ArcheoSciences
Revue d'archéométrie

45-1 | 2021
14th International Conference of Archaeological Prospection

Thesauri and Terminologies as Sources for the Interoperability of Archaeological Data: The Case of Prospection Vocabulary

Marie-Odile Rousset and Jean-Pierre Girard

Electronic version
URL: https://journals.openedition.org/archeosciences/9549
DOI: 10.4000/archeosciences.9549
ISSN: 2104-3728

Publisher
Presses universitaires de Rennes

Printed version
Date of publication: 16 August 2021
Number of pages: 191-195
ISBN: 978-2-7535-8587-4
ISSN: 1960-1360

Electronic reference
Marie-Odile Rousset and Jean-Pierre Girard, "Thesauri and Terminologies as Sources for the Interoperability of Archaeological Data: The Case of Prospection Vocabulary", *ArcheoSciences* [Online], 45-1 | 2021, Online since 16 August 2021, connection on 06 January 2023. URL: http://journals.openedition.org/archeosciences/9549 ; DOI: https://doi.org/10.4000/archeosciences.9549

Thesauri and Terminologies as Sources for the Interoperability of Archaeological Data: The Case of Prospection Vocabulary

Marie-Odile Rousset [a] and Jean-Pierre Girard [b]

Highlights:

- *Vocabulary-based approach of interoperability for archaeological data.*
- *Use of thesaurus (ISO 25964) as semantic tools for building a "hub" between multiple scientific vocabularies and perennial librarian repositories, to allow "machine to machine" dialogue.*
- *A bottom-up (from ground vocabulary to abstraction) and collective approach to build a disciplinary consensus.*

Keywords: archaeology, geophysical survey, terminologies, thesaurus, interoperability.

The challenge: sharing massively heterogeneous data

Archaeology shifts a large part of its field of study towards derived scientific objects, "traces of traces" that were for a long time material (notes, drawings, archives, maps, memoria), and that are now digital or digitized. The HyperThesau project, led by the Archéorient laboratory and financed by the Labex Intelligences des Mondes Urbains of the Université de Lyon, proposes an original approach to the problem of the structural and semantic heterogeneity of these archaeological data: the project intends to found the interoperability of the data on vocabulary and plans to build a thesaurus as a semantic hub for archaeology.

Expression of a data model and terminology

In the early years of the last decade, the STAR project led by English Heritage produced an extension to the CIDOC Conceptual Reference Model (CIDOC CRM – ISO 21127:2006 standard), which models the archaeological excavation and analysis process (CRM-EH). Following this work, several initiatives and programs (ARIADNE in Europe, MASA in France) have focused their efforts on accessing archaeological datasets and modeling the process of data acquisition in the field with CIDOC-CRM (Tudhope *et al.*, 2013; Doerr *et al.*, 2016). All of these approaches, however, rely ultimately on overcoming mismatches between all possible meanings/representations through the use, below the model, of "controlled" vocabularies (the terms being the concepts of a hierarchical thesaurus) to be aligned with when publishing the data.

A thesaurus is an organized list of controlled terms. This documentary tool seeks to solve the problem of equivocality in natural language (polysemy and homonymy). A descriptor or preferred term must univocally describe a concept. This type of vocabulary is used for indexing resources and aims to improve documentary researches by increasing the recall rate of relevant documents in relation to a query.

[a] *Corresponding author, Archéorient, 5/7 rue Raulin 69365 Lyon Cedex 07*
[b] *Archéorient, 5/7 rue Raulin 69365 Lyon Cedex 07*

Strictly defined by the ISO 25964-1 and 25964-2 standards (Hudon, 2012), the structure of a thesaurus allows the expression of all of the semantic relations of a concept (equivalence relation, hierarchical relation, associative relation).

Expressing and translating the singularity of the data

The thesaurus developed as part of the HyperThesau project (https://thesaurus.mom.fr) aims to harmonize the scientific and technical vocabulary of archaeology [1]. It is built with the multilingual thesaurus manager Opentheso (developed by the Semantic Web and thesauri technological platform of the Maison de l'Orient et de la Méditerranée Jean-Pouilloux) which allows an export in Skos, a standard format for the publication of thesauri in the semantic web. It is mainly based on the collection and processing of terminology used in the databases, resulting from archaeologists' practices. Nevertheless, it requires a huge amount of work in order to disambiguate terms by means of their definitions, which are attested by various quality sources (*Dictionnaire de l'Académie française*, *Littré*, *Trésor de la langue française*, archaeology manuals, scientific publications, indexes and glossaries) that are systematically referred to.

Expressing uncertainty and variability

At the same time let us consider the disruptive approach of the PeriodO chronological repository (Rabinowitz, 2014, https://perio.do): it aims to organize, not concepts of "periods" *in abstracto*, but sourced definitions of periods associated with a specific date range and geographic area through identified research work. Taking into account the ambiguity of the ontological status of the "period" (authority or concept), the repository thus constituted allows each person to situate his or her own periodization in relation to all the different definitions that may have been given. The "hub" becomes "a set of stable references of what authorities say about a period, rather than a thesaurus that seeks to impose a consensus on what periods are" (Rabinowitz, 2014). The semantic values of the descriptors that can be harvested are thus often both *close* and *dissimilar*. This creative vision of a "controlled vocabulary" is perfectly suitable to the multiple "points of view" that are characteristic of scientific analysis. Transforming this uncertainty into metadata means being able to expand or interpret the semantic and scientific environment of the data, to query it in an uncertain context.

Mapping to the subject repositories of the semantic web

The path chosen by HyperThesau consists of building a hub thesaurus in order to allow thematic alignments at a fine level, by "bouncing" between perennial repositories, outside the disciplinary community but endowed with the means and authority compatible with their universal vocation, and the multiple vocabularies of archaeologists. This thesaurus prepares the interoperability of archaeological data by providing links with international information systems and those perennial repositories, published in the web of data by the library community (for instance: data.bnf.fr, Library of Congress Subject Heading, VIAF). It is designed as a mediation tool between "local" vocabularies or idiolects and more general documentary vocabularies. With regard to French language, the aim is to make the concepts commonly used in archaeology coexist and communicate with the subject indexing language of the BnF called Rameau (Répertoire d'Autorité-Matière Encyclopédique et Alphabétique Unifié).

Designed for machine-machine interaction, this thesaurus obeys formal and logical constraints that are very different from scientific interpretation. The strict application of the hierarchical relation as a genus-species relation leads to the "cutting up" of a scientific object into several concepts that will maintain association relations between them. According to the principle of post-coordinated language, the indexing of a bracelet will call upon several descriptors: material, morphology, function, decoration, manufacturing technique, period, places. Similarly, the description of data acquisition associates several branches or sub-domains of the thesaurus: method, field of observation, instrumentation and documents produced. Eventually, alignment with Rameau and Library of Congress subject headings allows interconnection with other datasets through vocabularies and thus allows to enrich the data and produce new knowledge (Perrin *et al.*, 2020).

Application to geophysical vocabulary

The documentarization (method and material used, geographical space prospected, treatments, outputs, etc.) allowing the qualification of the data resulting from geophysical prospecting is, also in that field, a prerequisite for the sharing and the effective reuse of the data sets and/or their representations. This is why a HyperThesau working group,

1. This work was developed by Emmanuelle Perrin, a specialist in semantic web specially recruited thanks to the HyperThesau project.

bringing together the Archéorient, Chrono-Environnement and AOrOc French laboratories as well as the Bibracte European Archaeology Center, has been working since 2019 together with a specialist of Digital Humanities, on the creation of a common/consensual geophysical thesaurus (in French at first).

Based on educational (textbooks, etc.) and scientific material (articles, survey reports, etc.), a list of "candidate" concepts has been structured in a thesaurus, with detailed definitions (from a semantical point of view). The list was then discussed collectively, refined and clarified in depth and finally sourced.

This level of semantic standardization must be built within such a system of regulated cooperation that associates speakers of the language of the domain (i.e., specialists-experts; in this case, archaeologists-geophysicists), and a knowledge

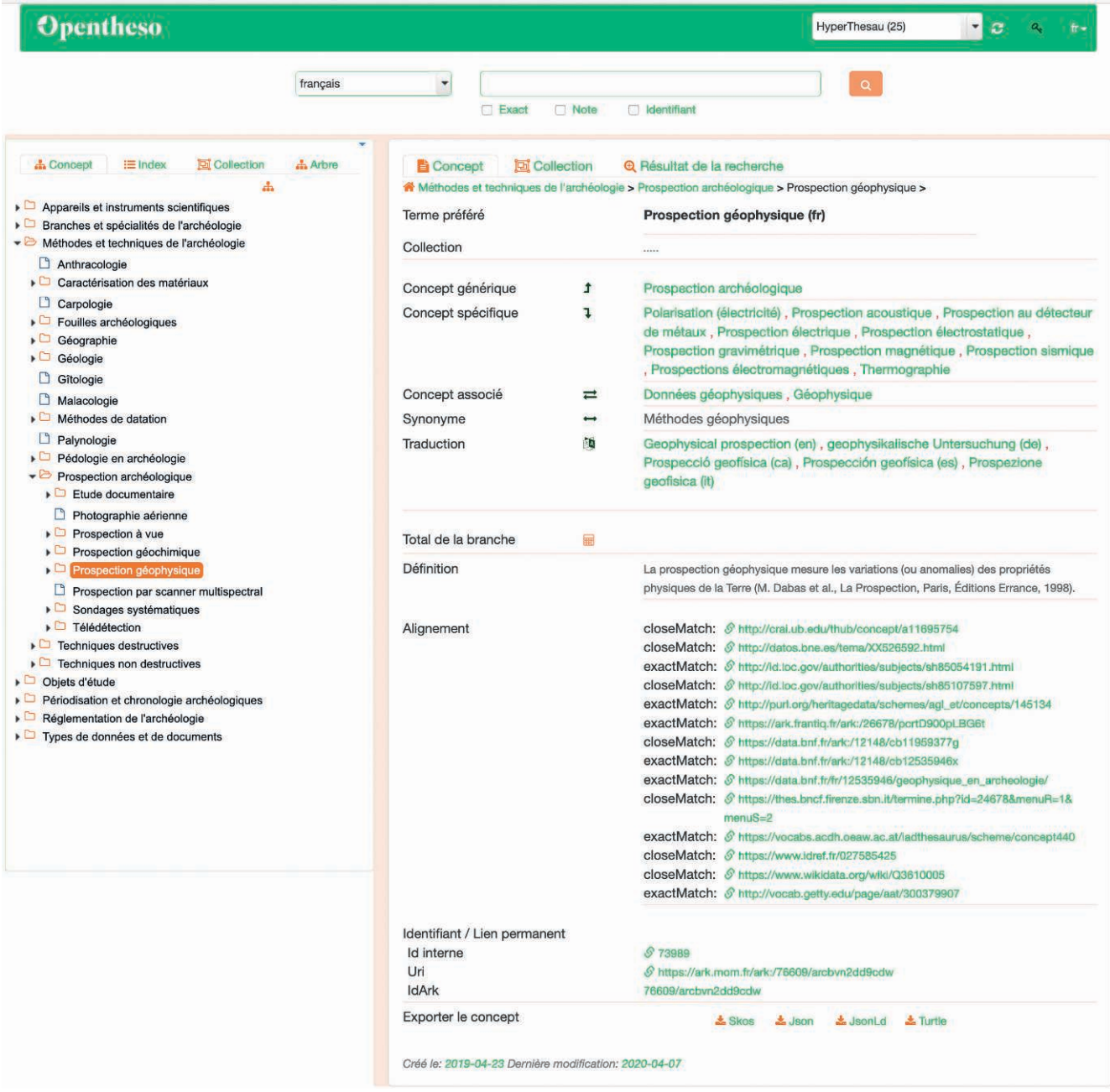

Figure 1. Page of the "Geophysical prospection" concept in HyperThesau thesaurus (Opentheso platform).

Figure 2. Graph of the "Geophysical prospection" concept in HyperThesau thesaurus (Opentheso platform).

engineer who masters the use of the standard, whose role is not that of arbitrating choices but of being a methodological facilitator of the elaboration of a truly collegial semantics (Bachimont, 2000).

At the end of 2020, the list has been published in open source on the Internet: https://ark.mom.fr/ark:/76609/arcbvn2dd9cdw (geophysical survey) and https://ark.mom.fr/ark:/76609/arcv63rxnc56x (geophysical data). Each of its concepts has been given a permanent address (URI) to guarantee its citability and reusability. As it stands, the "geophysical prospecting" branch of the HyperThesau thesaurus is an evolving, open and shared resource which, because it is aligned with the major repositories of the semantic web, is available to the entire scientific community to act as a hub for the interoperability of datasets and deliverables from any prospecting campaign (Fig. 1, Fig. 2).

We look forward to work with all our colleagues in order to augment it and reinforce a disciplinary consensus on a shared vocabulary.

References

Bachimont, B., 2000. Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en Ingénierie des connaissances. In J. Charlet, M. Zacklad,

G. Kassel, D. Bourigault (eds.), *Ingénierie des connaissances, évolutions récentes et nouveaux défis*, Eyrolles, Paris, 305-324.

Doerr, M., Theodoridou, M., Aspöck, E., Masur, A., 2016. Mapping archaeological databases to CIDOC-CRM. In S. Campana, R. Scopigno, G. Carpentiero, M. Cirillo (eds.), *CAA2015 Keep the revolution going – Proceedings of 43*rd *Annual Conference of Computer Applications and Quantitative Methods in Archaeology*, Archaeopress Archaeology, Oxford, 443-451.

Hudon, M., 2012. ISO 25964: pour le développement, la gestion et l'interopérabilité des langages documentaires. *Documentation et bibliothèques*, 58(3): 130-140.

Rabinowitz, A., 2014. It's about time: historical periodization and Linked Ancient World Data. In T. Elliott, S. Heath, J. Muccigrosso (eds.), *Current Practice in Linked Open Data for the Ancient World*, ISAW Papers, 7(22). Published online: http://dlib.nyu.edu/awdl/isaw/isaw-papers/7/rabinowitz/.

Perrin, E., Girard, J.-P., Rousset, M.-O., Durost, S., 2020. Thésaurus et terminologies aux sources de l'interopérabilité des données archéologiques. In C. Marinica, F. Guillet, F. Laroche, J. Velcin (eds.), *Actes de l'atelier DAHLIA. Conférence EGC*, Bruxelles, 19-21. Published online: https://www.egc.asso.fr/wp-content/uploads/egc2020_atelier_Dahlia.pdf.

Tudhope, D., Binding, C., May, K., Charno, M., 2013. Pattern based mapping and extraction via CIDOC-CRM. In V. Alexiev, V. Ivanov, M. Grinberg (eds.), *Proceedings of the Workshop Practical Experiences with CIDOC CRM and its Extensions (CRMEX 2013)*, Valetta, 23-36.