Derivation of thesauri in archaeological conservation through synergistic collaboration between domain experts and the use of semantic and NLP technologies with a FAIR and open publication

Introduction

Digital technology has significantly advanced the conservation of archaeological cultural heritage, underscoring the importance of structured data management. However, diverse innovations have led to various isolated solutions without uniform standards, creating challenges in interdisciplinary cooperation.

The consortium NFDI4Objects (Thiery et al., 2023; Bibby et al., 2023) aims to harmonise this fragmented data landscape concerning human legacies from cultural history, opening new potential for knowledge generation and interdisciplinary linking. Within the consortium, Task Area 4 - "Protecting" (Himmelmann et al., 2024; Fella, Mempel-Länger, and Witt, 2024) focuses on managing and protecting cultural heritage, enhancing the accessibility and exchange of conservation data and facilitating interdisciplinary research using data-driven methodologies. To achieve this, it is necessary to overcome barriers, such as using different specialised terminology.

Different terms can describe the same objects or conservation methods even within individual specialist communities, complicating clear communication and understanding of described entities. This challenge intensifies across disciplinary boundaries. Controlled vocabularies offer a solution by enabling uniform and precise communication, improving clarity and consistency in scientific collaboration. Semantic modelling and linking can connect disparate specialist vocabularies due to the impracticality of establishing a uniform vocabulary across all disciplines.

This project utilises the terminology used at the archaeological conservation of the Leibniz-Zentrum für Archäologie (LEIZA) to illustrate the process and demonstrate the potential for developing semantically networked, machine-readable specialist thesauri.

Material & Data

The project begins by indexing existing technical terms from previous LEIZA conservation reports and records. These "worksheets" (Werkblätter), constitute an archive of analogue and digital documentation that extends far back into LEIZA's (and former RGZMs) history of conservation. They consist of digitised handwritten index cards and digital reports as database entries, supplemented by attached digital documents in different formats. The LEIZA conservators themselves represent a second important source. In discussions with colleagues, who describe certain objects or techniques on a daily basis and use the terminology in practice, further terms are identified. This source also records in-house jargon and terminological imprecision, are recorded as "oral terms" and can be linked within the thesaurus to the correct scientific terminology in a professional context. Furthermore, previous considerations regarding terminologies at LEIZA offer valuable resources, including word lists, first tabular structuring drafts, discussion protocols, and notes.

Methodology

The methodology comprises the following three elements:

1. the capacity of the human mind to logically interpret and evaluate existing, specialised traditions and the vagueness and uncertainty inherent in terminology
2. semantic modelling and data enrichment using the Resource Description Framework (RDF)
3. automatic quantitative data extraction and support with the help of Natural Language Processing (NLP) technologies

A scientist specialising in archaeological conservation collects, validates and sorts subject-specific terms from the existing sources and documents them in a CSV table created according to the SKOS scheme (Miles and Bechhofer, 2009). The limitations of the Simple Knowledge Organisation System (SKOS) are offset by the incorporation of unique identifiers, preferred terms, synonyms, "oral terms", definitions and their sources.

A corpus analysis of text data from the database supports this work. Python scripts divide texts into sentences and words, perform text recognition, and lemmatise tokens to their original form. Common language is removed, and subject-related terms are visualised. Calculating co-occurrences – words that appear together in sentences – enables the identification of a certain semantic connection. Future work will include the expansion of the database, the generation of vector spaces (Mikolov et al., 2013; Pennington, Socher, and Manning, 2014) and the extraction of related topics using latent Dirichlet allocation (Blei, Ng, and Jordan, 2003) or machine learning (Grootendorst, 2022).

Results

The initial result is a vocabulary in tabular form (Fig. 1), developed in conjunction with the input of specialist expertise and technological guidance.

	A	C	D	E	F	G	H	I
1	identifier	level	type	prefLabel	altLabel	translation	description	parent
147	C4BCF8	0	Q2	Zustandserfassung	Zustandserhebung	Condition assessment@en	methodischen Untersuchung zur Ermittlung und Einschätzung des Erhaltungszustandes eines Objektes zu einem definierten Zeitpunkt	top
148	DBC278	1	Q2	Bisherige Lagerung			Bisheriger Verbleib und Aufbewahrung von Artefakten seit ihrer Entdeckung oder Ausgrabung bis zur gegenwärtigen Zeit.	C4BCF8
149	C7C3B7	2	Q2	Frischfund			Das Objekt wurde im direkten Anschluss seiner Bergung von der Grabung in die Restaurierungswerkstatt eingeliefert	DBC278
150	F9GB13	3	Q2	Blockbergung			Bergung des Objektes mit dem umgebenden Erdreich en-block	C7C3B7
151	G4AG5C	3	Q2	Scherbenpflaster			Bergung der einzelnen Scherben/Fragmente des Objektes unter Erhaltung der in-situ Anschlüsse zwischen den Fragmenten	C7C3B7
152	FAG415	3	Q2	Einzelentnahme			Einzelne Entnahme des Objektes/einzelner Fragmente aus dem Boden	C7C3B7
153	FC3322	2	Q2	Ausstellungsentnahme			Das Objekt war ausgestellt und wurde zur Bearbeitung aus der Ausstellung/der Vitirine entnommen.	DBC278
154	F52262	2	Q2	Depotlagerung			Das Objekt wurde bis zur Einlieferung in die Restaurierungswerkstatt im Museumsdepot aufbewahrt	DBC278
155	A8ABBA	2	Q2	klimatisierte Lagerung			Das Objekt wurde bis zur Einlieferung in die Restaurierungswerkstatt unter kontrollierten, klimatischen Bedingungen aufbewahrt.	DBC278
156	C948F2	3	Q2	Gefrierschrank Lagerung			Die bisherige Lagerung des Objektes erfolgte tiefgefroren in einem Gefrierschrank	A8ABBA
157	D21459	3	Q2	Kühlschrank Lagerung			Die bisherige Lagerung des Objektes erfolgte gekühlt in einem Kühlschrank	A8ABBA
158	C2241C	3	Q2	Klimaraum Lagerung	Klimakammer		Das Objekt lagerte bisher in einem Raum, der über eine kontrollierte Klimaregulierung verfügte und auf für das Objekt optimale klimatische Bedingungen eingestellt war	A8ABBA
159	A2BB42	3	Q2	klimatisierte Individualverpackung	Klimakiste|Klimabox		Das Objekt lagerte in einer individuellen Verpackung, die z.B. mittels Trockenmittel oder anderen Reagenzien auf einen speziellen Klimarwert eingestellt war.	A8ABBA
160	G79CA9	2	Q2	unklimatisierte Lagerung			Die Lagerung erfolgte ohne spezielle auf das Objekt abgestimmte, kontrollierte klimatische Bedingungen	DBC278
161	CA1BC5	1	Q2	Objektuntersuchung		investigation@en|examination@en	Beschaffung und Sammlung von Informationen über den Zustand eines Objektes, anhand derer das Restaurierungs-/Konservierungskonzept und/oder die weiteren Lagerungs- und Handhabungsempefehlungen formuliert werden	C4BCF8

Fig. 1 – Extract from the CSV file.

A web application supports further development, allowing the table to be uploaded and validated for coherence. Visualising the table using the d3.js library (Mempel-Länger, 2024c) enhances the clarity of the hierarchical organisation.

To achieve a readable representation for large thesauri, the CSV file is converted to RDF-Turtle using Python scripts. This allows data integration into a SkoHub repository (Mempel-Länger, 2024b), by generating a GitHub page.(Mempel-Länger, 2024a).

Fig. 2 – Word clouds of the most frequent terms in the analysed corpus (left) and the co-occurrences of the terms "shoe" (middle) and "glass" (right).

The tokens derived from the corpus analysis demonstrate its applicability. Figure 2 represents a genuine conservation vocabulary, with "object", "glass", and "fragment" as the most frequent terms displayed in large font sizes and "shoe" and "glass" co-occurrences and fitting semantic relationships. While "sole" and "leather" are terms that occur with greater frequency in the context of "shoe", "shard" and "part" are more typical companions of "glass".

To implement the LOD idea (Berners-Lee, 2006; Schmidt, Thiery, and Trognitz, 2022) and the FAIR principles (Wilkinson et al., 2016), a Wikibase (Vrandecic, 2013; Rossenova, Duchesne, and Blümel, 2022) instance at wikibase.cloud was set up, supporting the extended SKOS schema. This data model includes items (Thiery, Fella, and Mempel-Länger, 2024a) and properties (Thiery, Fella, and Mempel-Länger, 2024b).

Conclusion, Discussion & Outlook

This paper uses a practical example to show how a structured, human- and machine-readable and -understandable thesaurus can be developed from unstructured terms to describe comprehensive conservation processes. The primary objective of this study is to examine the potential for synergistic collaboration between specialised scientists and technical experts and identify the challenges and obstacles encountered during the development process. Moreover, it sets the stage for future innovations in the documentation and sharing of archaeological conservation practices. In the future, dealing with vagueness and uncertainties during the semantic alignment (Thiery and Mees, 2023), the technical possibilities (Unold, Thiery, and Mees, 2019; Tolle and Wigg-Wolf, 2015; Thiery et al., 2021; 2022; 2024) as well as extend the “predicate canon” to describe vague, uncertain or ambiguous vocabulary terms (Nation, 2017; Thiery and Engel, 2016; Piotrowski et al., 2014) is a common topic that has to be addressed within the modelling community, also within the NFDI structure (Thiery, Mees, and Arera-Rütenik, 2021; Thiery et al., 2021).

References

Berners-Lee, T. (2006), 'Linked Data', https://www.w3.org/DesignIssues/LinkedData.html [accessed 21 June 2024]

Bibby, D., K.-C. Bruhn, A. Busch, F. Dührkohp, and et al. (2023), 'NFDI4Objects - Proposal', *NFDI4Objects Zenodo Community*, https://doi.org/10.5281/zenodo.10409227

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003), 'Latent dirichlet allocation', *J. Mach. Learn. Res.*, 3, 993–1022

Fella, K., L. Mempel-Länger, and N. Witt (2024), 'NFDI4Objects - Community Cluster "Konservierung und Restaurierung/ Conservation Science"', https://doi.org/10.5281/zenodo.11370865

Grootendorst, M. (2022), 'BERTopic: Neural topic modeling with a class-based TF-IDF procedure' (arXiv), https://doi.org/10.48550/ARXIV.2203.05794

Himmelmann, U., R. Schwab, S. Metz, T. Krenscher, F. Thiery, J. Lefeldt, et al. (2024), 'Forschungsdatenmanagement im Bereich Denkmalpflege, Archäologie und Restaurierung', https://doi.org/10.5281/zenodo.10978098

Mempel-Länger, L. (2024a), 'LasseMempel/RestSkos-pages', *GitHub*, LasseMempel, https://github.com/LasseMempel/RestSkos-pages

——— (2024b), 'LasseMempel/RestSkos-pages/skosifyCSV', *GitHub*, LasseMempel, https://github.com/LasseMempel/RestSkos-pages/tree/main/skosifyCSV

——— (2024c), 'LasseMempel/ta4-conservation-dev', *GitHub*, LasseMempel, https://github.com/LasseMempel/ta4-conservation-dev

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013), 'Efficient Estimation of Word Representations in Vector Space' (arXiv), https://doi.org/10.48550/ARXIV.1301.3781

Miles, A., and S. Bechhofer (2009), 'SKOS Simple Knowledge Organization System Reference', https://www.w3.org/TR/2009/REC-skos-reference-20090818/ [accessed 21 June 2024]

Nation, Z. (2017), 'Perceptions of Probability and Numbers', https://github.com/zonination/perceptions [accessed 21 June 2024]

Pennington, J., R. Socher, and C. Manning (2014), 'Glove: Global Vectors for Word Representation', *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (presented at the Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar: Association for Computational Linguistics), pp. 1532–43, https://doi.org/10.3115/v1/D14-1162

Piotrowski, M., G. Colavizza, F. Thiery, and K.-C. Bruhn (2014), 'The Labeling System: A New Approach to Overcome the Vocabulary Bottleneck', *DH-CASE II: Collaborative Annotations on Shared Environments: Metadata, Tools and Techniques in the Digital Humanities - DH-CASE '14* (presented at the DH-CASE II: Collaborative Annotations, Fort Collins, CA, USA: ACM Press), pp. 1–6, https://doi.org/10.1145/2657480.2657482

Rossenova, L., P. Duchesne, and I. Blümel (2022), 'Wikidata and Wikibase as complementary research data management services for cultural heritage data', *Proceedings of the 3rd Wikidata Workshop 2022 Co-Located with the 21st International Semantic Web Conference (ISWC2022)*, https://ceur-ws.org/Vol-3262/paper15.pdf

Schmidt, S. C., F. Thiery, and M. Trognitz (2022), 'Practices of Linked Open Data in Archaeology and Their Realisation in Wikidata', *Digital*, 2:3, 333–64, https://doi.org/10.3390/digital2030019

Thiery, F., and T. Engel (2016), 'The Labeling System: The Labelling System: A Bottom-up Approach for Enriched Vocabularies in the Humanities', in S. Campana, R. Scopigno, G. Carpentiero, and M. Cirillo (eds), *CAA2015. Keep the Revolution Going. Proceedings of the 43rd Annual Conference on Computer Applications and Quantitative Methods in Archaeology.* (Oxford: Archaeopress), pp. 259–68

Thiery, F., K. Fella, and L. Mempel-Länger (2024a), 'N4o-ta4-dev.wikibase.cloud | items.csv', *GitHub*, leiza-scit, https://github.com/leiza-scit/n4o-ta4-dev.wikibase.cloud/blob/main/items.csv

——— (2024b), 'N4o-ta4-dev.wikibase.cloud | properties.csv', *GitHub*, leiza-scit, https://github.com/leiza-scit/n4o-ta4-dev.wikibase.cloud/blob/main/properties.csv

Thiery, F., and A. Mees (2023), 'Taming Ambiguity - Dealing with doubts in archaeological datasets using LOD', *CAA 2018: Human History and Digital Future*, https://doi.org/10.15496/PUBLIKATION-87762

Thiery, F., A. Mees, and T. Arera-Rütenik (2021), 'TRAIL 4.2: Implementing mapping processes for vocabularies related to site and object protection', https://doi.org/10.5281/zenodo.5849841

Thiery, F., A. Mees, K. Tolle, and D. Wigg-Wolf (2021), 'TRAIL 2.2: Evaluation of fuzziness and wobbliness in numismatics and ceramology', *NFDI4Objects TRAILS*, 2021, No. 2.2, https://doi.org/10.5281/zenodo.5654897

Thiery, F., A. W. Mees, K. Tolle, and D. G. Wigg-Wolf (2022), 'How to handle vagueness and uncertainty in graph-based LOD knowledge modelling? Dealing with archaeological numismatic and ceramological real world data.', *Squirrel Papers*, 4:1, No. 2, https://doi.org/10.5281/ZENODO.7184523

Thiery, F., A. W. Mees, B. Weisser, F. F. Schäfer, S. Baars, S. Nolte, et al. (2023), 'Object-Related Research Data Workflows Within NFDI4Objects and Beyond', *Proceedings of the Conference on Research Data Infrastructure*, 1, https://doi.org/10.52825/cordi.v1i.326

Thiery, F., F. Schenk, S. Baars, K. Tolle, and P. Thiery (2024), 'Modellierung von Fuzzyness / Wobbliness in Geodaten', *Tagungsband FOSSGIS-Konferenz 2024*, 2024, 64–73, https://doi.org/10.5281/zenodo.10571859

Tolle, K., and D. Wigg-Wolf (2015), 'Uncertainty Handling for Ancient Coinage', in F. Giligny, F. Djindjian, L. Costa, P. Moscati, and S. Robert (eds), *CAA2014. 21st Century Archaeology. Concepts, Methods and Tools. Proceedings of the 42nd Annual Conference on Computer Applications and Quantitative Methods in Archaeology.* (Oxford: Archaeopress), pp. 171–78

Unold, M., F. Thiery, and A. Mees (2019), 'Academic Meta Tool. Ein Web-Tool zur Modellierung von Vagheit', *ZfdG - Zeitschrift Für Digitale Geisteswissenschaften*, Die Modellierung des Zweifels – Schlüsselideen und-konzepte zur graphbasierten Modellierung von Unsicherheiten.:Sonderband 4, https://doi.org/10.17175/SB004_004

Vrandecic, D. (2013), 'The Rise of Wikidata', *IEEE Intelligent Systems*, 28:4, 90–95, https://doi.org/10.1109/MIS.2013.119