Elsevier Fingerprint Engine®

Based on state-of-the-art Natural Language Processing (NLP) techniques, the Elsevier Fingerprint Engine® is a back-end software system that extracts information from the unstructured text of scientific documents. It applies a domain-relevant thesaurus to each scientific publication to map text to semantic 'fingerprints', or collections of weighted key concepts.

By identifying and extracting new concepts, the Elsevier Fingerprint Engine can enrich each thesaurus and generate new vocabularies so it continuously improves the insights it delivers to researchers and funding bodies. It can be used as a back-office processing component of applications or as a stand-alone service.

Out-of-the-box text analytics

To create a Fingerprint® index – a set of standardized, domain-specific concepts that define the text – the Elsevier Fingerprint Engine mines a document's unstructured text, including publication abstracts, funding announcements and awards, project summaries, patents, proposals and applications. It enables institutions to look beyond metadata by aggregating and comparing Fingerprint indices. The ideas extracted from documents help users to identify trends, and expose and analyze valuable connections between groups (researchers, funders, reviewers), organizations (institutions, associations) or geographic areas.

Integrated across multiple solutions

As a key component in SciVal®, Pure and Expert Lookup, the Elsevier Fingerprint Engine computes semantic representations for publications and other data types to allow for presentation, navigation and reporting of scientific output. It also serves as the framework for customized modules that enable funding agencies to find reviewers, analyze grant portfolios and strategically plan which areas of research to fund next.

The system's flexibility allows it to be applied in various ways to help each institution answer its most pressing questions. It's used to present author profiles and match authors and funding opportunities in Pure and Profile Refinement Services, to analyze and match grants to reviewers in Expert Lookup, to showcase competencies and detect trends in research areas for SciVal.

Thesauri that cover a wide range of subjects

To support documents and applications across a wide variety of subject areas, the Elsevier Fingerprint Engine integrates a range of thesauri, as well as enriching existing thesauri and developing stand-alone vocabularies. It can also employ subsets of thesauri or vocabularies to continuously update and enhance its semantic fingerprints.

The thesauri used to power Fingerprint Engine account for approximately 500,000 unique concepts and terms across scientific domains:

Domain	Thesaurus/Vocabulary	Number of terms per thesaurus
Life Sciences	MeSH thesaurus	274,000
	Embase thesaurus	80,000
Physics	NASA thesaurus	19,000
Agriculture	NAL thesaurus	73,000
Economics	Eco Humanities vocabulary	21,000
Social Sciences	Gesis thesaurus	21,000
Mathematics	Cambridge Math thesaurus, Math vocabulary	21,000
Geosciences	Geobase thesaurus	14,000
Engineering	Compendex thesaurus	12,000
Humanities	Humanities vocabulary	39,000
Compounds (Chemistry)	Compendex thesaurus, MeSH thesaurus	256,000

ELSEVIER

Creating Fingerprint indices

To identify domain-relevant concepts in a text based on a thesaurus or vocabulary, the Elsevier Fingerprint Engine incorporates a suite of tools for:

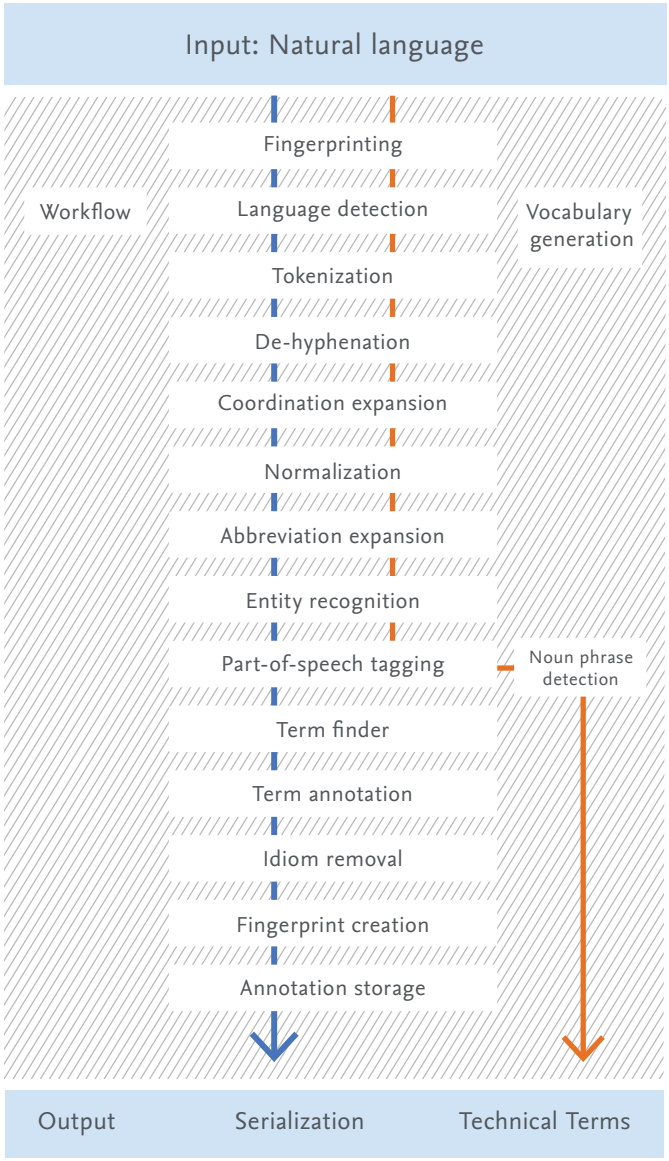
- pre-processing (sentence detection, eliminating hyphens)
- tokenization
- normalization
- expansion (abbreviations and coordinations)
- part-of-speech tagging
- pattern identification (chemical compounds, URLs, idioms)
- term disambiguation
- annotation

The tools enable the concept-finding algorithm to be sensitive to lexical and grammatical features (casing, word order) to distinguish different usages (for example, Windows® vs. windows, the noun form vs. the verb form of 'lead'). It also has the ability to ignore difference when it has no meaning (for example, tumour vs. tumor, kidney failure vs. failure of the kidney). It also considers the context of terms, looking to the neighboring text to avoid groupings that don't make sense (for example, non-Hodgkin Lymphoma with Hodgkin Lymphoma, seeing the 'tree of human ancestry' as a plant, or interpreting 'administration' as management in a text about a drug as a treatment for disease). Named entities, including the names of people and places, are identified for disambiguation across thesauri and vocabularies, and can be presented separately from the actual Fingerprint indices.

The document's concepts are weighted according to their frequency, occurrence in the text's title or body, and occurence in automatically detected sub-sections of a text's body. You can then aggregate the highest ranked or all ranked concepts of document Fingerprint indices into profiles of individual researchers, institutions or regions. While the Elsevier Fingerprint Engine is technically capable of handling all language, current applications focus on English as the scientific lingua franca.

Enriching terminology resources

In addition to identifying concepts, the Fingerprint Engine can help to enrich existing thesauri or vocabularies or to build new ones from scratch. Using a subset of the engine's NLP components, a Noun Phrase Detector extracts technical terms from document collections of specific domains.



ELSEVIER

Elsevier Fingerprint Engine®

For more information please contact your Elsevier representative.