Using ChatGPT for generating SKOS thesauri from handwritten sketches

Kraus, Felix
felix.kraus@kit.edu
Karlsruhe Institute of Technology (KIT), Germany

Blumenröhr, Nicolas
nicolas.blumenroehr@kit.edu
Karlsruhe Institute of Technology (KIT), Germany

Introduction and Motivation

Thesauri and controlled vocabularies play an important role in organizing and structuring knowledge (Hyvönen 2020), especially when handling humanities data collections with their diverse types of research objects. For example, thesauri-structured metadata enables researchers to use computer software to query linked data and use it, e.g. for data annotation. The development of such thesauri is not a straight-forward process. Using editors for SKOS (Simple Knowledge Organization System) thesauri (SKOS is the most common data model used to represent machine-readable thesauri (Conway et al. 2016)) such as Protégé (Musen 2015), VocBench (Stellato et al. 2017), or our editor EVOKS (Ernst et al. 2023) rather hinders than amplifies the creativity and flexibility needed especially for the initial drafts that often times is an important part of the research.

In this paper, we explore utilizing ChatGPT (https://chatgpt.com/), which uses a Large Language Model (LLM) designed by OpenAI as base, for transforming automated generation of SKOS thesauri based on hand-drawn sketches or digital drafts created in tools like drawio (https://www.drawio.com/). This method can offer a pragmatic solution for reducing the initial workload, especially when there is little prior experience in creating SKOS thesauri. If additional research or data management software is used that requires SKOS thesauri, e.g. for (meta)data management, it can be quickly evaluated if the developed thesaurus structure fulfils requirements posed by such tools. Naturally, it is possible anytime to import the resulting SKOS output of ChatGPT into an editor for further refinements. Another benefit over conventional editors is the possibility to support the user in the learning process of SKOS when making use of the interactive nature of ChatGPT, e.g. because selected features of the code can be explained to the user.

Methodology

To evaluate the results coming out of ChatGPT, we created sketches, both handdrawn and using drawio. The sketches are based on the DHA taxonomy (https://vocabs.dariah.eu/dha_taxonomy/en/ , CC BY 4.0, Creators: ACDHOEAW Team) and on TaDiRAH: Taxonomy of Digital Research Activities in the Humanities (https://vocabs.dariah.eu/tadirah/en/ , CC0, Creators: Luise Borek, Canan Hastik, Vera Khramova, Jonathan Geiger). To closely resemble real-world applications of this approach, we removed all properties from the terms except the URI, the English label (skos:prefLabel), hierarchy relations (skos:narrower and skos:broader) and the membership of the concept scheme (skos:inScheme) as well as the declaration of the concept scheme itself. We then removed selected hierarchy branches to decrease the number of terms using our python code (published with MIT licence on GitHub (https://github.com/FelixFrizzy/rdf-tools/tree/main/ hierarchy-subranches , DOI: 10.5281/zenodo.12731609)). This led to thesauri small enough to draft by hand in a mind map-like structure (see fig. 1) which were then digitized. Both, the images from the hand-drawn sketch and the draft created with drawio were then fed into ChatGPT together with the prompt as shown in fig. 2, resulting in the output of SKOS files.

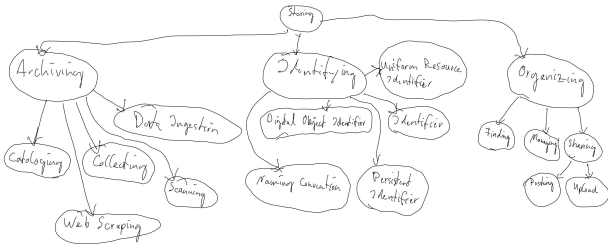

Fig.1. Initial thesaurus structure transferred into hand-drawn figure.

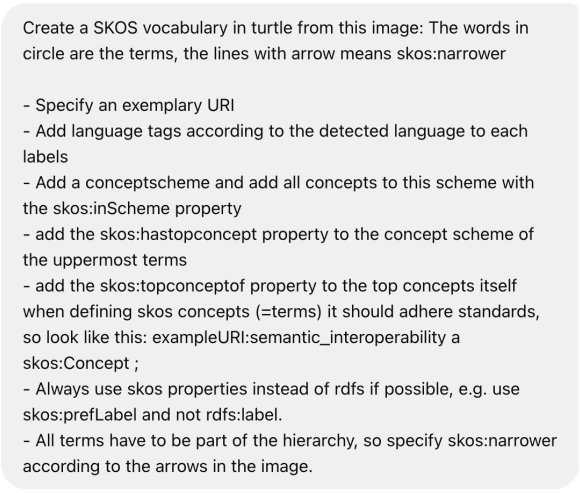

Fig.2. The identical prompt used to generate a SKOS thesaurus using mind map-like hand-drawn drafts.

Additionally, we created 50 random fictional language words (like eumkauh or otrkor) using a tool (Shack 2011) to create a hand-drawn, random hierarchy out of it. This prevents that the GPT model already knows the content of the other two publicly available thesauri because they might have been used for training the model.

To replicate our experiments, we published our data as a supplement to this paper (https://doi.org/10.5281/zenodo.14290535 (CC-BY-SA)). We used the OpenAI 4o-model, only a single prompt input without refining the output, we used a new chat for every prompt and we disabled the memory function. Finally, we compared the SKOS output of ChatGPT with the thesauri which the hand-drawn sketches were based on.

Results and Limitations

When looking at the results for the two hand-drawn thesauri of the size-decreased DHA and TaDiRAH thesauri, the resulting SKOS file closely matches the one we used as source. The main differences were an introduced capitalization, change

of the label of the uppermost term, and three terms one level too deep in the hierarchy, but still in the right branch. All other 35 labels as well as the broader and narrower relationships were identical.

For the drawio case, the original uppermost term was one level too deep in the hierarchy and a new one was created. Furthermore, one label was altered from ”Upload” to ”Uploading”, which would then be more consistent with the term ”Posting” in the same level. All other labels and relations were correct.

Examining the ChatGPT output when using the hand-drawn image with fictional language terms as input, we find that all relations are correct. Different from the other cases, most of the labels are misspelled. This indicates that ChatGPT is using a combination of handwritten text recognition and LLM-based error correction to get the best results. The latter naturally fails when using fictional words. It has to be mentioned that identifying the correct letters in the image is also challenging for humans, which can be assessed in the published paper supplement.

While these results look promising, it is important to keep in mind that LLMs are not deterministic and can change the data and hallucinate. Especially in these cases, LLMs might violate the copyright of data. In our case, this is a minor issue because existing content is transformed, not created.

We found that the prompt generation will stop for a higher number of terms (in our case, more than about 100) which poses a big limitation. This can be easily prevented by using the OpenAI API or by using any other LLM that accepts image input.

Outlook and Conclusion

We plan to conduct further work on using ChatGPT for validation of SKOS files, automated translation or adding descriptions and relations to e.g. Wikidata items which would enhance the usability of our proposed approach.

To summarize, our experiments strongly suggest that ChatGPT or similar can accelerate the process of creating a SKOS vocabulary from handwritten or digital drafts. Setting aside the required prior knowledge for setting up the base structure of a valid SKOS file, using an editor requires manually entering all terms. Even in cases where this can be automized by handwritten text recognition or similar, adding a term relation would still require at least one click or pasting and adjusting one line of code. This process takes far longer than the few seconds that it took ChatGPT to create the result. On the downside, this also poses the danger of misunderstanding important properties of the SKOS data model and therefore creating data model violations. However, by far the biggest benefit is that the barrier of creating Findable, Accessible, Interoperable and Reusable (FAIR, https://www.go-fair.org/fair-principles/) data is tremendously lowered and the learning curve of using dedicated editors can be flattened by supporting the researchers with interactive help.

Bibliography

Conway, M. / **Khojoyan, A.** / **Fana, F.** / **Scuba, W.** / **Castine, M.** / **Mowery, D.** / **Chapman,W.** / **Jupp, S.**: Developing a web-based SKOS editor. Journal of Biomedical Semantics 7(1), 5 (Apr 2016). https://doi.org/10.1186/s13326-015-0043-z

Ernst, F. / **Frank, L.** / **Götzelmann, G.**: EVOKS - Benutzerfreundliche Erstellung kontrollierter Vokabulare für die Geisteswissenschaften. In: FORGE 2023 - Forschungsdaten in Den Geisteswissenschaften: Anything Goes?! Forschungsdaten in Den Geisteswissenschaften - Kritisch Betrachtet. Konferenzabstracts. Tübingen, Germany (Oct 2023). https://doi.org/10.5281/zenodo.8386468

Hyvönen, E.: Using the Semantic Web in digital humanities: Shift from data publishing to data-analysis and serendipitous knowledge discovery. Semantic Web 11(1), 187–193 (Jan 2020). https://doi.org/10.3233/SW-190386

Musen, M.A.: The Protégé Project: A Look Back and a Look Forward. AI matters 1(4), 4–12 (Jun 2015). https://doi.org/10.1145/2757001.2757003

Shack, E.: Random Word Generation for Fictional Languages. Wolfram (2011)

Stellato, A. / **Turbati, A.** / **Fiorelli, M.** / **Lorenzetti, T.** / **Costetchi, E.** / **Laaboudi, C.** / **Gemert, W.V.** / **Keizer, J.**: Towards VocBench 3: Pushing Collaborative Development of Thesauri and Ontologies Further Beyond. In: 17th European Networked Knowledge Organization Systems (NKOS) Workshop. vol. 1937, pp. 39–52. CEUR Workshop Proceedings, Thessaloniki, Greece (Sep 2017)