Vision for modular taxonomy production at Elsevier: The VOICE project

Wytze J. Vlietstra[1,*], Matthias Albus[1], Nick Drummond[2], Simon Jupp[2] and George Georghiou[1]

[1]Elsevier B.V., Radarweg 29, Amsterdam, Noord-Holland, 1043 NX Netherlands

[2]SciBite Limited, BioData Innovation Centre, Wellcome Genome Campus Hinxton, Cambridge CB10 1DR, United Kingdom

Abstract

Elsevier aims to streamline taxonomy production by creating a shared infrastructure supported by automation. In this presentation we will explain the components of this infrastructure, which include a candidate pool for incoming candidate terms. Here candidates are enriched with tools such as a synonym suggestion classifier, a term categorization classifier, an ambiguity scorer, and a hierarchical relationship suggestor. In the future, we want to move to a domain-based architecture, in which pre-built branches for specific scientific domains are maintained, and the taxonomy compiler, which chooses from these ”modules” to create taxonomies for specific products.

Keywords

Taxonomies, Taxonomy production, Taxonomy tooling

1. Background

Taxonomies drive many of Elsevier’s products. They both support searching through our scientific literature corpora, as well as extracting knowledge from publications and patents with NLP techniques. Taxonomies group synonymous terms together to represent concepts, potentially further enriching them with commonly used identifiers such as UniProt identifiers. Concepts within taxonomies are hierarchically organized, allowing some flexibility of what the hierarchical relationship represents exactly.

For each product, Elsevier currently develops and maintains a separate taxonomy, supported by a dedicated team of subject matter experts (SMEs). Up until now, these teams have worked in a siloed manner, reusing relatively little taxonomy data curated by other teams, and each developing their own set of tools. As a result, taxonomy production processes were poorly supported by automation, leading to many tasks being performed manually by SMEs.

To improve taxonomy production and reuse of their contents, the VOICE project (Vision for Ontological Interoperability & Content Enhancement) project was started. Based on an analysis of all processes around the production of taxonomies, four objectives were defined: Maintaining the high quality of our taxonomies, creating a shared taxonomy production pipeline supported by state-of-the-art automation, improve reuse of existing curated taxonomy data, and ensuring their FAIR compliance.

SSWAT4HCLS 2024: The 15th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences, February 26–29, 2024, Leiden, The Netherlands

[*]Corresponding author.

w.vlietstra@elsevier.com (W. J. Vlietstra)
0000-0002-1096-8563 (W. J. Vlietstra)

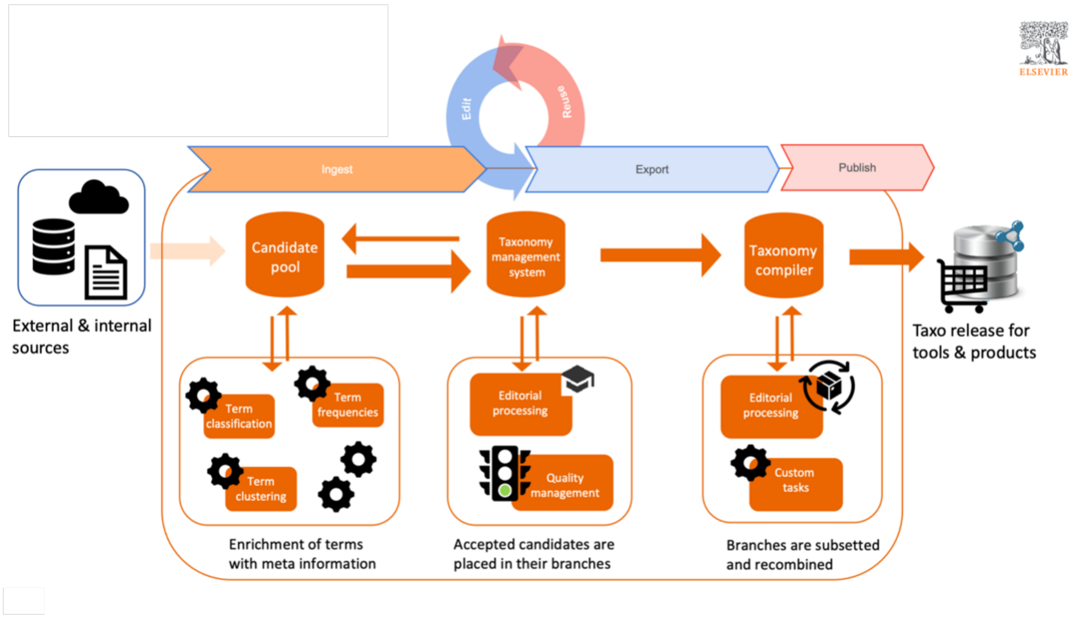
ELSEVIER
Edit
Reuse
Ingest
Export
Publish
External & internal sources
Candidate pool
Taxonomy management system
Taxonomy compiler
Taxo release for tools & products
Term classification
Term frequencies
Term clustering
Enrichment of terms with meta information
Editorial processing
Quality management
Accepted candidates are placed in their branches
Editorial processing
Custom tasks
Branches are subsetted and recombined

Figure 1: Schematic overview of our vision for taxonomy production. It consists of three processes: 1) Gathering and enriching candidate terms from all the various in a candidate pool. The output of the candidate pool are so-called proto-concepts. 2) Placing proto-concepts within a taxonomy for further editorial processing. 3) Compiling taxonomies based on pre-specified rules to create a product-taxonomy. Please note that the figure describes processes, which may be supported by the same system.

2. Candidate pool & services

Our initial focus was on the part of the taxonomy production process where we estimated most could be gained: developing a shared infrastructure and set of services for processing candidate terms. This infrastructure consists of a so-called candidate pool, shown in Figure 1, which is a triple store that stores each candidate term supplied to us, along with several services that enrich these candidate terms. Storing candidate terms and their enrichments in a candidate pool enables memorization of previous assessments of candidate terms, thereby allowing for quick comparisons of new candidates with existing data and eliminating the need for their repeated assessment. The candidate term enrichment services enable normalizing terms to the lexical variant preferred by specific taxonomies, counting their frequencies in our literature corpora, categorizing them to different scientific domains to efficiently assign them to the SME specialized in that domain, and clustering them with their synonymous terms. Additional services, such as hierarchy suggestion and ambiguity scoring are currently on our roadmap. The output of the candidate pool are so-called proto-concepts, which are collections of synonymous terms, which ideally also contain a suggestion on where they should be placed within a specific taxonomy in the taxonomy management system.

3. Domain-oriented taxonomy architecture

To improve the reuse of existing taxonomy data, we aim to move from a product-oriented architecture for our taxonomies to a domain oriented one. In a domain-oriented architecture, each scientific domain would be represented by a single pre-built taxonomy branch. Product taxonomies would then be able to select their required subset (i.e. concepts and labels) from these pre-built branches, combining them with selections from other pre-built branches covering other domains. The result would then be an equivalent product taxonomy as currently is being produced but eliminating duplicated taxonomy curation efforts. To support different needs of different product taxonomies, such a domain-based architecture would require a number of advanced features. For example, to support different granularities of concepts, concepts would need to be specified at their maximum granularity by default (e.g. different brand names of drugs are not considered to be synonymous to each other). Coarser granularities of concepts could then be achieved by "rolling up" child concepts to a pre-defined parent. Other modifications would include filtering out specific subsets of labels, using different preferred labels, and automatically adding qualifiers to terms that occur in multiple scientific domains and will therefore be ambiguous. Many of these features would be supported by what we refer to as a taxonomy compiler, shown in Figure 1, which would perform these operations based on pre-specified rules, which sometimes require flags to be assigned to concepts or labels.

4. Outlook

Ultimately, the VOICE project should lead to comprehensive and up to date taxonomies, the quality of which is guaranteed by Elsevier SMEs, which are produced with such an efficient process that we can be highly responsive to new use cases or customer requests. Although there remain to be open questions around e.g. the feasibility of the domain-based architecture of taxonomy data, we believe our achievements up until now have put us firmly on the path to reaching these goals.