



Elsevier Fingerprint Engine

Thesauri and Vocabularies

May 2024

1.1 Introduction

The Elsevier Fingerprint Engine is a back-end software system using state-of-the-art Natural Language Processing (NLP) techniques to extract information from unstructured text. Leveraging domain-relevant vocabularies and thesauri, it turns scientific publications of various types into semantic 'fingerprints': collections of weighted key concepts. Extracted fingerprints are subsequently used to power Elsevier products such as [Pure](https://www.elsevier.com/solutions/pure)¹, [SciVal](https://www.elsevier.com/solutions/expert-lookup)², or [ScienceDirect](https://www.elsevier.com/solutions/sciencedirect/topics)³: to summarize research output, offer concept-based search and matching, or present summarized knowledge on important scientific concepts.

To facilitate concept extraction and fingerprint generation across all sciences, a much-requested capability, Elsevier's *OmniScience taxonomy* is leveraged. Applications requiring highly specific subject areas are supported with well-established *special interest thesauri*, for instance with the [IAEA INIS thesaurus](https://inis.iaea.org/search/thesaurus.aspx)⁴ covering nuclear science and technology. Terminology that has not yet made it to expert-curated thesauri and taxonomies is captured using keyphrase extraction, which relies on natural language processing and machine learning to identify previously unseen concepts. Terminology that is captured in this manner constitute the *Keyphrase Thesaurus*, which supplements regular thesauri, vocabularies, and taxonomies.

This document provides further details about three major sources for concept annotation with the Fingerprint Engine: *OmniScience* (with all its subject categories), *IAEA INIS thesaurus*, and the *Keyphrase Thesaurus*.

OmniScience

OmniScience is an all-science taxonomy that has been developed by Elsevier to facilitate concept annotation of research content across Elsevier platforms and products. It connects to terms from established external and Elsevier-owned vocabularies, taxonomies, and thesauri across all domains of scientific research. For instance, OmniScience incorporates terms from the popular [Medical Subject Headings](https://www.nlm.nih.gov/mesh/meshhome.html)⁵ (MeSH), the [NASA thesaurus](https://www.sti.nasa.gov/nasa-thesaurus/)⁶, and the [UNESCO thesaurus](https://vocabularies.unesco.org/browser/thesaurus/en/)⁷, as well as terms from the specialist taxonomies used to power other Elsevier-owned databases such as [Ei Compendex](https://www.elsevier.com/solutions/engineering-village/content/compendex)⁸ (Engineering), [Reaxys](https://www.elsevier.com/solutions/reaxys)⁹ (Chemistry), or [Emtree](https://www.elsevier.com/solutions/embase-biomedical-research/emtree)¹⁰ (Life Sciences). Terms are also derived from other sources, like book indexes and article abstracts or author keywords from [Scopus](https://www.elsevier.com/solutions/scopus)¹¹ and further curated by a team of in-house taxonomists and subject matter experts on an on-going basis to ensure actuality.

The full taxonomy is updated quarterly. Updates usually involve newly added terminology, further improved scientific domain coverage, or replaced or removed terms that have fallen out of use or became antiquated.

¹ <https://www.elsevier.com/solutions/pure>

² <https://www.elsevier.com/solutions/expert-lookup>

³ <https://www.elsevier.com/solutions/sciencedirect/topics>

⁴ <https://inis.iaea.org/search/thesaurus.aspx>

⁵ <https://www.nlm.nih.gov/mesh/meshhome.html>

⁶ <https://www.sti.nasa.gov/nasa-thesaurus/>

⁷ <https://vocabularies.unesco.org/browser/thesaurus/en/>

⁸ <https://www.elsevier.com/solutions/engineering-village/content/compendex>

⁹ <https://www.elsevier.com/solutions/reaxys>

¹⁰ <https://www.elsevier.com/solutions/embase-biomedical-research/emtree>

¹¹ <https://www.elsevier.com/solutions/scopus>

The taxonomy is structured into four large domains, sub-divided into 21 subject categories that have proven to be useful in describing sciences and specific areas of interest. As per 1st quarter of 2024 (version OS24Q2), these categories contain around 869 thousand concepts (567 thousand unique concepts across categories). The domains and subject categories distinguished are the following:

Parent Domain	Subject Category	
Health Sciences	NUR	Nursing and Health Professions
	MED	Medicine and Dentistry
	PHA	Pharmacology, Toxicology and Pharmaceutical Science
	VET	Veterinary Science and Veterinary Medicine
Life Sciences	AGR	Agricultural and Biological Sciences
	NEU	Neuroscience
	BGM	Biochemistry, Genetics and Molecular Biology
	IMM	Immunology and Microbiology
	FSC	Food Science
Physical Sciences & Engineering	CHE	Chemical Engineering
	CHM	Chemistry
	EPS	Earth and Planetary Sciences
	MTS	Materials Science
	COS	Computer Science
	ENG	Engineering
	MAT	Mathematics
	PHY	Physics and Astronomy
Social Sciences & Humanities	ECO	Economics, Econometrics and Finance
	PSY	Psychology
	SOC	Social Sciences
	HUM	Arts & Humanities

Concept extraction from text can be restricted to any set of subject categories, to zoom in to a specific topic and avoid concepts from relevant subject categories to be annotated. With much content, for instance with articles from highly specific journals, the subject category is known beforehand. If this aspect is unknown, a common procedure is to first classify the text into subject categories, to then limit the concept extraction accordingly.

INIS Thesaurus

The International Nuclear Information System (INIS) thesaurus is published and maintained by the International Atomic Energy Agency (IAEA) and receives yearly updates on FPS. The current version used by the Elsevier Fingerprint Engine counts around 31 thousand concepts.

INIS itself is the world's leading database on the peaceful uses of nuclear science and technology, containing over 3 million bibliographic records. The subjects covered by the thesaurus range from nuclear engineering, safeguards, and non-proliferation to applications in agriculture and health.

The following subject areas are covered by the INIS thesaurus:

S01 Coal, lignite, and peat
S02 Petroleum
S03 Natural gas
S04 Oil shales and tar sands
S07 Isotopes and radiation sources
S08 Hydrogen
S09 Biomass fuels
S10 Synthetic fuels
S11 Nuclear fuel cycle and fuel materials
S12 Management of radioactive wastes, and non-radioactive wastes from nuclear facilities
S13 Hydro energy
S14 Solar energy
S15 Geothermal energy
S16 Tidal and wave power
S17 Wind energy
S20 Fossil-fueled power plants
S21 Specific nuclear reactors and associated plants
S22 General studies of nuclear reactors
S24 Power transmission and distribution
S25 Energy storage
S29 Energy planning, policy and economy
S30 Direct energy conversion
S32 Energy conservation, consumption, and utilization
S33 Advanced propulsion systems
S36 Materials science
S37 Inorganic, organic, physical and analytical chemistry
S38 Radiation chemistry, radiochemistry and nuclear chemistry
S42 Engineering
S43 Particle accelerators
S46 Instrumentation related to nuclear science and technology
S47 Other instrumentation
S54 Environmental sciences
S58 Geosciences
S60 Applied life sciences
S61 Radiation protection and dosimetry
S62 Radiology and nuclear medicine
S63 Radiation, thermal, and other environmental pollutant effects on living organisms and biological materials
S70 Plasma physics and fusion technology
S71 Classical and quantum mechanics, general physics
S72 Physics of elementary particles and fields
S73 Nuclear physics and radiation physics
S74 Atomic and molecular physics
S75 Condensed matter physics, superconductivity and superfluidity
S77 Nanoscience and nanotechnology
S79 Astrophysics, cosmology and astronomy
S96 Knowledge management and preservation
S97 Mathematical methods and computing
S98 Nuclear disarmament, safeguards and physical protection

Keyphrase Thesaurus

The Keyphrase Thesaurus is a fully data-driven, automatically generated all-domain thesaurus. Its main objective is to provide access to more detailed as well as more recent scientific terminology compared to traditional, manually curated thesauri. It can be used either on its own or alongside existing domain thesauri to enhance and enrich fingerprints.

It uses a combination of state-of-the-art machine learning and rule-based approaches to find and filter candidate terms, cluster them into concepts and select appropriate preferred labels. The main source of candidate terms are author keywords assigned to research articles in Elsevier's Scopus database.

The Keyphrase Thesaurus is updated annually.

The current version contains more than 2 million concepts.