Elsevier Fingerprint Engine

Thesauri and Vocabularies

October 2022

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[1], Expert Lookup[2], or ScienceDirect[3]: to summarize research output, offer concept-based search and matching, or present summarized knowledge on important scientific concepts. The system also provides capabilities to identify and extract novel concepts from text, to enrich existing thesauri or generate completely new vocabularies.

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[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 two major sources for concept annotation with the Fingerprint Engine: *OmniScience* (with all its subject categories) and the *IAEA INIS 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[5] (MeSH), the NASA thesaurus[6], and the UNESCO thesaurus[7], as well as terms from the specialist taxonomies used to power other Elsevier-owned databases such as Ei Compendex[8] (Engineering) , Reaxys[9] (Chemistry), or Emtree[10] (Life Sciences). Terms are also derived from other sources, like book indexes and article abstracts or author keywords from 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 adding new terminology, further improving scientific domain coverage, or replacing or removing 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 which can be sub-divided into 21 subject categories that have proven to be useful in describing sciences and specific areas of interest. As per the 4th quarter of 2022, these categories contain 823,175 concepts (532,849 unique concepts across categories) described by 1,910,956 terms. The subject categories with corresponding unique concept counts are the following:

Parent Domain	Subject Category		Concepts
Health Sciences	NUR	Nursing and Health Professions	58,601
	MED	Medicine and Dentistry	80,538
	PHA	Pharmacology, Toxicology and Pharmaceutical Science	56,372
	VET	Veterinary Science and Veterinary Medicine	8,315
Life Sciences	AGR	Agricultural and Biological Sciences	72,142
	NEU	Neuroscience	28,857
	BGM	Biochemistry, Genetics and Molecular Biology	34,421
	IMM	Immunology and Microbiology	31,896
	FSC	Food Science	2,473
Physical Sciences & Engineering	CHE	Chemical Engineering	2,322
	CHM	Chemistry	124,179
	EPS	Earth and Planetary Sciences	31,498
	MTS	Materials Science	7,075
	COS	Computer Science	69,791
	ENG	Engineering	90,987
	MAT	Mathematics	28,781
	PHY	Physics and Astronomy	13,658
Social Sciences & Humanities	ECO	Economics, Econometrics and Finance	5,445
	PSY	Psychology	9,197
	SOC	Social Sciences	22,834
	HUM	Arts & Humanities	43,793

Concept extraction from text can be restricted to any set of subject categories, which could be helpful 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.

The OmniScience taxonomy has a poly-hierarchical nature: a concept can be a child to more than a single parent concept. For instance, a concept like “Medical Imaging” is not only included in subject category Neuroscience but also included in subject category Medicine and Dentistry. Consequently, subject categories can have a certain concept overlap with each other. This is especially the case with adjacent or highly related areas. Moreover, some fields tend to be represented by multiple categories. Linguistics, for instance, is generally covered by Arts & Humanities, but concepts related to Sociolinguistics can also be found in the Social Sciences subject category.

The following matrix shows to what extent any two subject categories have concepts in common, where cells indicate the percentage of all concepts combined that is shared. Cells for pairs of subject categories with less than 1% overlap were left blank. For instance, the subject categories Immunology and Microbiology (IMM) and Biochemistry, Genetics and Molecular Biology (BGM) have – not surprisingly – about 45% of concepts in common.

	NUR	MED	PHA	VET	AGR	NEU	BGM	IMM	FSC	CHE	CHM	EPS	MTS	COS	ENG	MAT	PHY	ECO	PSY	SOC
MED	53.5																			
PHA	59.7	50.7																		
VET	2.6	3.6	3.2																	
AGR	3.9	7.0	9.7	4.0																
NEU	25.8	27.6	26.8	8.7	5.8															
BGM	10.6	17.5	21.8	5.3	14.5	18.6														
IMM	11.0	22.5	21.6	4.4	10.2	14.8	45.3													
FSC	0.5	0.9	0.8	0.4	1.8	1.1	1.6	1.0												
CHE	1.0	0.8	1.1	0.2	0.4	0.8	1.4	0.8	1.7											
CHM	4.6	4.7	4.9	0.6	1.6	3.1	2.9	1.2	0.4	1.3										
EPS	2.4	2.5	2.0	0.6	3.5	1.4	3.2	2.6	0.6	2.7	1.9									
MTS	1.0	1.6	1.1	0.2	0.6	1.3	3.0	2.1	3.3	8.7	2.5	4.4								
COS	0.4	0.4			0.2	0.2	0.4	0.3		0.1	0.2	1.2	0.4							
ENG	0.7	0.8	0.6	0.1	0.7	0.3	0.9	0.7	0.1	1.5	1.6	3.6	3.4	5.0						
MAT	0.3	0.2	0.1		0.2	0.2	0.4	0.3		0.2	0.2	1.3	0.3	3.9	2.8					
PHY	1.1	1.4	1.1	0.5	1.0	1.2	2.5	1.6	0.7	3.8	1.8	21.4	11.2	1.1	3.2	1.0				
ECO	0.2	0.2			0.5	0.5	0.2			0.1		3.0	0.5	0.3	0.3	1.3	0.3			
PSY	1.4	2.4	0.7	1.0	0.6	3.7	2.0	1.2	1.0	0.5	0.3	1.1	0.8	0.8	0.3	1.1	0.9	2.6		
SOC	1.6	2.4	0.6	0.3	1.6	2.6	2.0	1.2	0.6	0.6	0.5	8.4	2.1	1.6	1.2	1.3	2.3	15.4	10.5	
HUM	0.7	0.8	0.3	0.1	0.6	0.5	0.7	0.5		0.1	0.2	1.7	0.2	1.3	0.5	0.6	0.4	0.7	1.0	4.8

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 (Jan 2022) used by the Elsevier Fingerprint Engine counts 31,139 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