire les objets transformés par l'activité humaine [5].

7 Définies comme « tout élément trouvé dans la fouille », les données produites par l'archéologie sont donc de natures variées : « mobilier, objet, couches, ainsi que les liens entre ces éléments, des observations et des interprétations [que l'archéologue] peut faire et aussi tout document permettant de garder une trace visuelle de ces objets, de ces couches et des liens comme des photographies ou des dessins » (Chaillou 2003, 19). D'aucuns vont différencier les données primaires qui proviennent de l'observation directe du site et de la fouille, des données traitées ou dérivées, résultats d'études menées sur les données primaires, assemblage de plans, datation et regroupement des données en faits, structures ou entités archéologiques (Chaillou 2003, 21).

8 La problématique du traitement informatique et documentaire de ces données implique que le langage devienne lui-même un objet d'étude et que le discours soit considéré comme une « production scientifique » (Dufal 2010). Or, dans l'archéologie, le langage possède un « statut à la fois central et distinct » (Plutniak 2017b, 10). Les procédures d'analyse de cette discipline portent en effet sur les « propriétés non discursives des objets » (Plutniak 2017b, 11). Son positionnement du côté de la pratique et de la culture matérielle induit une forme de « naturalisation », comme l'observe Blaise Dufal dans un article sur la position singulière qu'occupe l'archéologie dans le champ des sciences humaines et sociales, intitulé « L'archéologie enfermée dehors » (Dufal 2010). La pratique des archéologues réserve une large place à la description minutieuse des données si bien que l'on peut considérer que, pour toute une partie de leur activité, ils ne travaillent pas « sur des objets matériels mais sur leur représentation linguistique » (Ginouvès et Guimier-Sorbets 1978, 14). Il y a, dans les opérations d'observation ou de description, une forme souvent trompeuse « d'invisibilisation » du travail autour des données (Denis et Goëta 2013 ; Malingre *et al.* 2019). Pourtant les observations ne sont jamais directes et complètes [6], ni une simple lecture de la réalité : la description d'une fouille ne consiste pas « à lire comme les pages d'un livre, pages écrites une fois pour toutes et que nous devrions seulement déchiffrer (pages d'ailleurs que le fouilleur arracherait en même temps – ce qui est la partie la plus exacte de la métaphore) » (Ginouvès et Guimier-Sorbets 1978, 10). Dans la démarche archéologique, la phase descriptive et la phase interprétative entretiennent un rapport dialectique et c'est l'interprétation – comme système scientifique de référence – qui rend l'objet observable. C'est pourquoi, dans ses *Essais d'analyse du discours archéologique*, Jean-Claude Gardin incite à subordonner toute description à une visée clairement définie : la formation et la validation d'une hypothèse (Gardin et Lagrange 1975, 23). Par ailleurs, en dépit d'un usage pionnier et massif des bases de données en archéologie – « secteur scientifique le plus riche dans ce domaine : 12 % des bases tous secteurs confondus soit ¼ des sciences humaines et sociales à lui seul », dès les années 1980 (Cacaly 1989, 148) –, leur multiplication et leur construction à l'échelle du chercheur ou du petit groupe (Ginouvès 1989), « sans réel intérêt pour l'échange avec d'autres bases de données existantes », auraient « isolé une partie des archéologues des grands courants documentaires » (Chaillou 2003, 24-25).

9 Partant du constat que « les descriptions traditionnelles écrites dans le "langage naturel" des archéologues [étaient] entachées de synonymies, polysémies, irrégularités de toutes sortes, [utilisaient] des concepts non définis, et [donnaient] des informations diversement incomplètes » (Ginouvès et Guimier-Sorbets 1978, 14), des chercheurs pionniers dans l'application du calcul et de l'informatique à l'archéologie, se sont, dès les années 1960-1970, intéressés aux langages et aux codes descriptifs. Il faut citer les publications de Jean-Claude Gardin (1925-2013) et de son équipe du Centre d'analyse documentaire pour l'archéologie, en différents domaines : cylindres orientaux (Digard 1975), monnaie (Le Rider 1975), monuments religieux (Nivelle 1975), monuments civils (Lagrange 1975), formes des poteries (Gardin 1976), analyse des textes orientaux (Salomé 1978), ornements (Gardin 1978), iconographie des vases grecs (Salomé 1980). Autre exemple, la description de l'industrie

lithique a bénéficié des travaux d'André Leroi-Gourhan *et al.* (1966), de Michel Brézillon (1983) et de la « typologie analytique » de Georges Laplace (1918-2004), qui associe un lexique descriptif, des indices métriques et une méthode pour leur représentation graphique (Plutniak 2017a, 4). Dans les années 1970, René Ginouvès (1926-1994) s'est, quant à lui, beaucoup intéressé aux bases de données et à l'application de l'informatique à l'archéologie (Ginouvès 1971, 1989 ; Ginouvès et Guimier-Sorbets 1978).

10 Le déploiement des technologies numériques et du Web sémantique, la question de l'interopérabilité des données archéologiques redonnent à ces travaux une place centrale dans l'élaboration « d'ontologies et de standards pour l'archéologie » (Djindjian 2013, 176). Seul un langage documentaire est susceptible d'apporter une « représentation stable de la réalité » (Ginouvès et Guimier-Sorbets 1978, 12) et de faire correspondre les termes de l'indexation avec ceux de l'interrogation des données.

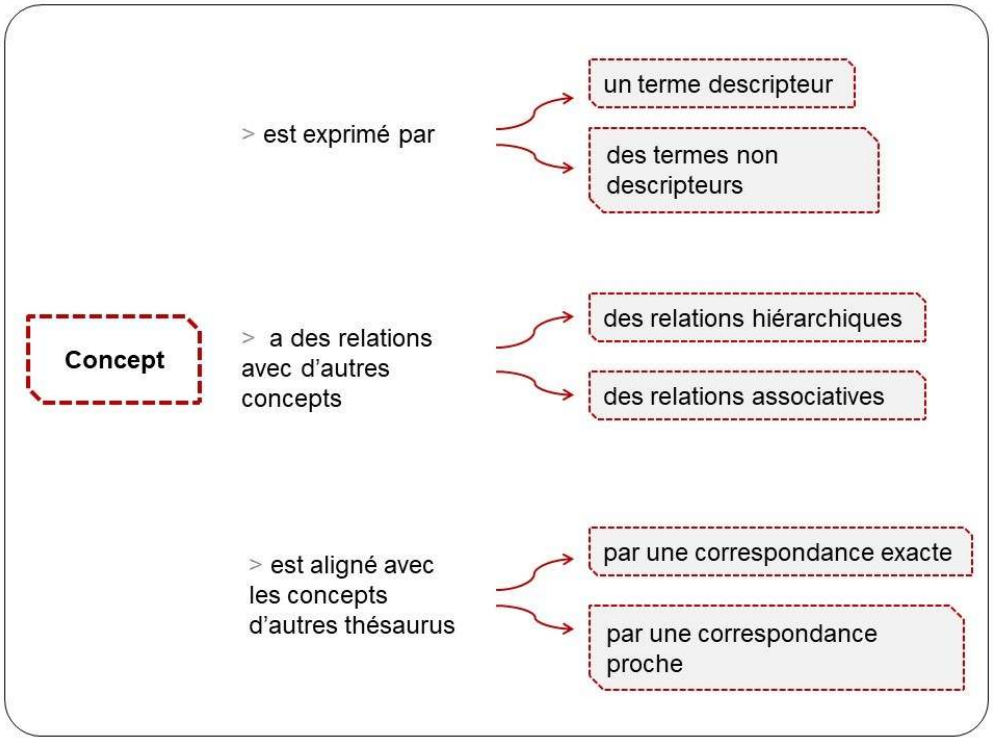
Interopérabilité et Web sémantique : le rôle des langages documentaires

11 Un langage documentaire est un vocabulaire contrôlé construit pour répondre à l'équivoque du langage naturel et aux problèmes que posent la polysémie et l'homonymie. Il sert à indexer et à décrire le contenu d'une ressource et vise à améliorer la recherche documentaire en augmentant le taux de rappel de documents pertinents au regard d'une requête. « L'objectif est de parvenir à un vocabulaire exact dans son contenu, homogène dans sa forme et univoque dans son utilisation (où un descripteur = un concept et un concept = un descripteur), tout en multipliant les points d'accès aux formes retenues à partir de formes rejetées grâce à des renvois d'équivalence » (Mingam 2015, 125). On distingue trois grands types de langages documentaires : les classifications, les listes d'autorité ou de vedettes matière et les thésaurus. Les classifications, comme la classification décimale de Dewey, proposent une organisation hiérarchique des concepts. Les listes d'autorité ou de vedettes sont des listes de termes normalisés, qui doivent être obligatoirement utilisés dans le catalogage ou l'indexation, pour décrire le contenu des documents. Enfin, un thésaurus structure un vocabulaire contrôlé en décrivant l'ensemble des relations sémantiques qui relient des concepts (figure 1).

12 Le caractère structuré d'un thésaurus est généralement méconnu : une simple liste de mots-clés « à plat » ou un module de « gestion de thésaurus » dans une application informatique sont fréquemment considérés comme tels. De même, la notion de « vocabulaire contrôlé » renvoie souvent à une liste fermée de mots mise au point dans le cadre d'un projet sans relation avec un référentiel documentaire extérieur. La construction d'un thésaurus obéit à une norme [7]. Il faut distinguer entre le concept, « représentation mentale », et le terme, expression linguistique d'un concept dans une langue donnée. Un concept est exprimé par un terme descripteur ou préféré et des termes alternatifs. Il a des relations hiérarchiques et associatives avec d'autres concepts. Il est aligné

avec les concepts d'autres thésaurus par des correspondances exactes ou proches. Il est défini par une note d'application, brève explication précisant les modalités d'emploi du descripteur.

FIGURE 1. LA STRUCTURE D'UN THÉSAURUS



La structure d'un thésaurus permet d'exprimer l'ensemble des relations sémantiques d'un concept.

Schéma : E. Perrin

13 Le terme préféré (ou retenu ou préférentiel) doit décrire de manière univoque un concept. Un thésaurus n'admet donc qu'un seul terme descripteur par concept et par langue. Un même terme descripteur ne peut pas désigner plusieurs concepts. En cas d'homonymie, il convient de désambiguïser le champ d'application du terme en ajoutant un qualificatif entre parenthèses. Le terme descripteur et les termes non descripteurs (ou rejetés, ou non préférentiels) qui représentent un même concept sont liés entre eux par une relation d'équivalence. Elle permet le contrôle des synonymes. La fonction première des descripteurs dans un langage documentaire est de donner accès au document, « afin que l'utilisateur final puisse ensuite le trouver à partir des termes de sa requête, les formes rejetées acquièrent une importance égale à celle de la forme retenue, au sein d'un ensemble indissociable que l'on peut appeler la "grappe" ou le "bouquet" terminologique » (Mingam 2015, 126). Selon les normes de construction d'un thésaurus, sont considérés comme des synonymes, le nom complet et son abréviation ou sigle, le nom courant et le nom scientifique, le nom scientifique et le nom de marque, les termes d'origines linguistiques différentes, les variantes orthographiques d'un même terme, les appellations anciennes et modernes.

14 La relation hiérarchique est à la fois celle qui est le plus spontanément associée à un thésaurus et celle qui est la plus malmenée. Une relation hiérarchique (ou générique-spécifique ou parents-enfants ou *broader-narrower*) doit logiquement s'exprimer sur le modèle de la relation genre-espèce. Elle relie des concepts de même catégorie – entités, activi-

tés, agents, propriétés – des plus généraux ou abstraits aux plus précis pour permettre une navigation verticale. Les relations hiérarchiques incluent également des relations partitives (tout-partie) uniquement si elles concernent les systèmes et organes du corps, les lieux géographiques, les organisations sociales hiérarchiques et les disciplines et leurs spécialités. Les relations hiérarchiques admettent enfin les relations classe-instances. Cette relation caractérise le lien entre une catégorie ou une classe désignée par un nom commun et une instance de cette catégorie, représentée par un nom propre. Par exemple, « Alpes » est une instance de la classe « Montagnes ». L'application stricte de la relation hiérarchique semble poser de nombreux problèmes lors de la construction des thésaurus et incite à leur révision. Les « relations hiérarchiques bancales (A ne peut plus avoir pour termes génériques B et C si B a lui-même pour terme générique C) » (Mingam 2015, 126) ont ainsi été bannies des autorités Rameau [8] au profit des relations associatives. De même, la consolidation de la structure du thésaurus PACTOLS [9] passe par le respect « des règles rigoureuses dans les relations. Par exemple, la relation hiérarchique « « est un/est un type de ou « est une partie de » (éperon barré EST UN TYPE DE fortification ; Anubis EST UNE divinité égyptienne ; doigt EST UNE PARTIE DE corps) [10] ».

15 La relation associative décrit une relation horizontale entre deux concepts. Elle permet d'élargir une requête au champ sémantique d'un concept. Pour associer deux concepts, la norme veut que l'un soit un élément nécessaire à la définition de l'autre. Cette relation décrit la cause et l'effet (infection et maladie infectieuse), le tout et sa partie (livre et reliure), l'action et l'agent ou l'instrument (prospection électrique et résistivimètre), l'action et le produit de cette action (édition et livre), l'action et l'objet qui subit l'action (analyse de données et données), l'action et le lieu de l'action (enseignement et école), une science et son objet (paléontologie et fossiles), l'objet ou l'action et sa propriété (poison et toxicité), un objet ou une action et son rôle ou son but (citerne et stockage de l'eau de pluie), le matériau et son produit (peaux et cuir), un processus et son agent neutralisant (inflammation et anti-inflammatoire), un objet et son origine (puits et eaux souterraines), un objet ou un concept et une unité ou un mécanisme de mesure (température et thermomètre), une profession et la personne qui exerce cette profession (maçonnerie et maçon).

16 Les vocabulaires contrôlés ont gagné une importance nouvelle avec le déploiement du Web sémantique. Ce système conçu pour permettre aux machines de répondre aux requêtes humaines a besoin de s'appuyer sur des entités conceptuelles, plutôt que sur des chaînes de caractères. Dans cette perspective, les référentiels, constitués de notices d'autorité, forment la « clef de voûte du Web de données » (Bermès 2013, 44). En tant que « réservoir d'URI [11] réutilisables », ils offrent des « points d'ancrage qui vont transformer les graphes isolés de chaque institution en un graphe global, relié par des points de contacts communs » (Bermès 2013, 44). Le Web de données propose ainsi une forme d'interopérabilité qui repose sur la création d'un espace global d'information, constitué de liens pour naviguer d'une ressource à l'autre.

17 La solution *HyperThesau* propose ainsi un thésaurus pivot appuyé sur les grands référentiels du Web de données, conçu comme un outil de médiation entre des vocabulaires « locaux » ou « maison » et des vocabu-

laires documentaires plus généraux. Sur le modèle de la roue et de l'essieu (*hub and spoke*), l'idée est d'aligner le vocabulaire archéologique hétérogène sur un référentiel externe qui procure un identifiant du concept, indépendant du vocabulaire utilisé dans les jeux de données (figure 2). Ce thésaurus joue ainsi un rôle de pivot pour l'interopérabilité des données archéologiques en ménageant des convergences avec les grands référentiels publiés dans le Web de données par la communauté des bibliothèques.

FIGURE 2. CRÉER UN POINT DE CONTACT ENTRE JEUX DE DONNÉES PAR LES VOCABULAIRES : LE MODÈLE DE LA ROUE ET DE L'ESSIEU (*HUB AND SPOKE*)

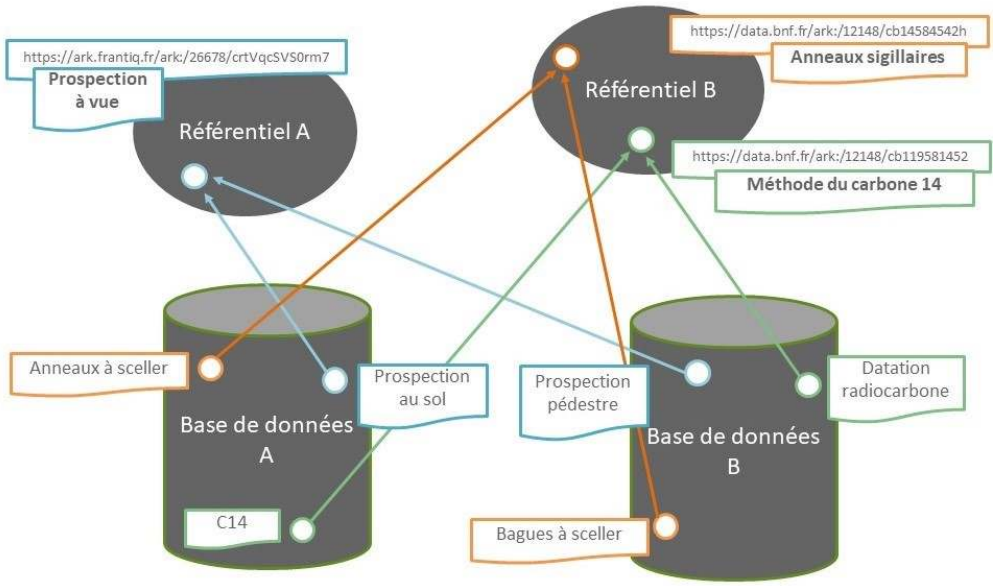


Schéma : E. Perrin d'après Bermès 2013

18 Le thésaurus d'*HyperThesau* est construit avec le gestionnaire de thésaurus multilingue Opentheso, développé par la plateforme technologique Web sémantique et thésauri, de la Maison de l'Orient et de la Méditerranée à Lyon [12]. Opentheso permet notamment d'exporter un thésaurus au format SKOS (*Simple knowledge organization system* [13]), qui est une recommandation du W3C pour représenter des langages documentaires. Ce format s'appuie sur le modèle de données RDF et il permet la publication et l'utilisation des vocabulaires structurés dans le cadre du Web sémantique.

Recueil et définition du vocabulaire de l'archéologie

Articuler recherche et sciences de l'information

19 Le travail de réflexion engagé sur le vocabulaire de l'archéologie dans le cadre du projet *HyperThesau* est relativement bien accueilli pour les questions de rétrodocumentation. Il paraît utile pour bâtir des liens avec des données anciennes, quand, par exemple, la dénomination et la localisation des secteurs de fouille ont évolué avec le temps, comme c'est le cas pour les sites d'Ullastret et de Bibracte. Le besoin de normalisation se fait logiquement davantage ressentir quand une base de données est renseignée par un grand nombre de chercheurs issus d'institutions et de programmes différents que dans le cadre d'un projet dont les membres peuvent s'accorder de manière orale et informelle sur leur vocabulaire de description. On relève cependant une certaine réticence vis-à-vis des outils documentaires, dont la mission centrale est pourtant l'accès à « l'ensemble des informations produites par la recherche et nécessaires à l'activité scientifique[14] ». Le vocabulaire normalisé des bibliothèques, quoiqu'établi à partir de l'indexation des publications scientifiques, paraît paradoxalement n'être jamais adapté, ou être insuffisamment ou trop précis[15]. Il est parfois reproché au thésaurus PACTOLS d'être issu des bibliothèques et non du terrain, alors même qu'il répond de plus en plus aux enjeux de l'interopérabilité et du signalement des publications par son intégration dans les principales structures des humanités numériques (Isidore, Métopes, OpenEdition, DARIAH-EU, Wikidata) et de l'archéologie (consortium MASA, INRAP, ARIADNE).

20 Le travail de normalisation et de standardisation du vocabulaire alimente la crainte d'un appauvrissement des données. L'argument de l'originalité du terrain conduit à produire un vocabulaire « maison » fermé sur lui-même pour chaque projet de recherche. Les partenaires du projet *HyperThesau* revendiquent souvent la singularité de leurs méthodes de travail et de leur lexique. Ainsi la société Archéodunum laisse la terminologie d'un chantier au libre choix du responsable d'opération, chacun ayant son propre bagage scientifique. Les discussions autour du vocabulaire ont lieu de manière informelle entre membres d'une équipe. La spécificité de la base de données Artefacts est de présenter des descriptions typologiques originales et inédites. La mise en place d'un protocole homogène pour les prospections géophysiques du site de Bibracte, proposée dès 1997, s'est heurtée aux habitudes de travail des différents intervenants qui utilisaient leurs propres techniques (Sanchez 2019). Le musée d'archéologie de Catalogne dispose d'un thésaurus que les différents établissements qu'il regroupe – dont le site d'Ullastret partenaire du projet – ont modifié et adapté à leurs besoins, pour faire du « sur-mesure à partir d'une racine commune[16] », en rupture avec la notion de vocabulaire partagé que promeut cet outil. Cette spécificité revendiquée apparaît également au sujet du groupement de service (GDS) Frantiq (Fédération et ressources sur l'Antiquité, créée en 1984). Afin d'offrir un accès centralisé à des ressources documentaires spécialisées, Frantiq

gère un catalogue collectif indexé (CCI), qui se présente comme l'unique catalogue national interétablissements réservé à l'archéologie. Le thésaurus PACTOLS permet l'indexation des notices du CCI. Il s'est ainsi développé isolément des agences bibliographiques françaises (Bibliothèque nationale de France [BNF] et Agence bibliographique de l'enseignement supérieur [ABES]) et le rapprochement n'a été envisagé qu'en 2020[17]. Que deviennent les données au vocabulaire spécifique et donc hétérogène produites par un projet ? On ne peut ni les centraliser, ni les comparer, ni les réutiliser. Sans décloisonnement, les données ne peuvent pas circuler dans le circuit de la recherche.

Fouille de texte et analyse de données textuelles

21 En l'absence de liste préétablie de termes, il existe deux méthodes à combiner pour recueillir le vocabulaire d'un thésaurus, en lien avec les habitudes descriptives et les besoins interprétatifs d'une communauté scientifique. La première est la méthode synthétique : on recherche des termes significatifs dans des sources de référence qui contiennent des listes lexicales, comme les index, les tables des matières, les manuels, les bases de données, les documents administratifs, les guides de bonnes pratiques, les fiches d'inventaire, etc. La seconde méthode de recueil de vocabulaire est la méthode analytique, qui consiste à recueillir les mots significatifs du langage naturel à partir de sources textuelles. Dans le cadre du projet *HyperThesau*, ont été testés des outils de fouille de texte et d'analyse de données textuelles. Deux méthodes d'exploration d'un corpus de 134 publications (monographies et articles) ont été utilisées. La première est basée sur la quantification des termes du corpus avec une analyse de type LDA (*Latent Dirichlet allocation*) pour en détecter les thématiques (*topic modeling*). La seconde se fonde sur une analyse morphosyntaxique (lemmatisation, mots-clés). Après révision de l'OCR, un recueil d'articles (Bulliot 1899) a servi à tester le logiciel de textométrie TXM[18]. Dans le cadre du projet *Bulliot, Bibracte et moi*[19], volet participatif du projet *HyperThesau*, AntConc, un autre logiciel d'analyse textuelle, a également été utilisé pour extraire les toponymes du mont Beuvray de différentes sources archéologiques[20]. La fouille de texte et l'analyse de données textuelles ouvrent de riches perspectives pour la recherche d'information et le repérage des entités nommées : le contrôle des synonymes avec la détection des termes discriminants d'un thème donné, l'analyse des différences de niveaux et de qualité de vocabulaire entre les notes de terrain (idiolecte), les bases de données et les publications, l'étude de la variation du lexique en fonction des auteurs, des lieux et des périodes.

« Déloger l'implicite » : objectiver les définitions

22 L'absence d'un lexique archéologique stable et défini est depuis longtemps observée au sein de cette discipline. L'archéologie « déploie une déconcertante collection de termes et de qualificatifs pour les objets, les types, les lieux et les périodes » (Rabinowitz *et al.* 2016, 49). Des notions identiques s'y trouvent désignées par des termes différents, et des notions différentes, par des termes identiques (Ginouvès et Guimier-Sorbets 1978, 12). René Ginouvès et Anne-Marie Guimier-Sorbets font appel à la notion d'intersubjectivité pour expliquer comment l'objectivité d'une description se trouve liée au consensus autour d'un système de référence partagé. Ainsi une notion admise, tel un chapiteau dorique ou un aryballe, possède une désignation sans ambiguïté. Le flottement de langage ne ferait que traduire les flottements du système de référence (Ginouvès et Guimier-Sorbets 1978, 14-15). Pour Jean-Claude Gardin, l'archéologie « ne serait pas tant loquace *et* vague, que loquace *parce que* vague [21] » (Gardin et Lagrange 1975, 3). Le problème de la description de la morphologie des objets illustre bien la tension qui existe entre le caractère continu de l'évolution des formes et celui, nécessairement discontinu, du vocabulaire : le découpage des vases, par exemple, « met en jeu des concepts (ceux qu'on désigne par les mots de "col", "lèvre", "bord", etc.) dont le contenu est difficile à fixer, car dans la pratique courante il repose seulement sur des intuitions plus ou moins partagées et varie ainsi d'auteur à auteur, parfois même de description à description à l'intérieur d'une même publication » (Ginouvès et Guimier-Sorbets 1978, 33 ; voir aussi 63 et 69). Si la coexistence, à l'intérieur d'une même discipline, de vocabulaires spécifiques paraît inévitable, le contexte de la publication, de l'échange et de l'agrégation des données de la recherche invite à articuler ces différents systèmes de description entre eux afin qu'ils soient complémentaires plutôt que concurrents.

23 En raison des divergences sur l'acception de certains mots qui existent au sein de la communauté des archéologues, la composition de ce thésaurus a nécessité un important travail de retour sur les définitions des termes archéologiques, qui doivent être attestées par des sources de qualité (grands dictionnaires de la langue française, manuels d'archéologie, publications scientifiques) qui sont systématiquement citées. Ces notes d'application ont l'avantage d'objectiver les définitions et de fournir un point d'appui aux discussions. Elles permettent aussi de lever les ambiguïtés sur le sens des concepts et de faciliter leur réutilisation. En ce qui concerne le mobilier archéologique par exemple, bien souvent la description morphologique prime sur la définition fonctionnelle de l'objet. On constate également un emploi assez vague de termes généraux. Par exemple, les ouvrages d'archéologie font un usage très restrictif de la notion de « télédétection » qui est assimilée à la télédétection par satellite alors que, selon les dictionnaires, il s'agit de l'ensemble « des techniques permettant de déterminer certaines caractéristiques physiques et biologiques de points observés à partir de mesures effectuées à distance, sans contact matériel avec ceux-ci », ce qui inclut la reconnaissance aérienne, la thermographie et l'usage des radars [22]. À l'inverse, le terme « citerne » est en archéologie utilisé avec un sens plus large que celui de « réservoir, souterrain ou non, construit pour recueillir et conserver les

eaux pluviales [23] ». Un cas intéressant est celui des sources, qui jaillissent d'un lieu naturel, et des fontaines, qui sont des aménagements hydrauliques. En l'absence de fouille, il est, par exemple, difficile de savoir ce qu'il en est des nombreuses sources et fontaines du mont Beuvray, certaines résurgences pouvant être liées à des aménagements souterrains. On observe également l'usage de termes techniques très spécifiques issus de publications anciennes (bélière [24], furgeoire [25], barbacane [26]). Établir le sens des termes spécialisés, souvent implicite, nécessite la compilation de plusieurs documents, dont très peu présentent des glossaires [27], et demande donc un temps supplémentaire de travail.

24 L'alignement sur les grands référentiels, publiés dans le Web de données par la communauté des bibliothèques, est un autre moyen d'objectiver la définition des concepts. Cette opération permet de récupérer de nouvelles informations, comme des identifiants uniques et pérennes, des traductions ou des alignements avec d'autres référentiels [28]. On relève l'emploi assez général de termes désuets comme « parure », sans équivalent dans Rameau, le langage d'indexation matière de la BNF, qui renvoie seulement à deux concepts, « peignes de coiffure » et « épingles comme bijoux ». « Bassin » en tant que structure hydraulique ne figure pas dans le vocabulaire Rameau, qui va préconiser l'emploi de « réservoirs ». À l'inverse, Rameau utilise comme descripteur l'expression « fouilles de sauvetage » alors que la réglementation de l'archéologie parle de « fouilles préventives ». Il existe ainsi différents degrés d'équivalence pour l'alignement des concepts.

25 Le tableau 1 illustre les problèmes de documentation rencontrés lors du recueil et du traitement du vocabulaire de la prospection archéologique pour le thésaurus *HyperThesau*. Ce lexique paraît isolé en l'absence de définition publiée ou d'alignement avec un référentiel extérieur (valeur 0 dans le tableau 1).

TABLEAU 1. UN VOCABULAIRE ISOLÉ ? LE LEXIQUE DE LA PROSPECTION ARCHÉOLOGIQUE DANS DES RÉFÉRENTIELS COURANTS

descripteurs	définition	traduction (en)	alignement	data.bnf	LCSH	AAT	Pactols	Wikidata
Cartes de susceptibilité magnétique	0	1	0	0	0	0	0	0
Configuration Wenner	0	1	0	0	0	0	0	0
Configuration Wenner-Schlumberger	0	1	0	0	0	0	0	0
Géoradars	0	1	1	1	1	1	0	1
Imagerie 3D	0	1	1	1	1	0	0	0
Imagerie acoustique	0	1	1	1	1	0	0	0
Imagerie infrarouge	0	1	1	1	1	0	0	0
Magnétomètres à pompage optique	0	1	1	0	0	0	0	1
Magnétomètres à résonance de protons	0	1	1	0	1	0	0	1
Prospection au détecteurs de métaux	0	1	1	0	0	0	1	0
Radargrammes	0	0	0	0	0	0	0	0
Radars à visée latérale	0	1	1	0	0	0	0	1
Radars spatiaux	0	1	1	1	1	0	0	0
Sismogrammes	0	1	1	1	1	0	0	1
Sismomètres	0	1	1	1	1	1	0	0
Susceptibilimètres	0	0	0	0	0	0	0	0
Caméras thermiques infrarouge	0	0	0	0	0	0	0	0
Conductimètres	0	1	0	0	0	0	0	0
Configuration dipôle-dipôle	0	1	0	0	0	0	0	0
Imagerie de résistivité électrique	0	1	1	1	0	0	0	0
Images de télédétection	0	1	1	1	1	0	0	1

Ce tableau montre les problèmes que pose la documentation du lexique de la prospection archéologique en l'absence de définition et d'alignement sur des référentiels extérieurs (valeur 0) tels que Rameau (data.bnf.fr), la Bibliothèque du Congrès (*LCSH* : *Library of Congress subject headings*), Art & Architecture Thesaurus du Getty Research Institute (AAT), le thésaurus PACTOLS et la base de données de Wikipédia, Wikidata.

E. Perrin ; source : thésaurus *HyperThesau*

Structurer le vocabulaire archéologique

De l'objet aux concepts

26 Destiné à l'interaction machine-machine, le thésaurus *HyperThesau* obéit à des contraintes formelles et logiques qui sont très différentes de l'interprétation scientifique. L'application stricte de la relation hiérarchique comme relation genre-espèce conduit notamment à « découper » un objet scientifique en plusieurs concepts qui vont entretenir entre eux des relations d'associations. Selon le principe du langage postcoordonné, l'indexation d'un objet fera appel à plusieurs descripteurs : matériau, morphologie, fonction, décor, technique de fabrication, chronologie, répartition géographique. De même, la description de l'acquisition des données associe plusieurs branches ou sous-domaines du thésaurus : « Méthodes et techniques de l'archéologie » (prospection électrique), « Appareils et instruments scientifiques » (résistivimètre) et « Types de données et de documents produits » (cartes de résistivité électrique). La structuration du thésaurus amène aussi à faire une distinction entre les objets, leurs composants et leur matière. Par exemple, le concept de « Ceintures », terme spécifique de la branche « Vêtements », aura une relation associative avec le concept de « Ceintures (composants) [29] », terme spécifique de la branche « Composants ». Les possibilités combinatoires des thésaurus répondent ainsi particulièrement bien aux besoins exprimés par les archéologues d'utiliser un système documentaire souple et adaptable, susceptible de créer des classifications selon différents critères et de rendre compte de l'emboîtement des différents niveaux ou champs d'une analyse. Enfin, la structuration du vocabulaire offre de combiner des descripteurs synthétiques avec des descripteurs analytiques. Des descripteurs synthétiques comme les noms d'outils (hache), les noms de divinité (Athéna) ou de type d'ornement (dents de scie) pourront être associés à la description analytique et explicite de leurs éléments ou attributs constitutifs (Ginouvès et Guimier-Sorbets 1978, 36-37).

Décloisonner les systèmes de classement

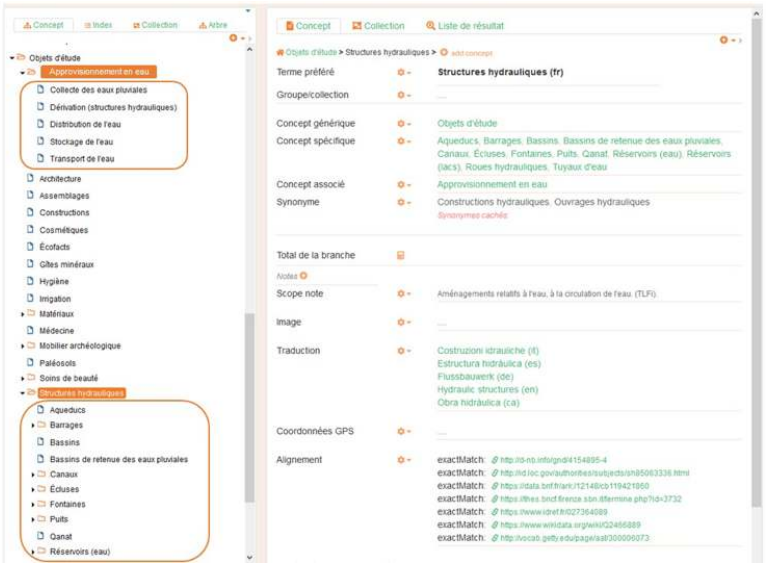
27 À l'instar du vocabulaire, les différents systèmes de classement proposés par les archéologues supposent le même travail de structuration en thésaurus et d'alignement avec les concepts de référentiels. On constate tout d'abord de grandes divergences, par exemple au sujet de la prospection aérienne qui, selon les auteurs, n'inclut pas les mêmes techniques (tableau 2).

TABLEAU 2. LA PROSPECTION AÉRIENNE ET LES DIFFÉRENTES TECHNIQUES QUI LUI SONT ASSOCIÉES

Sources	**Dabas *et al.* 1998**	**Demoule *et al.* 2002**	**Djindjian 2013**	archaeologydataservice.ac.uk
Titre du chapitre	Prospection aérienne à basse altitude	Prospection par observation	Prospection aérienne	Aerial survey and remote sensing
Techniques associées	Photographie aérienne	Prospection aérienne	Photographie aérienne	Aerial photographs
			Télédétection par image satellite	Optical imagery from space, radar imagery from space
			Prospection thermique	Airborne thermography
			Prospection par radar aéroporté (SLAR)	
			Prospection par laser scanner aéroporté (Lidar)	
				Airborne multispectral scanners
				Ground-based imagery

28 L'élaboration d'un microthésaurus sur la géophysique appliquée à l'archéologie a été discutée lors de deux séances de travail, dans un groupe réunissant des archéologues géophysiciens issus de différentes institutions [30], afin de mettre en place des protocoles unifiés pour l'enregistrement des données de terrain. Les propositions concernant la tomographie de résistivité électrique, subdivisée en deux sous-groupes avec les données 2D et les données 3D, sont caractéristiques de ces relations hiérarchiques « bancales » qui ne relient pas des concepts de même nature. Les données 2D et 3D sont un résultat obtenu avec les différentes méthodes de prospection électrique et non pas une méthode à proprement parler. Il a donc fallu remplacer la relation hiérarchique par une relation d'association qui permet de relier une action avec le produit de cette action et donc une méthode de prospection avec les documents qu'elle produit. On remarque de même au sujet des structures hydrauliques décrites dans la base de données PaléoSyr [31] que des « actions » (extraction/captage, distribution, dérivation, stockage) sont hiérarchiquement reliées à des « objets » (fontaines, puits, bassins, aqueducs) alors que le lien entre un objet et sa fonction forme une relation associative (figure 3). On observe aussi que les classements fonctionnels du mobilier archéologique ont tendance à regrouper des fonctions que l'alignement sur les référentiels demande de séparer (tableau 3). Par exemple, dans le classement proposé par Briand *et al.* (2013), la catégorie « agro-pastorale » correspond à trois concepts du langage Rameau : « Agriculture », « Élevage » et « Outils agricoles ». Il en va de même pour la catégorie « Croyances et funéraire » qui renvoie dans Rameau aux « Rites et cérémonies funéraires », aux « Religions » et aux « Cultes » (Briand *et al.* 2013).

FIGURE 3. STRUCTURES HYDRAULIQUES ET APPROVISIONNEMENT EN EAU



Les objets sont reliés à leur fonction par une relation associative.
Source : thésaurus *HyperThesau*

TABLEAU 3. LE CLASSEMENT FONCTIONNEL DES MOBILIERS D'*INSTRUMENTUM* : DES CATÉGORIES À ALIGNER SUR UN SYSTÈME D'ORGANISATION DES CONNAISSANCES DE TYPE RAMEAU

	CATÉGORIE	DÉFINITION	FONCTION
1	Agro-pastoral	Exploitation du sol, entretien paysager et des espaces, élevage, agriculture ; comprend les outils agro-pastoraux	Produire de la nourriture, entretenir
2	Chasse et pêche	Objets dédiés exclusivement à la pêche et à la chasse (et non à la guerre)	Se procurer de la nourriture
3	Artisanat	Extraction, fabrication et transformation des ressources naturelles et des produits agro-pastoraux ; inclut les matières premières, les outils et ustensiles artisanaux	Extraire, transformer, fabriquer
4	Divers production	Objets déterminés mais impossibles à classer en « agro-pastoral » ou « artisanat » (outils, déchets…)	
5	Activité culinaire	Objets de la sphère culinaire, de la préparation à la consommation (ustensiles, vaisselle, stockage)	Conserver, préparer/cuire, manger/boire
6	Éclairage, chauffage	Éléments liés à la production de chaleur et de lumière	Éclairer, chauffer
7	Ameublement	Regroupe les meubles, leurs décors, ainsi que leurs éléments d'assemblage (charnières, ferrures…) et leurs systèmes de fermeture (serrure, cadenas, clés, bagues-clés…)	Aménager, ranger
8	Gros-œuvre	Clouterie et pièces de quincaillerie utilisées pour la construction (élévations, couvertures) et l'aménagement du bâti	S'abriter, habiter
9	Huisserie	Clouterie et pièces de quincaillerie dont l'utilisation pour la réalisation et le fonctionnement des ouvrants est avérée	Se protéger, circuler
10	Hydraulique	Objets liés à la gestion de l'eau	Conduire, stocker l'eau
11	Parure, vêtement	Comprend les éléments de costume et les accessoires	Se vêtir, paraître
12	Soin du corps	Objets de la toilette, de l'hygiène et de la médecine	(S')entretenir, (se) soigner
13	Attelage, véhicule	Pièces d'attelage, de charonnerie et véhicules terrestres divers	Se déplacer, transporter
14	Équipement lié à l'animal	Harnachement (dont les éperons), ferrures animales	
15	Navigation	Vestiges mobiliers liés au transport par voie d'eau	
16	Compte, mesure, échange	Objets liés au commerce et mesures (balances, poids, plombs fiscaux, jetons…)	Commercer, échanger
17	Écriture	Objets liés à l'écriture	Écrire, dessiner, communiquer
18	Divertissement	Pièces de jeu, jouets, pipes…	Se divertir, fêter
19	Musique	Instruments	Se divertir, croire/honorer, signaler/fédérer
20	Équipement militaire	Inclut l'équipement et les armes, y compris celles pouvant être utilisées pour la chasse	S'équiper, combattre
21	Statuaire	Représentations en ronde-bosse de toutes dimensions	Orner, représenter, croire/honorer
22	Croyances et funéraire	Objets liés à la magie, divination, religion, rites, pratiques cultuelles et funéraires	Croire, rendre hommage, prier, consacrer
23	Éléments d'assemblage et de serrurerie	Objets pour lesquels il est impossible de trancher entre « ameublement » et « huisserie » (certaines ferrures, pentures, clés, certains éléments de serrure…)	
24	Divers, polyvalent	Objets identifiés mais dont la fonction précise n'est pas déterminée (certains contenants, couteaux…)	
25	Indéterminés	Objets dont l'identification est inconnue	

Source : Briand *et al.* 2013, 17

29 Pour conclure sur les réalisations du projet, la démarche exposée dans cet article a permis de rassembler, de définir, de traduire, d'aligner et d'associer 476 concepts ou descripteurs et 379 non-descripteurs (ou synonymes) pour le thésaurus d'*HyperThesau*[32]. Les « Méthodes et tech-

niques de l'archéologie » (155 concepts) comprennent principalement les méthodes de prospection (55 concepts), de caractérisation des matériaux (52 concepts) et de datation (14 concepts). Les « Appareils et instruments scientifiques » comptent 33 concepts, et les « Types de données et de documents », 58. La branche « Objets d'étude » regroupe 270 concepts, avec notamment les matériaux (108 concepts) et le mobilier archéologique (109 concepts). Afin que ce thésaurus joue un réel rôle de vocabulaire pivot vers le Web sémantique et le réseau des données ouvertes liées (*linked open data*), les termes ont été systématiquement et manuellement alignés avec plusieurs référentiels : les notices d'autorités de la BNF et les vedettes matière du vocabulaire Rameau [33], les identifiants et référentiels pour l'enseignement supérieur et la recherche de l'ABES (IdRef [34]), les vedettes matière de la Bibliothèque américaine du Congrès (*Library of Congress subject headings* [35]), le Art & Architecture Thesaurus du Getty Research Institute (AAT [36]), le catalogue bibliographique et les autorités de la Bibliothèque nationale d'Espagne [37], le fichier d'autorités intégré de la Bibliothèque nationale d'Allemagne (Deutsche NationalBibliothek [38]), le thésaurus de l'université de Barcelone (THUB [39]), les notices d'autorité de la Bibliothèque nationale centrale de Florence (Nuovo Soggettario [40]), le thésaurus de la Fédération et ressources sur l'Antiquité (PACTOLS [41]), le Iron-Age-Danube Thesaurus publié par l'Austrian Centre for Digital Humanities and Cultural Heritage [42], les vocabulaires publiés par Heritage Data (*linked data vocabularies for cultural heritage* [43]), les thésaurus et vocabulaires contrôlés du German Archaeological Institute (*IDAI for archaeological research* [44]), Periodo pour les périodes historiques [45], le thésaurus de la FAO (Agrovoc [46]) et la base de données de Wikipédia (Wikidata [47]).

30 Pour les domaines de recherche spécialisés, utiliser un vocabulaire commun et contribuer à son enrichissement peut poser de manière récurrente le problème de l'alignement avec les grands référentiels documentaires, dans lesquels les données traitées sont parfois absentes. La diffusion des données de la recherche nécessite la mise en place de nouvelles formes de collaboration entre les institutions productrices de vocabulaires normalisés et les laboratoires. L'application IdRef offre un exemple heureux de convergence entre les données de la recherche et celle des bibliothèques. Développée par l'ABES, elle a pour objectif la mutualisation et la réutilisation des données d'autorité. En exploitant IdRef, on peut récupérer une notice d'autorité, son identifiant unique et pérenne et la forme normalisée d'une appellation. L'exploitation de cet identifiant permet l'interopérabilité et l'agrégation des ressources signalées dans différentes bases de données et donne une plus large visibilité à ces ressources. Parallèlement, IdRef ouvre la production des autorités à ses partenaires de l'enseignement supérieur et la recherche. Il est ainsi possible, dans le cadre d'un projet de recherche, de compléter, de modifier et de créer des notices. La multiplication des expertises contribue ainsi à l'enrichissement et à la consolidation des données.

31 Les tâches liées à la documentation des données peuvent apparaître aux chercheurs comme une contrainte supplémentaire. Le temps à investir pour la description des données n'est pas encore pris en compte dans l'évaluation de leurs travaux. Si les chercheurs se trouvent placés au centre du dispositif, ils ne peuvent assumer seuls toutes les opérations nécessaires à la description et au partage de jeux de données, qui de-

mandent la collaboration de plusieurs acteurs, dont des professionnels de la documentation (Malingre *et al.* 2019). Le travail de réflexion sur le vocabulaire de l'archéologie engagé dans le projet *HyperThesau* montre une fois encore la nécessité de promouvoir la gestion des données comme un indéniable service d'accompagnement à la recherche.

32 La composition de ce thésaurus participe également aux questionnements méthodologiques de l'archéologie. Elle permet de déloger l'implicite pour définir explicitement des objets d'étude et amène à réfléchir à des modèles de données à la fois scientifiquement valides et interopérables. Ce thésaurus engage enfin à une réflexion historiographique sur la construction du lexique de l'archéologie, pour s'interroger sur la place singulière qu'elle occupe parmi les sciences humaines, en tant que discipline vouée à la matérialité des objets et à leur interprétation. Il serait pertinent de considérer son vocabulaire, à l'instar de ses unités d'enregistrement et de ses documents de fouille, comme un véritable objet méthodologique.

Bibliographie

ANR (Agence nationale de la recherche). 2019. « Modèle de Plan de gestion des données (PGD) ». Paris : ANR. https://anr.fr/fileadmin/documents/2019/ANR-modele-PGD.pdf.

ANR (Agence nationale de la recherche). 2021. « La science ouverte ». ANR. https://anr.fr/fr/lanr-et-la-recherche/engagements-et-valeurs/la-science-ouverte/.

Bermès, Emmanuelle. 2013. *Le Web sémantique en bibliothèque*. Paris : Éditions du cercle de la librairie. https://www.cairn.info/le-web-semantique-en-bibliotheque--9782765414179.htm.

Brézillon, Michel. 1983. *La Dénomination des objets de pierre taillée : matériaux pour un vocabulaire des préhistoriens de langue française*. 3e éd. Paris : Éditions du CNRS.

Briand, Aline, Émilie Dubreucq, Aurélie Ducreux, Michel Feugère, Céline Galtier, Benjamin Girard, Didier Josset, Agathe Mulot, Valérie Taillandier et Nicolas Tisserand. 2013. « Le classement fonctionnel des mobiliers d'instrumentum ». *Les Nouvelles de l'archéologie* 131 : 14-19. https://doi.org/10.4000/nda.1764.

Bulliot, Jacques Gabriel. 1899. *Fouilles du mont Beuvray (ancienne Bibracte) de 1867 à 1895*. 2 vol. Autun : Dejussieu.

Cacaly, Serge. 1989. « Les banques de données de la recherche en archéologie : aperçu historique et problématique ». *BRISES. Bulletin de recherches sur l'information en sciences économiques humaines et sociales* 15 : 147-153.

Chaillou, Anne. 2003. « Nature, statut et traitements informatisés des données en archéologie : les enjeux des systèmes d'informations archéologiques ». Thèse de doctorat en sciences humaines et sociales, université Lyon 2. http://theses.univ-lyon2.fr/documents/lyon2/2003/chaillou_a#p=0&a=top.

Chebance, Marie et Esther Magnière. 2020. « Chantier à venir : l'alignement du catalogue indexé avec les catalogues de référence ». *Lettre d'information de Frantiq* 12 (hiver). https://www.frantiq.fr/wp-content/uploads/2020/04/frantiq-newsletter-hiver2020.pdf.

CNRS (Centre national de la recherche scientifique). 2020. « CNRS : un plan ambitieux pour des données accessibles et réutilisables ». CNRS. 16 novembre. https://www.cnrs.fr/fr/cnrsinfo/cnrs-un-plan-ambitieux-pour-des-donnees-accessibles-et-reutilisables/.

Dabas, Michel, Henri Delétang, Alain Ferdière, Cécile Jung et W. Haio Zimmermann. 1998. *La Prospection*. Paris : Éditions Errance.

Darmont, Jérôme, Cécile Favre, Sabine Loudcher et Camille Noûs. 2020. « Data Lakes for Digital Humanities ». Dans *Digital Tools & Uses Congress (DTUC'20). Proceedings of the 2nd International Conference on Digital Tools & Uses Congress*, édité par E. Reyes, G. Kembellec, F. Siala-Kallel, L. Sfaxi, M. Ghenima, I. Saleh, 1-4. New York : Association for Computing Machinery. https://doi.org/10.1145/3423603.3424004.

De Lucca, Erri. 2020. « L'ouïe est l'arbre maître de la connaissance ». *Affaires culturelles*. France Culture. 3 novembre. https://www.franceculture.fr/emissions/affaires-culturelles/affaires-culturelles-emission-du-mardi-03-novembre-2020.

Demoule, Jean-Paul, François Giligny, Anne Lehöerff et Alain Schnapp. 2002. *Guide des méthodes de l'archéologie*. Paris : La Découverte.

Denis, Jérôme, et Samuel Goëta. 2013. « La fabrique des données brutes. Le travail en coulisses de l'open data ». Communication présentée au colloque *Penser l'écosystème des données. Les enjeux scientifiques et politiques des données numériques*, Paris, 14 février. https://halshs.archives-ouvertes.fr/halshs-00990771.

Digard, Françoise. 1975. *Répertoire analytique des cylindres orientaux publiés dans des sources bibliographiques éparses*. Paris : Éditions du CNRS.

Djindjian, François. 2013. *Manuel d'archéologie*. Paris : Armand Colin.

Dufal, Blaise. 2010. « L'archéologie enfermée dehors. Retour sur un malentendu français ». *L'Atelier du Centre de recherches historiques. Revue électronique du CRH* 06 (mai). https://doi.org/10.4000/acrh.2597.

Enlart, Camille. 1916. *Manuel d'archéologie française depuis les temps mérovingiens jusqu'à la Renaissance. Tome III. Le costume*. Paris : A. Picard. https://archive.org/details/manuel-darcheolog03enla.

Gardin, Jean-Claude. 1976. *Code pour l'analyse des formes de poteries*. Paris : Éditions du CNRS.

Gardin, Jean-Claude. 1978. *Code pour l'analyse des ornements*. Paris : Éditions du CNRS.

Gardin, Jean-Claude et Marie-Salomé Lagrange. 1975. *Essais d'analyse du discours archéologique*. Valbonne : Centre de recherches archéologiques.

Gay, Victor, et Marcel Aubert. 1928. *Glossaire archéologique du Moyen Âge et de la Renaissance*. 2 vol. Paris : Éditions Auguste Picard.

Ginouvès, René. 1971. « Archéographie, archéométrie, archéologie. Pour une informatique de l'archéologie gréco-romaine ». *Revue archéologique* 1 : 93-126.

Ginouvès, René. 1989. « Des banques de données pour l'archéologie ? ». *Brises. Bulletin de recherches sur l'information en sciences économiques, humaines et sociales* 15 : 97-107.

Ginouvès, René et Anne Marie Guimier-Sorbets. 1978. *La Constitution des données en archéologie classique. Recherches et expériences en vue de la préparation de bases de données*. Paris : Éditions du CNRS.

Kaeser, Marc-Antoine. 2015. « La muséologie et l'objet de l'archéologie. Le rôle des collections face au paradoxe des rebuts du contexte ». *Les Nouvelles de l'archéologie* 139 (avril) : 37-44. https://doi.org/10.4000/nda.2873.

Lagrange, Marie-Salomé. 1975. *Code pour l'analyse des monuments civils*. Paris : Éditions du CNRS.

Le Rider, Georges. 1975. *Code pour l'analyse des monnaies*. Paris : Éditions du CNRS.

Leroi-Gourhan, André, Gérard Bailloud, Jean Chavaillon et Annette Laming-Emperaire. 1966. *La Préhistoire*. Paris : Presses universitaires de France.

Malingre, Marie-Laure, Morgane Mignon, Cécile Pierre et Alexandre Serres. 2019. « Construction(s) et contradictions des données de recherche en SHS ». *Recherche d'information, document et Web sémantique* 2 (1). https://doi.org/10.21494/ISTE.OP.2019.0336.

MESRI (ministère de la recherche, de l'enseignement supérieur et de l'innovation). 2018. « Le Plan national pour la science ouverte : les résultats de la recherche scientifique ouverts à tous, sans entrave, sans délai, sans paiement ». MESRI. 4 juillet. https://www.enseignementsup-recherche.gouv.fr/cid132529/le-plan-national-pour-la-science-ouverte-les-resultats-de-la-recherche-scientifique-ouverts-a-tous-sans-entrave-sans-delai-sans-paiement.html.

MESRI (ministère de la recherche, de l'enseignement supérieur et de l'innovation). 2021. « Horizon 2020. Le portail français du programme européen pour la recherche et l'innovation. Le libre accès aux publications et aux données de la recherche ». MESRI. https://www.horizon2020.gouv.fr/cid82025/le-libre-acces-aux-publications-aux-donnees-recherche.html.

Mingam, Michel. 2015. « Rameau, les catalogues, le Web ». *Bulletin des bibliothèques de France (BBF)* 5 (avril) : 120-131.

Nivelle, Nicole. 1975. *Code pour l'analyse des monuments religieux*. Paris : Éditions du CNRS.

Perrin, Emmanuelle. 2020. « Bonnes pratiques pour structurer un thésaurus ». *Opentheso* (blog). 28 septembre. https://opentheso.hypotheses.org/67.

Plutniak, Sébastien. 2017a. « L'innovation méthodologique, entre bifurcation personnelle et formation des disciplines : les entrées en archéologie de Georges Laplace et de Jean-Claude Gardin ». *Revue d'histoire des sciences humaines* 31 : 113-139. https://doi.org/10.4000/rhsh.435.

Plutniak, Sébastien. 2017b. « Une contribution archéologique à la théorie des sciences sociales est-elle possible ? Faits et concepts archéologiques entre Jean-Claude Gardin et Jean-Claude Passeron ». *Palethnologie. Archéologie et sciences humaines* 9 (décembre). https://doi.org/10.4000/palethnologie.279.

Rabinowitz, Adam, Ryan Shaw, Sarah Buchanan, Patrick Golden et Eric Kansa. 2016. « Making Sense of the Ways We Make Sense of the Past : The Periodo Project ». *Bulletin of the Institute of Classical Studies* 59 (2) : 42-55. https://doi.org/10.1111/j.2041-5370.2016.12037.x.

Rabot, Alexandre. 2017. « Missions et archives de fouille. Entre la production et la conservation ». Dans *Les Archives de fouilles : modes d'emploi*, édité par Jean-Pierre Brun, Martine Denoyelle, Pierre Rouillard, Stéphane Verger et Sandra Zanella. Paris : Collège de France. http://books.openedition.org/cdf/4909.

Salomé, Marie Rose. 1978. *Code pour l'analyse des textes orientaux*. Paris : Éditions du CNRS.

Salomé, Marie-Rose. 1980. *Code pour l'analyse des représentations figurées sur les vases grecs*. Paris : Éditions du CNRS.

Sanchez, Christelle. 2019. « Reprise de la documentation géophysique sur l'oppidum de Bibracte ». Bibracte EPCC (rapport inédit).

Veyne, Paul. 1978. *Comment on écrit l'histoire*. Paris : Le Seuil.

Zeng, Marcia. 2019. « Interoperability ». *Knowledge Organization* 46 (2) : 122-146. https://www.isko.org/cyclo/interoperability.

Notes

1 Le libre accès aux données de la recherche est soutenu par le Programme européen pour la recherche et l'innovation Horizon 2020 lancé en 2014 (MESRI 2021), la loi pour une République numérique (loi n° 2016-1321 du 7 octobre 2016 pour une République numérique, https://www.legifrance.gouv.fr/dossierlegislatif/JORFDOLE000031589829/), le Plan national pour la science ouverte (MESRI 2018), l'Agence nationale pour la recherche (ANR 2021) et, plus récemment, en novembre 2020, par le plan « Données de la recherche » du CNRS (CNRS 2020).

2 FAIR est l'acronyme de *Findable, Accessible, Interoperable, Reusable*.

3 Selon la définition proposée par la norme ISO 25964 « Thésaurus et interopérabilité avec d'autres vocabulaires » (2011-2013).

4 Voir, par exemple, la présentation des principes FAIR sur le site Doranum : https://doranum.fr/2019/12/04/presentation-des-principes-fair-nouvelle-ressource/. L'emploi de vocabulaires contrôlés est également une recommandation de l'Agence nationale de la recherche pour l'évaluation de la qualité des données dans les plans de gestion des données (ANR 2019).

5 Telle que la définit la réglementation de l'archéologie dans l'arrêté du 16 septembre 2004 portant définition des normes d'identification, d'inventaire, de classement et de conditionnement de la documentation scientifique et du mobilier issu des diagnostics et fouilles archéologiques, https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000000627559/.

6 Comme le rappelle Paul Veyne (1978, 14), un événement est toujours saisi « incomplètement et latéralement » à travers des traces.

7 ISO 25964-1:2011 Information et documentation – Thésaurus et interopérabilité avec d'autres vocabulaires – Partie 1 : Thésaurus pour la recherche documentaire et ISO 25964-2:2013 Information et documentation – Thésaurus et interopérabilité avec d'autres vocabulaires – Partie 2 : Interopérabilité avec d'autres vocabulaires. Voir aussi Perrin 2020.

8 Rameau (Répertoire d'autorité matière encyclopédique et alphabétique unifié) est un langage documentaire utilisé pour l'indexation matière par la Bibliothèque nationale de France, les bibliothèques universitaires et des nombreuses bibliothèques de recherche et de lecture publiques. https://rameau.bnf.fr.

9 Le thésaurus PACTOLS, créé en 1987, sert à l'indexation du catalogue collectif de la Fédération et ressources sur l'Antiquité (Frantiq), qui offre un accès centralisé aux ressources documentaires des bibliothèques d'archéologie partenaires. Il est organisé en six domaines, qui forment son acronyme : Peuples et cultures, Anthroponymes, Chronologie, Toponymes, Œuvres, Lieux et Sujets. Voir aussi ci-dessous et https://www.frantiq.fr/pactols/le-thesaurus/.

10 https://www.frantiq.fr/pactols/le-thesaurus/. C'est l'auteur qui souligne.

11 *Uniform resource identifier* (identifiant uniforme de ressource) : cette chaîne de caractères identifie de façon unique et permanente une ressource sur un réseau (https://digithum.huma-num.fr/ressources/glossaire/).

12 https://opentheso.hypotheses.org.

13 https://www.w3.org/TR/skos-reference/.

14 Selon la définition des missions de l'information scientifique et technique donnée par le ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation : https://www.enseignementsup-recherche.gouv.fr/fr/information-scientifique-et-technique-51161/.

15 Certains archéologues réclament des hyperonymes, dont le caractère général permet d'éviter le recours impropre au vocabulaire contemporain et de désigner la fonction d'une structure au sens large : structure funéraire pour « des vestiges laissés par un processus funéraire quel qu'il soit », structure d'habitat pour les vestiges de construction ayant servi d'habitation aux hommes et à leurs animaux domestiques (remblais, amas d'os de mammouth, abris naturels, trous de poteaux, pierres de calage, pieux, murets) (Djindjian 2013, 172).

16 Gabriel de Prado, responsable du site d'Ullastret, entretien du 29 mai 2019.

17 La moitié des notices des monographies recensées ne comporte pas d'identifiant de type ISBN pour effectuer l'alignement (Chebance et Magnière 2020).

18 http://textometrie.ens-lyon.fr.

19 https://bbm.hypotheses.org.

20 http://explorationdecorpus.corpusecrits.huma-num.fr/antconc/.

21 C'est l'auteur qui souligne.

22 Les définitions citées en notes 22, 23 et 24 sont issues du *Trésor informatisé de la langue française*, consulté sur le site du CNRTL. https://www.cnrtl.fr/definition/télédétection/.

23 https://www.cnrtl.fr/definition/citerne/.

24 Anneau qui maintient le battant d'une cloche, par analogie, anneau de suspension d'une lampe d'église, d'une montre, d'une boucle d'oreille, d'un fourreau de sabre (*CNRTL*).

25 Nom donné à divers outils de toilette (cure-dents, cure-oreille et autres) réunis en manière de trousse, ils se suspendaient parfois à la ceinture (Gay et Aubert 1928).

26 Barbacane ou porte d'agrafe : anneau destiné à recevoir une agrafe (Enlart 1916).

27 R. Ginouvès et A.-M. Guimier-Sorbets, dans leur travail sur la constitution des données archéologiques, observaient déjà que « la définition de chaque terme n'est pas (ou n'est qu'exceptionnellement) indiquée dans les publications ». Ils avaient dû comparer les descriptions d'un même auteur pour connaître les critères déterminants de l'emploi d'un terme ou d'un autre (Ginouvès et Guimier-Sorbets 1978, 35). Selon François Djindjian (2013, 173), chaque publication devrait posséder en annexe un lexique donnant la définition précise de chaque mot.

28 Les notices de data.bnf.fr incluent régulièrement des alignements avec la Bibliothèque du Congrès, la Bibliothèque nationale allemande, la Bibliothèque nationale d'Espagne et la Bibliothèque nationale centrale de Florence.

29 Les éléments composant une ceinture sont, par exemple, les agrafes, les boucles, les chapes, les contre-plaques, les mordants. Le caractère fragmentaire du mobilier archéologique donne toute son importance à la branche des « Composants », qui sont souvent les uniques vestiges d'un objet : perles d'un collier, cabochons, clous d'une chaussure, mors, etc.

30 Archéorient (UMR 5133), Artehis (UMR 6298), Aoroc (UMR 8546), Bibracte EPCC, Chrono-environnement (UMR 6249), Masaryk University.

31 https://www.cepam.cnrs.fr/ressources/bases-de-donnees/paleosyr-paleolib/.

32 Le thésaurus est publié sur la plateforme : https://thesaurus.mom.fr/opentheso/?idt=25. Veuillez sélectionner *HyperThesau* dans le menu déroulant « Choisir un thésaurus ».

33 https://data.bnf.fr.

34 https://www.idref.fr.

35 https://id.loc.gov/authorities/subjects.html.

36 http://www.getty.edu/research/tools/vocabularies/aat/.

37 http://datos.bne.es.

38 https://portal.dnb.de.

39 https://vocabularis.crai.ub.edu/fr/.

40 https://thes.bncf.firenze.sbn.it/ricerca.php.

41 https://pactols.frantiq.fr/opentheso/index.xhtml.

42 https://vocabs.dariah.eu/iad_thesaurus/de/?clang=en/.

43 https://www.heritagedata.org/blog/vocabularies-provided/.

44 https://idai.world.

45 https://client.perio.do.

46 http://www.fao.org/agrovoc/fr/.

47 https://www.wikidata.org.

Auteur

Emmanuelle Perrin

UMR 5133 Archéorient, CNRS, Lyon, France
Emmanuelle Perrin est docteur en histoire et formée aux humanités numériques. En tant que post-doctorante dans le cadre du projet *HyperThesau*, elle était chargée de la mise en œuvre d'un thésaurus pivot pour aligner les concepts de l'archéologie sur les systèmes d'information internationaux.
ORCID 0000-0001-7571-7194
emmanuelle.perrin.touche@gmail.com

Droits d'auteur

Les contenus de la revue *Humanités numériques* sont mis à disposition selon les termes de la Licence Creative Commons Attribution 4.0 International.