Thesaurus Newsletter

Quarterly update from the CAB Thesaurus management team

Issue H2 – 2021

Upcoming meetings

20th International Semantic Web Conference: 24-28 October 2021 (online) iswc2021.semanticweb.org/

KMIS 2021 - 13th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management: 25-27 October 2021, Valletta, Malta. http://www.kmis.ic3k.org/

Bite-Sized Taxonomy Boot Camp, 10 November 2021, London, UK (online) www.taxonomybootcamp.com/London/2021/

Taxonomy Boot Camp Connect: 15-16 November 2021, Washington DC, USA. www.taxonomybootcamp.com/

KM World 2021: 15-18 November 2021, Washington DC, USA. https://www.kmworld.com/Conference/

V Congresso ISKO Espanha-Portugal: 25-26 November 2021, Lisbon, Portugal. Theme: Organização do Conhecimento no Horizonte 2030/Organización del conocimiento en el Horizonte 2030. www.isko2021.letras.ulisboa.pt

KO-ED KOS Workshop: Introduction to Universal Decimal Classification: 25 November-9 December 2021 (online) https://www.iskouk.org/event-4516930

TOTh 2022: 2-3 June 2022, University Savoie Mont-Blanc, Chambery, France. Theme: Terminology & Ontology – theories and applications. toth.condillac.org/conference

17th International ISKO Conference: 6-8 July 2022, Aalborg, Denmark. Theme: Knowledge Organization Across Disciplines, Domains, Services and Technologies www.communication.aau.dk/events/

CABT Q3 2021 published

The CAB Thesaurus Q3 2021 edition went live on 20th October. Combining all languages, the thesaurus now exceeds 3.1 million terms (labels).

Since releasing the previous quarterly edition on 5th July 2021:

- 4093 terms were updated
- 1518 new terms were added in English
- 966 translations were added from English, particularly into German, Russian, French, Spanish, Dutch and Portuguese.
- Revision of dicot plant families continued, using the current APG IV system of classification. Another 33 families were updated, filling gaps especially in the tropical flora, including important forest trees, edible wild fruits, and ornamentals.
- The current developmental and economic statuses of all countries were revised according to the classification by World Bank and United Nations Development Programme. Older terminology was deprecated.
- Authority files needed for re-indexing the entire CAB Abstracts and Global Health databases were generated directly from the thesaurus. For further details, please see the article on page 2.

The full report is available online via the thesaurus web site.

Dealing with homonyms

Homonyms or homographs are words that have identical spelling but have different meanings. They are very common. In CABT there are nearly 800. Thus, homonym disambiguation is important in applications that use thesauri, such as indexing, particularly in automated systems that use natural-language processing, and other knowledge organization systems.

CABT usually adds disambiguating terms in parentheses as well as using differing subject category codes. Examples are Alabama, the US state, and Alabama (Lepidoptera). Also, ammonia and Ammonia (Protozoa). Sometimes CABT uses different words to make the distinction clear. For example, Cancer is used for a genus of crabs, but 'neoplasms' for the disease.

However, the great majority of homonyms in CABT are names of organisms, mostly at generic rank. For example, Coniophora (Diptera) and Coniophora (Fungi). Since there are separate rules for botanical and zoological nomenclature, the same name can legitimately be used for an animal and a plant or fungus, but not if both are plants, fungi or animals. In the latter case rules of priority are applied. Generally, but not always, the oldest validly published name is the one that should be used.

Obtaining thesaurus files

If your organization is interested in obtaining thesaurus data for your projects they are available in multiple formats, including plain text, CSV, XML and SKOS. If you wish to see beforehand what to expect sample data are available for download via the thesaurus web site. Please contact us to discuss your requirements. CABI has offices in a dozen countries.

Thesaurus contact

Email: Anton Doroszenko

CABI databases re-indexed

To keep the CAB Abstracts and Global Health databases aligned with CAB Thesaurus they are periodically re-indexed. The last time this was done was 18 months ago. However, henceforth the re-indexing will happen every quarter to coincide with every new edition of the thesaurus.

The indexing scheme in CABI databases is complex. There are separate indices for organism name descriptors, other descriptors, geographic locations, and broader terms generated from organism descriptors and geographic locations. These are all preferred terms in the thesaurus. In addition, all other non-preferred thesaurus terms are placed in the identifiers index field.

Numerous authority files are needed for this exercise, six of which are generated directly from the thesaurus.

Briefly, (1) a synonym file adds the preferred taxonomic name of an organism if the non-preferred taxonomic name is present as a descriptor; (2) a file identifies all preferred terms of organisms in order to place them correctly in the organism descriptor field; (3) a list of all geographic terms; (4) a list of all preferred thesaurus terms, including organisms; (5) a file that allows us to switch all descriptors that are non-preferred non-taxonomic terms to the corresponding preferred term; and (6) a file that adds the full hierarchy of broader-term descriptors for geographic terms and preferred organism names.

As a result, in the week beginning 13th September 2021, about 18% of 14.1 million records in the CAB Abstracts and Global Health databases (approximately 2.5 million records) were updated.