Elsevier Fingerprint Engine

The Elsevier Fingerprint Engine is a back-end software system based on state-of-the-art Natural Language Processing (NLP) techniques to extract information from unstructured text. Applying a domain-relevant thesaurus to each scientific publication, the fingerprint engine maps text to semantic 'fingerprints', collections of weighted key concepts.

By identification and extraction of new concepts the Elsevier Fingerprint Engine can enrich each thesaurus and generate new vocabularies.

The Elsevier Fingerprint Engine can be used as a back-office processing component of applications or as a stand-alone service.

Out-of-the-box text analytics functionality

The Fingerprint Engine mines the unstructured text of scientific documents – publication abstracts, funding announcements and awards, project summaries, patents, proposals, applications and other sources – to map it to a ranked set of standardized, domain-specific concepts that define the text, known as a Fingerprint. By aggregating and comparing fingerprints, the engine enables institutions to look beyond metadata. Based on ideas extracted from documents, users can identify trends and expose and analyze valuable connections between groups like people (researchers, funders, reviewers etc.), organizations (institutions, associations) or geographic areas.

Used as a key component in several Elsevier products, such as SciVal®, Pure and SciVal Analytics, the Fingerprint Engine computes semantic representations for publications and other data types to allow for presentation, navigation and reporting on scientific output. The Elsevier Fingerprint Engine also serves as the framework for customized modules which enable funding agencies to find reviewers, analyze grant portfolios and strategically plan which areas of research to fund next. A flexible platform, the Elsevier Fingerprint Engine can be applied in various ways to help each institution answer its most significant questions.

Covering a wide range of subject areas with a collection of thesauri

The Elsevier Fingerprint Engine integrates a range of thesauri to support documents and applications pertaining to different subject areas like the Medical Subject Headings (MeSH), the National Agriculture Library's (NAL) thesaurus and Elsevier's Compendex thesaurus. To improve coverage we use the Fingerprint Engine to enrich existing thesauri (Cambridge Math thesaurus, Gesis thesaurus for the social sciences) and develop stand-alone vocabularies (e.g., for the humanities).

In its current standard configuration, the Elsevier Fingerprint Engine covers the following domains:

Domain	Thesaurus/Vocabulary
Life Sciences	MeSH thesaurus
Physics	NASA thesaurus
Agriculture	NAL thesaurus
Economics	Eco Humanities vocabulary
Social Sciences	Gesis thesaurus
Mathematics	Cambridge Math thesaurus, Math vocabulary
Geosciences	Geobase thesaurus
Engineering	Compendex thesaurus
Humanities	Humanities vocabulary
Compounds (Chemistry)	Compendex thesaurus, MeSH thesaurus

Subsets of thesauri/vocabularies can be employed and terminology sources provided by institutions can be implemented. Thesauri and vocabularies are continuously updated and enhanced.

Natural Language Processing (NLP) Modules

The Elsevier Fingerprint Engine framework facilitates configuration of a processing workflow. Multiple NLP modules are executed sequentially, using processing results generated by previous modules.

The standard NLP facilities of the Elsevier Fingerprint Engine can be complemented by third party text analytics modules, unlimited in type and number. The infrastructure enabling that consists of a .Net platform, a collection of text analysis modules and a host process.

Workflow: Fingerprinting

The Elsevier Fingerprint Engine identifies domain-relevant concepts in a text based on a thesaurus or vocabulary making use of a suite of tools for preprocessing (e.g. sentence detection, dehyphenization), tokenization and normalization, expansion (e.g. of abbreviations and coordinations), part-of-speech tagging, pattern identification (e.g. chemical compounds, urls, idioms), term disambiguation and, eventually, annotation.

By virtue of those tools, the concept finding algorithm is sensitive to lexical and grammatical features - casing, word order, part-of-speech and others when it must be - e.g., to distinguish Windows® from windows, the noun from the verb 'lead', etc. At the same time it ignores differences when they have no meaning - e.g., the differences between 'tumour' and 'tumor', between 'kidney failure' and 'failure of the kidney' etc.

In addition, concept finding takes into account the context of terms. It looks at their neighbors and will, e.g., not identify a "non-Hodgkin Lymphoma" as a Hodgkin Lymphoma or the ' tree of human ancestry' as a plant, but also at their wider environment and will not interpret 'administration' as management in a text about a drug as a treatment for a disease.

Concepts found in documents are weighted according to their frequency, their occurrence in a text's title or text body and, in a recent solution for Funding Opportunity Announcements, according to their occurrence in automatically detected subsections of a text's body.

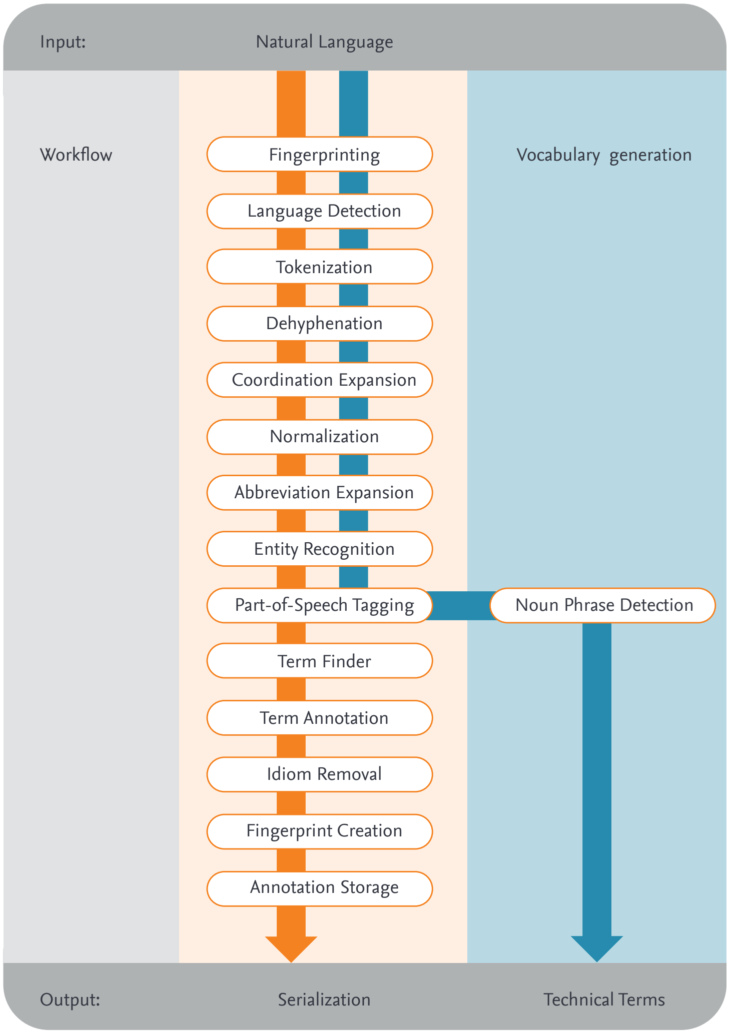
The most highly ranked or all ranked concepts of document fingerprints can be aggregated to profiles of individual researchers, institutions, regions etc. (see above).

So-called Named Entities like the names of people ('John O'Keefe') and places ('Philadelphia, Pennsylvania') are identified and disambiguated across thesauri and vocabularies and can be presented separate from fingerprints proper.

While the Fingerprint Engine is a technology that can technically handle all languages, current applications focus on English language support as the scientific lingua franca.

Workflow: Generation of Controlled Vocabularies and Enrichment of Thesauri

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



Applications of the Fingerprint Engine

All Elsevier Research Management products contain fingerprints for semantic enrichment and presentation purposes: Pure and Profile Refinement Services for presenting author profiles and matching authors and funding opportunities; Expert Lookup for analyzing and matching grants to reviewers; SciVal for showcasing competencies and detecting trends in research areas; SciVal Analytics for showing trends, profiling authors, institutions and departments, classifying research across several dimensions. End-users of the Fingerprint Engine include a number of funding bodies who use the tool for research classification – to report and gain deep insight into their funding efforts.

For more information please contact your Elsevier representative.

Empowering Knowledge™

ELSEVIER