# Building the DEFC app thesaurus

Seta Štuhec, Ksenia Zaysteva, Edeltraud Aspöck

## Introduction

Thesauri, along with gazetteers, lexical database, ontologies, taxonomies and classification systems, are Knowledge Organization Systems (KOS) that are used to model the underlying semantic structure of a domain for the purposes of data retrieval (Tudhope & Lykke Nielsen, 2006). Depending on their purpose, KOSs differ from each other based on the degree of vocabulary control, richness of semantic relationships and formality (Hodge, 2000).

Most DEFC app entries are a part of controlled vocabulary in a form of drop-down lists filled with terms provided by the AAPP research group (site types, find types etc.). Because different people might use different terms to describe the same concept, it is important to unify the terminology to reduce the description and retrieval ambiguity caused by free text entries. However, DEFC word lists many times include terms that are not on the same hierarchical level, i.e. the same word list contains hypernyms and hyponyms. Furthermore, the meaning of some terms is sometimes similar but not the same and the same concept appears in different parts of the database. For this reason we are using thesaurus as a reference work that lists and contextualizes terms from DEFC word lists based on their meaning, to present and explain the data structure.

## Contextualizing the DEFC app vocabulary

Defining the terms

Hierarchical arrangement

Adding AT, RT, close match, etc.

### Defining the terms

The first step of building the thesaurus for the DEFC app was to create a hierarchical arrangement of terms and link them to each other based on their meaning. To do this, each term that appears in controlled vocabulary of the DEFC app database was first defined to prevent any misconceptions and to serve as a tool for easier hierarchical and other arrangement of the terms. The definitions were formed in a close collaboration with the AAPP research group as well as by using archaeological literature and online resources (online database, thesauri, glossary and encyclopaedia):

- Historic England Thesaurus
- Encyclopaedia Britannica
- Oxford dictionary

- DBpedia
- Encyclopaedia of Life
- Kipfer, B. A. 2000, Encyclopedic Dictionary of Archaeology. Kluwer Academic/Plenum Publishers, New York.
- Kipfer, B. A. 2007, Dictionary of Artifacts. Blackwell Publishing, Oxford.
- Mollett, J. W. 1966, An Illustrated Dictionary of Art and Archaeology. American Archives of World Art, New York.
- Shaw, I., Jameson, R. (eds.) 1999, A Dicitionary of Archaeology. Blackwell Publishers, Oxford.
- Inizan, M.-L., M. Reduron-Ballinger, H. Roche, J. Tixier 1999, Technology and Terminology of Knapped Stone. Nanterre, CREP.,
- D. Srejovic (ed.) 1997, Arheološki Leksikon.
- Peregrine, P. N., Ember, M. (eds.) 2002. Encyclopedia of Prehistory. Vol. 8: Sourth and Southwest Asia. Kluwer Academic/Plenum Pulishers, New York, Boston, Dortrecht, London, Moscow.
- Gopher, A. 1994. Arrowheads of the Neolithic Levant. Wiona Lake, Eisenbrauns.
- Biers, W. R. 1969. Excavations at Phlius, 1924 the prehistoric deposits. Hesperia: The journal of the American School of Classical Studies at Athens. Vol. 38, No.4. pp. 443-458.
- ...


## Classes and top concepts – building hierarchy

Based on their meaning the terms have been arranged in a hierarchical order using broader terms (BT) (and narrower terms) following the DEFC app database structure. Each top class of the data model (site, research event, area, finds, interpretation) presents a top class in the DEFC app thesaurus.

Top class SITE is dealing with the information about the topographic location where a certain archaeological activity took place (coastline, hill...).

Top class AREA holds concepts referring to archaeological areas: cave or rock shelter, settlement, quarry, cemetery or grave. The top concept "settlement" is further described using terminology concerning separate buildings (e. g. building techniques, types, shapes), enclosure types, settlement layout, settlement type and other archaeological features (e. g. installations, storage, ritual structures). Top concept "cemetery or grave" holds information about grave types, human remains, grave disturbances and other mortuary evidence. The top concept "cave or rock shelter" mostly includes the terms related to the evidence of occupations, whereas the top concept "quarry" gathers the terms that describe different types of extraction itself as well as the extracted raw material.

Top class RESEARCH EVENT holds concepts related to the type of research conducted on an archaeological site (such as archaeological excavation, field survey, remote sensing) or finds and soil (such as dating, sourcing analysis, isotope analysis, soil analysis).

Top class FINDS holds concepts about different types of archaeological artefacts including small finds (tools, jewellery, figurines), lithics, pottery, plant remains and animal remains. Each of them holds narrower terms with different level of detail (e. g. lithic types) and descriptive information (e.g. pottery decoration).

Top class INTERPRETATION holds top concepts about past human activities that have been recognized on a certain site including production (e. g. textile, metal, pottery production), subsistence (e.g. fishing, farming, hunting).

Throughout the DEFC controlled vocabulary terms such as "unidentified", "unknown", "undetermined" appear. These terms describe entities that were recognized, but either could not be further identified during the archaeological recording, were not further described in the available documentation or were not further investigated for various reasons. However, it has been recognized as essential to acknowledge their presence and document it in the database.

## Adding synonyms, related terms etc.

After the basic hierarchical structure of the concepts has been established the individual concepts were linked to each other and/or complemented by additional terms.

Since the terms are originally written in English, they have been translated also to German. Additionally, terms for plant and animal remains were specified in Latin. Greek and Turkish language and alphabet were used to denote geographic names of corresponding regions and districts.

Original terms (in English) were then complemented with alternative terms (AT) that included different spelling (such as "tel" and "tell") and synonyms (such as "perforator", "borer", and "drill"). When possible, also new alternative names were translated into German and additional German synonyms and different spellings were added to the thesaurus. Several terms from the database whose meaning is somehow related such as "textile equipment" and "textile production" were linked through related terms label (RT). This means that the AT terms are external, whereas RT are already included in the database.

## Practical implementation

There are several available tools used to build a thesaurus such as TemaTres and CLAVAS & CLARIN concept registry. However, to keep a clear overall view over the thesaurus structure in the making we found easiest to first draft the structure using one of many mind mapping tools. After the general structure has been created the terms were organized using tables and spreadsheets in the Microsoft Excel environment. One term per row was presented as a concept with broader and narrower terms. A general table was constituted from the following fields:

- ID (ID of the term derived from the database)
- Hierarchy fields (1st order, 2nd order…: following from broader terms on the left to narrower terms on the right)
- Translation (a separate column for each additional language)
- Alternative terms (synonyms and different spelling)
- Related terms (terms from the database that have relatable meaning)
- Broader terms (to avoid duplication, another "broader term" column was added to the table)
- Close match (terms from the database that have almost the same meaning)
- Exact match (terms from the database that have exactly the same meaning)
- Definition (definition of the term)
- Source (source of the term definition)

### Table example: Hierarchy

| dc_class | dc_ID | 1st order | 2nd order | 3rd order | 4th order |
|---|---|---|---|---|---|
| | | finds material | | | |
| dc_finds_material | 1 | finds material | stone | | |
| dc_finds_material | 2 | finds material | stone | obsidian | |
| dc_finds_material | 12 | finds material | stone | chert | |
| dc_finds_material | 11 | finds material | stone | chert | flint |
| dc_finds_material | 18 | finds material | stone | chert | radiolarite |

dc_class and dc_ID together work as a unique database constructed identifier of the term. When the field is left blank, the term was added for the purpose of the thesaurus, but cannot be find in the database vocabulary. At the later stage, IDs were also added to new terms.

First order represents the top concept – each further order to the right (2nd, 3rd…) is a narrower term of the term on the left.

### Table example: translations, alternative and related terms

| 1st order | 1st order @de | AT 1st order | RT 1st order |
|---|---|---|---|
| Radiocarbon dating | C14-Datierung | C14 | Dating material |

The translation as well as the alternative and related term refers to the "first order" original term. If the table will be mapped to SKOS using OPEN Refine (as it is described further on), it is important to note, that AT, RT and all other complementary terms should be added in separate columns for each individual order in hierarchy (and for each language separately). Because such table can be rather extensive and not very transparent to work with (if for example we want to assign RT and AT to every order in several languages), we can construct a table as follows: we enter RT and AT column for each language only once. The term in this field would refer to the narrowest term in the hierarchy. When the table is completed we use Excel formulas to re-distribute the RT and AT columns to the hierarchical order it belongs to (the last filled out field).

### Table example: broader terms

| 2rd order | BT |
|---|---|
| grave | settlement; grave & cemetery |
| fireplace | cave; rock shelter |

The label "broader term" has been introduced in the thesaurus to avoid additional work in duplicating terms with more than one top concept. For example, terms describing graves (such as human remains,

mortuary feature, grave location etc.) were used to describe "settlement" (AREA type concept) as well as "cemetery or grave" (AREA type concept) in the database. A "broader term" label therefore assigns the terms to any additional broader term. However, one has to make sure that the entered broader term is actually higher in the hierarchy. As the "cave" and "rock shelter" are 1st order terms the term "fireplace" must be placed a level below – to the 2nd order.

# SKOS mapping

SKOS

Data preparation

Mapping with OPEN Refine

## SKOS

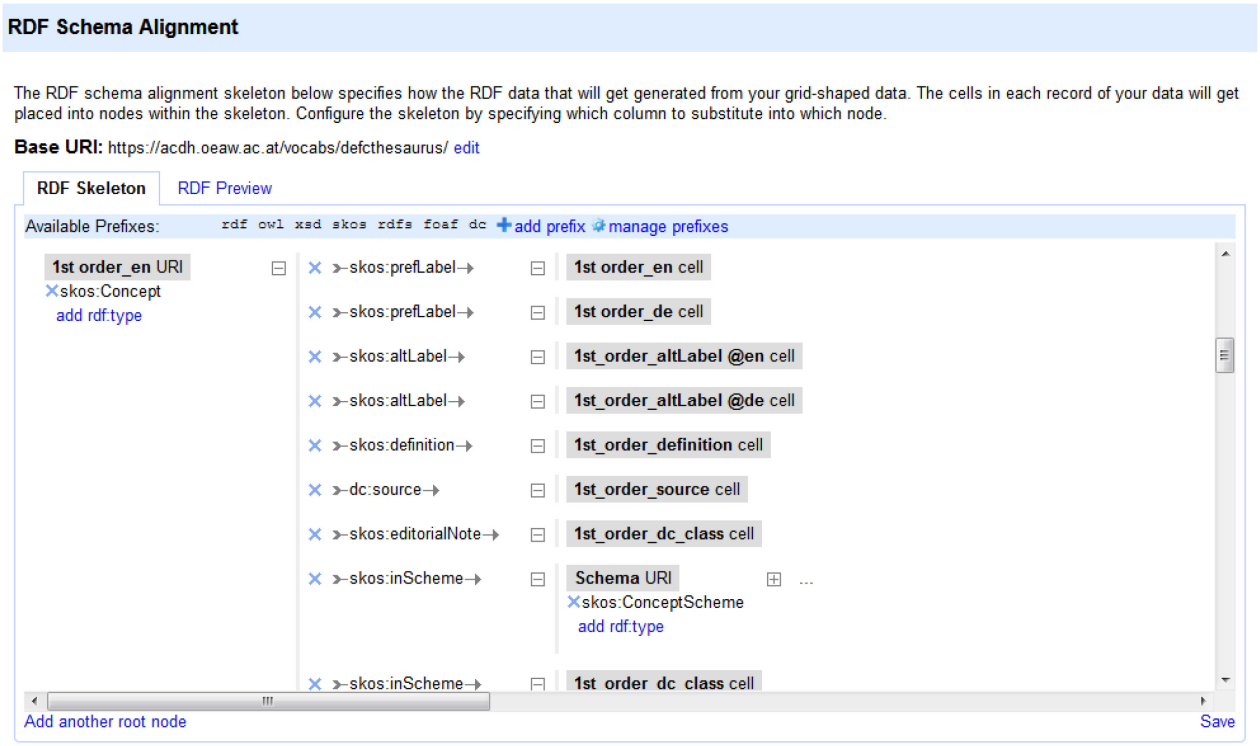
SKOS stands for a Simple Knowledge Organization System and is a w3C recommended data model for sharing and linking knowledge organization systems such as thesauri and other types of controlled vocabulary (classification schemes, subject heading lists and taxonomies) within the framework of Semantic Web.

SKOS uses Resource Description Framework (RDF) to encode the information, which is the standard model for data interchange on the Web.

## Mapping with OPEN Refine

Open Refine (previously known as Google Refine) is a free tool for cleaning, transforming and extending the data. Using the RDF extension the tool has been used to build an RDF skeleton using standard vocabularies of SKOS for the thesaurus concepts and Dublin Core for their metadata (source of the term's definition).

1. The table of each top class (site, area, research event, finds, interpretation) was imported separately and was first cleaned to make the data consistent (all English names start with a lower case, deleting 0 values etc.).
2. When building the RDF skeleton, first the concept schema has to be defined. A skos:ConceptScheme is an umbrella that includes all other concepts (in our case, the concept schema is represented by the top class, e.g. site, research event…). This means that an additional column holding this information had to be added to the table.
3. Mapping to SKOS started with defining the broadest (1st order) SKOS concept and assigning the preferred and alternative labels (English and German) as well as its definition and its source.
4. Finally the skos:Concept has been linked to the ConceptScheme it belongs to and to its narrower concept (2nd order).
5. Once the 2nd order concept has been added, the procedure described in points 3 and 4 has been repeated for all narrower concepts consecutively. Additionally, not only the narrower but also the broader concept had to be specified (e.g. both: *2nd order-has narrower term-3rd order* as well as *2nd order-has broader term-1st order*).




When all concepts got assigned their unique URIs, related and exact matches were created in a separated mapping.

Several terms appear several times in the controlled vocabularies (e. g. stone: building material/find material, fireplace: installation/occupation evidence), but represent the same concept and are therefore represented by the same URI. Because these terms have several broader terms they appear several times in a hierarchical view of the thesaurus, however, only one concept has been defined for all of them.

All mappings were imported as RDF files and joined in one file which was validated via Skosify (https://code.google.com/archive/p/skosify/) tool. SKOS play service was then used to visualize the thesaurus as a hierarchical tree on the DEFC homepage.

## Future work

The structure of DEFC thesaurus follows the DEFC app structure, because it aims to explain and contextualize DEFC terminology and simplify data retrieval process. In this form it can be used also by other projects dealing with similar subject, however, slight restructuring might be needed. We are hoping that a new, general thesaurus including DEFC terminology will be built in the framework of future projects.

## References

Hodge, G. (2000). Systems of Knowledge Organization for Digital Libraries: Beyond traditional authority files. *Report of The Digital Library Federation Council on Library and Information Resources*.

Tudhope, D., & Lykke Nielsen, M. (2006). Introduction to Knowledge Organization Systems and Services. *New Review of Hypermedia and Multimedia, 12*(1), 3-9. doi: 10.1080/13614560600856433