Semantic Indexing of Open Scientific Literature to Help Users Discover and Navigate through Publications Networks

Franck Michel[‡], Anne Toulet[§], Anna Bobasheva[|], Marie-Claude Deboin[§], Sébastien Dupré[§], Aline Menin[|], Marco Winckler[|], Andon Tchechmedjiev[¶]

‡ University Cote d'Azur, CNRS, Inria, Sophia-Antipolis, France
§ CIRAD, Montpellier, France
| University Cote d'Azur, Inria, CNRS, Sophia Antipolis, France
¶ Euromov Digital Health in Motion, Univ Montpellier, IMT Mines Ales, Ales, France

Corresponding author: Franck Michel (franck.michel@inria.fr)

Abstract

In recent years, several evolutions have drastically transformed the way researchers as well as scientific and technical information (STI) services interact with scientific literature. The amount and pace of publications are skyrocketing, whether in journals and conferences or through pre-publication repositories (e.g., arxiv.org), such that it is increasingly difficult to keep up, find and make sense of relevant articles. Furthermore, the specialization of research communities makes it difficult to discover cross-disciplinary knowledge, which is essential to meet the growing demand of funding agencies for interdisciplinary projects. Scientific open archives are central in this landscape, however the keyword-based search services that they usually provide fail to grasp the semantic relationships between articles. Therefore, it is necessary to develop new tools that allow users to find their way in this mass of knowledge.

In this talk, we wish to present the methods, tools and services implemented in the ISSA*[2] project to address these needs, and discuss how they could fit and be deployed in the biodiversity area. Guided by the open science goals and embracing the FAIR*[1] principles, the project aims to:

1. provide a generic, transferable and extensible pipeline for the analysis and processing of the articles of an open scientific archive;
2. turn the processing results into a semantic index stored and published as a public RDF knowledge graph;
3. develop innovative search and visualization services that leverage this semantic index to allow researchers, decision makers or STI professionals to explore thematic association rules, networks of co-publications, articles with co-occurring topics, etc.

The semantic index construction process involves several artificial intelligence techniques: natural language processing, knowledge engineering and Semantic Web. These techniques are used to process the publications' metadata and text to automatically extract thematic descriptors and named entities. These descriptors and named entities are linked to knowledge bases such as Wikidata, DBpedia and GeoNames, or domain-specific terminological resources suited to the archive's domain. The semantic index linked with the third-party resources serves as a keystone to support the development of rich search and visualization tools aimed at researchers and/or STI professionals.

We demonstrated the effectiveness of this solution in the use case of Agritrop, an institutional archive of 110,000+ resources among which are 12,000 open access articles, specialized in the fields of agronomy, biodiversity and sustainable development. In this context, the Agrovoc multilingual thesaurus was used as a domain-specific reference vocabulary. Fig. 1 illustrates how the concepts mentioned in the articles of the archive can be used to discover and visualize association rules. In this example, articles mentioning concepts *COVID-19* and *food security* (a) frequently mention concept *pandemics* (b). Fig. 2 shows how other visualization techniques can help users search articles mentioning concept *health* or any of its sub-concepts (a and b), discover that it is often co-mentioned with *climate change* (c), and get the list of related publications (d) and their time distribution (e).

Being designed as a generic, transferable solution, the pipeline and visualization tools delivered by ISSA could be easily adapted to open archives of biodiversity literature. Typically, terminological references such as Darwin Core Terms, Access to Biological Collection Data (ABCD), open Digital Specimens (openDS), Audubon Core Metadata Schema as well as various taxonomic registries, could be considered for the description of an article's metadata or the linking of thematic descriptors and named entities. From there, the proposed visualization techniques could easily be reconfigured to explore the articles from a biodiversity open archive to answer various competency questions, for instance: what are the articles that mention a taxon or any of its child taxa? What are the museums/ institutions that are more frequently mentioned together with certain taxonomic groups? What are the research topics that frequently co-occur with climate change, and how do these topics evolve through the years? What public policies frequently occur in articles that mention endangered species? Furthermore, the pipeline could be extended by including existing third-party tools to carry out e.g., the extraction of relationships between entities or the reconciliation of authors' names.

Keywords

data indexing, knowledge graph, data visualization, scientific archive

Presenting author

Franck Michel

Presented at

TDWG 2022

Conflicts of interest

Endnotes

[*1] Findable, Accessible, Interoperable, Reusable

[*2] ISSA stands for *Semantic Indexing of a scientific archive Associated Services*

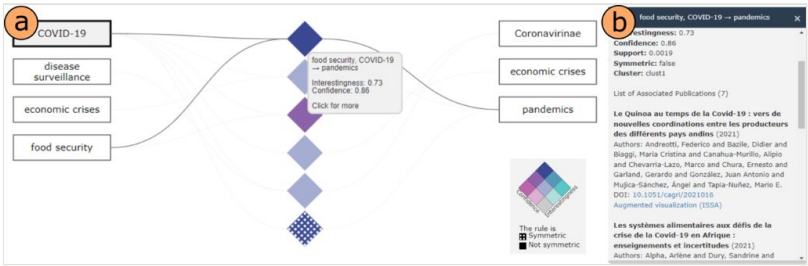

Figure 1.

Association rule stating that articles mentioning concepts *COVID-19* and *food security* (a) also frequently mention the *pandemics* concept (b).

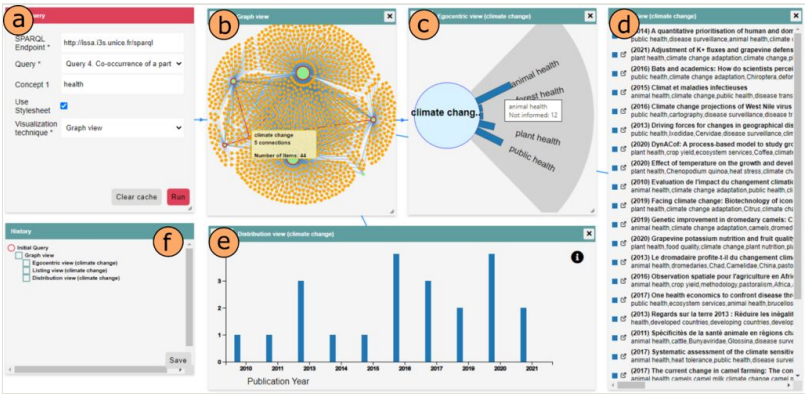

Figure 2.

Exploration of the relationship between concepts *health* and *climate change* or any of their sub-concepts.