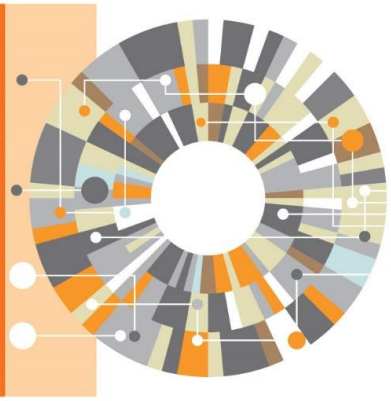
ELSEVIER

Elsevier Fingerprint Engine



May 2016

Elsevier Research Intelligence

Contents

Elsevier Fingerprint Engine™

The Elsevier Fingerprint Engine is a back-end software system of state-of-the-art Natural Language Processing (NLP) techniques to extract information from unstructured text. Applying domain-relevant thesauri to scientific publications of various types, the fingerprint engine creates an index of weighted terms, called concepts, which defines the text, known as a Fingerprint™ . Through the identification and extraction of new concepts the Elsevier Fingerprint Engine can enrich thesauri and generate new vocabularies.

By aggregating and comparing Fingerprints, the Elsevier Fingerprint Engine enables institutions to look beyond metadata and expose valuable connections among people, publications, funding opportunities and ideas.

The Elsevier Fingerprint Engine can be used as a back-office processing component of applications, as it is for a number of Elsevier products, or as a stand-alone service.

Out-of-the-box text analytics functionality

The Elsevier Fingerprint Engine offers out-of-the-box text analytics functionality that can be adapted to meet each institution's needs. Recognizing a growing need for more in- depth insight to support strategic research-related decisions, Elsevier develops inventive semantic solutions for academic and government institutions using the Elsevier Fingerprint Engine as an enabling technology. The Elsevier 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 Elsevier Fingerprint Engine enables institutions to look beyond metadata. Based on ideas extracted from documents users can identify trends and expose and analyze valuable connections between people (researchers, funders, reviewers etc.), organizations (institutions, associations), geographic areas etc.

Used as a key component in several Elsevier products, such as SciVal®, Pure and SciVal Analytics, the Elsevier 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 automatically generates author profiles for Pure (hosted edition) by scanning and analyzing publications from the Scopus® database and additional content provided by the institution. A directory of research expertise, Pure enables researchers, administrators and managers to enable collaboration within and outside of the organization. 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.

A wide range of subject areas are covered by a collection of thesauri

The Elsevier Fingerprint Engine integrates a range of thesauri to support applications pertaining to different subject areas, including a number of traditional popular ones 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, Geobase thesaurus) 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	Economics 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, terminology sources provided by institutions can be implemented. Thesauri and vocabularies are continuously updated and enhanced.

1. A look inside the Elsevier Fingerprint Engine

1.1. Workflow: Fingerprinting

The Elsevier Fingerprint Engine identifies relevant technical concepts in a text based on a thesaurus or vocabulary.

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, e.g., 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 separated from fingerprints proper.

While the Fingerprint Engine is language insensitive technology and can technically handle all language input, current applications focus on English language support as the scientific lingua franca.

1.2. Workflow: Generation of Controlled Vocabularies and Enrichment of Thesauri

In addition to identifying the concepts of given thesauri or vocabularies (see above) 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.

1.3. Framework and Natural Language Processing (NLP) Modules

The NLP workbench framework facilitates configuration of a processing workflow where multiple modules are executed sequentially, using processing results generated by previous modules.

The standard NLP facilities of the Elsevier Fingerprint Engine, described in more detail below, can be complemented by third party text analytic 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.

1.4. Key Natural Language Processing components of the Elsevier Fingerprint Engine

Language Detection: Identifies the language in which a text is written.

Tokenization: Splits text in tokens like words, punctuation marks and sentences.

Dehyphenation: Recognizes sentence-final hyphenations and reconstructs the original words, for instance replaces “dehyph- enation” with “dehyphenation”.

Coordination Expansion: Detects abbreviated coordinations and reconstructs full forms. For instance, the phrase “intra- and extramural” is expanded to “intramural and extramural”. Similarly, full noun phrases are reconstructed from compacted ones, e.g. “Hepatitis A and B” is expanded to “Hepatitis A and Hepatitis B”.

Normalization: Produces normal forms converting plural to singular forms (children > child) and British to American spelling variants (gynaecology > gynecology).

Abbreviation Expansion: Detects and expands abbreviations that are defined in a text. For instance, if the phrase “Blood group (BG)” is detected all occurrences of “BG” in the same text are expanded to “Blood group”.

Entity Recognition: Recognizes specific entities like email addresses, URLs, citations and chemicals using regular expressions. For example “\b[a-z]+kinase\b” recognizes simple enzymes while “[A-Z][a-z]+ \([0-9]+\)” recognizes simple citations.

Part-of-Speech Tagging: Tags tokens as linguistic parts of speech (verb, noun etc.) depending on their context. E.g., the word ‘lead’ will be tagged as verb in “This lead to the conclusion...”, as a noun in “....where lead concentration was high”.

Noun Phrase Detection [Alternative Workflow, see above]: Detects noun phrases in preanalyzed text (i.a. by the modules described above), the first step of suggesting new terms for an existing or a new thesaurus or vocabulary. For instance, processing “The novel association at the ABO locus provides evidence” produces “novel association”, “ABO locus” and “evidence”.

Term Finder: Finds occurrences of the terms of a thesaurus or vocabulary in preanalyzed text (i.a. by the modules described above).

Term Annotation: Marks thesaurus terms identified in text with flags providing further information about them, most notably to exclude terms from concept assignment (see below) when disambiguation routines found its meaning in the given context to differ from its meaning in the applied thesaurus.

Idiom Removal: Excludes known idioms from concept assignment. For instance, “on the other hand” will not produce the concept “hand”.

Fingerprint Creation: Assigns concepts to the remaining found terms and, based on a set of criteria (see above), assigns a weight to each concept.

2. Example solutions powered by the Elsevier Fingerprint Engine

2.1. National Institutes of Health

At the request of US Congress, in 2008 the National Institutes of Health (NIH) implemented a process to provide more consistency and transparency in the reporting of its funded research. This new process uses the Research, Condition, and Disease Categorization (RCDC) system, a custom solution powered by the Elsevier Fingerprint Engine, to mine the text of grant titles, abstracts, and specific aims, extract the terms used to describe the research being performed, and apply these terms to match projects to NIH-wide category definitions. The total funding for all projects in each research category represents NIH's best estimate based on the category definition. The NIH uses the system to categorize approximately $29 billion USD of research funding per year.

2.2. Wellcome Trust

The Wellcome Trust uses solutions driven by the Elsevier Fingerprint Engine to enhance its reporting capabilities. Restricted by inflexible, inconsistent and static reporting based on manually applied keyword terms, the Wellcome Trust introduced Elsevier's Fingerprint technology to allow them to define their own search categories for up to date, accurate, flexible and consistent reporting. The Wellcome Trust has also used the Reviewer Finder to mine the text of grant proposals, extract the terms used to describe each submission, and apply these terms to match applications to potential reviewers.