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 maps text to semantic 'fingerprints', collections of weighted key concepts. By identification and extraction of new concepts the Elsevier Fingerprint Engine can also enrich thesauri and generate new vocabularies.

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. This document gives a general description, provides instructions how to use the RESTful web API, and contains the following sections:

1. Covering a wide range of subject areas with 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, STW thesaurus for Economics) and develop stand-alone vocabularies (e.g., for the humanities). Together, the thesauri and controlled vocabularies allow to annotate text with concepts and terms for specific scientific domains. Where a domain is described by a thesaurus, which in contrast to vocabularies structure concepts in a hierarchy of broader and narrower concepts, the term annotations contain identifiers that link to concepts in the hierarchy, as exemplified in Figure 1 for MeSH based indexing.

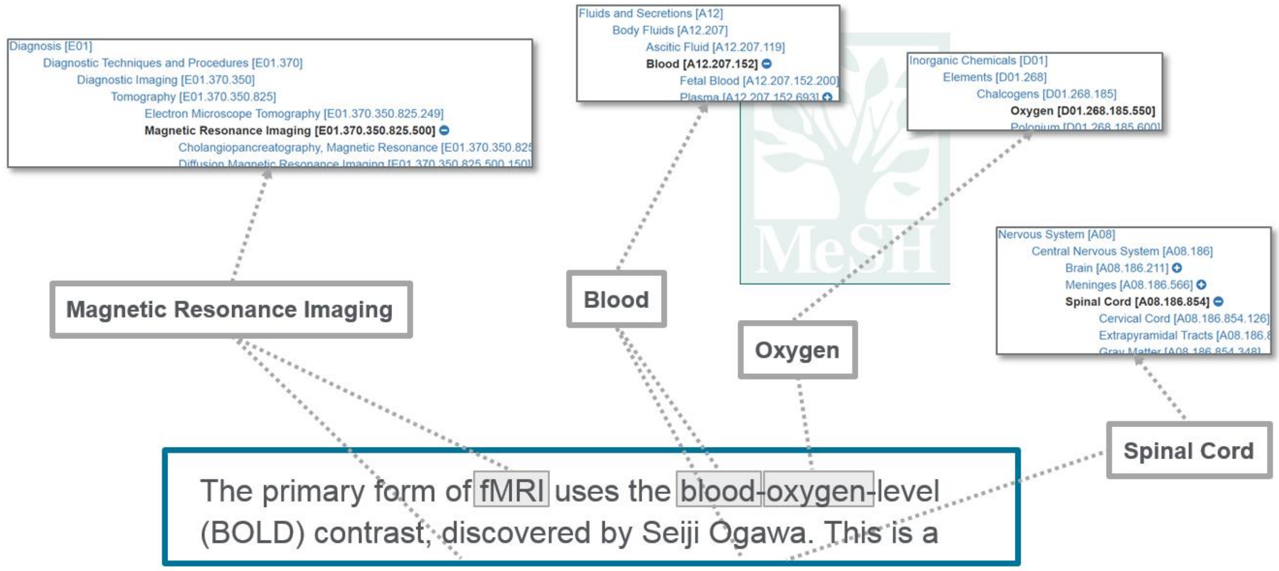

Diagnosis [E01]
Diagnostic Techniques and Procedures [E01.370]
Diagnostic Imaging [E01.370.350]
Tomography [E01.370.350.825]
Electron Microscope Tomography [E01.370.350.825.249]
Magnetic Resonance Imaging [E01.370.350.825.500]
Cholangiopancreatography, Magnetic Resonance [E01.370.350.825
Fluids and Secretions [A12]
Body Fluids [A12.207]
Ascitic Fluid [A12.207.119]
Blood [A12.207.152]
Fetal Blood [A12.207.152.200]
Inorganic Chemicals [D01]
Elements [D01.268]
Chalcogens [D01.268.185]
Oxygen [D01.268.185.550]
MeSH
Nervous System [A08]
Central Nervous System [A08.186]
Brain [A08.186.211]
Meninges [A08.186.566]
Spinal Cord [A08.186.854]
Cervical Cord [A08.186.854.126]
Extrapyramidal Tracts [A08.186.
Magnetic Resonance Imaging
Blood
Oxygen
Spinal Cord
The primary form of fMRI uses the blood-oxygen-level
(BOLD) contrast, discovered by Seiji Ogawa. This is a

Figure 1: Text fragment with identified MeSH terms.

In its current standard configuration, the Elsevier Fingerprint Engine includes a set of thesauri or controlled vocabularies and corresponding domains that is listed in Appendix 1.

2. A look inside the Elsevier Fingerprint Engine

2.1 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 or according to their occurrence in automatically detected subsections of a text's body.

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.

2.2 Natural Language Processing components of the EFE

In order to capture all possible term instance (maximizing recall) while keeping erroneous annotations minimal (maximizing precision), the Fingerprint Engine relies on a processing pipeline of natural language processing components. The most important components are the following:

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”.

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.

3. Getting started with the Fingerprint Engine’s RESTful web API

3.1 General

Upon obtaining a subscription or test account from Elsevier, authentication for any call requires credentials to be provided using BASIC AUTHENTICATION. The API functionality can be subdivided in *thesaurus based indexing*, *thesaurus querying*, and *classification*. The payload provided with POST calls generally cannot exceed 25MB of text data.

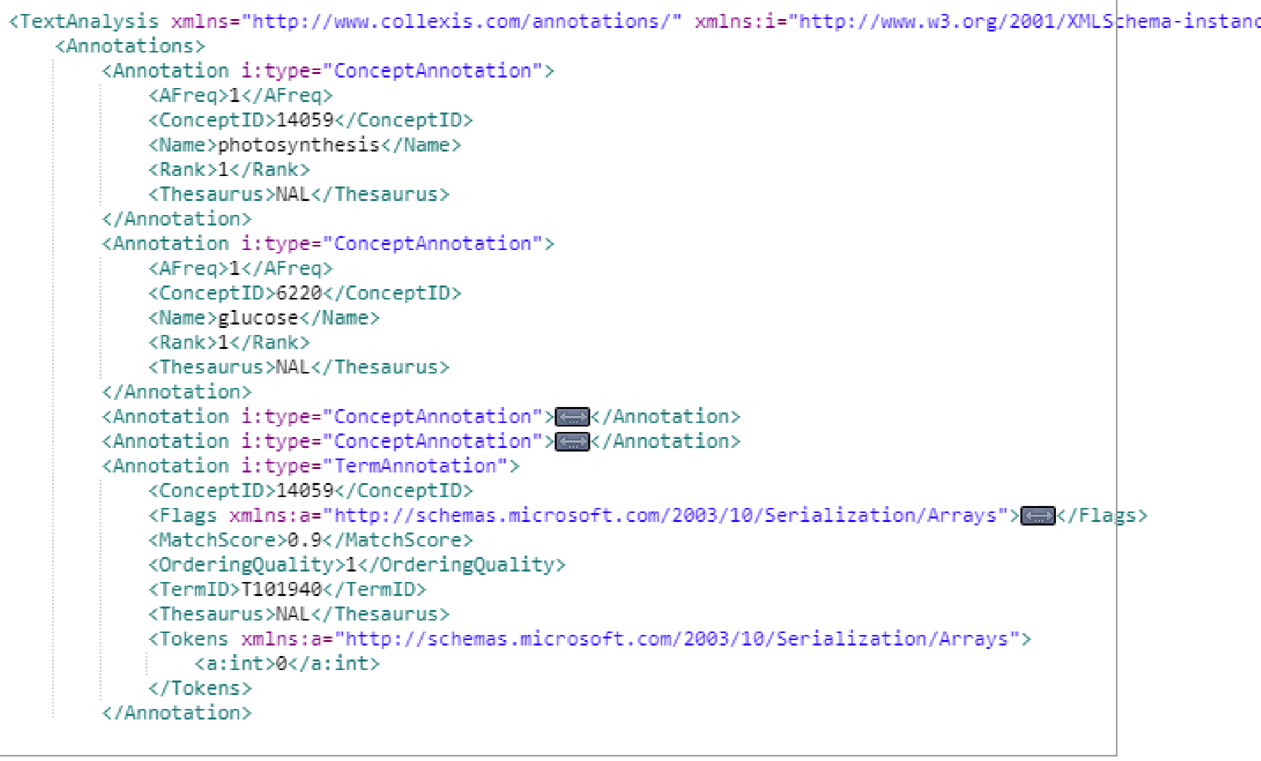
3.2 Thesaurus based indexing

Each thesaurus has its own endpoint, and can be called in the following way:

URI	https://fingerprintengine.scivalcontent.com/TACO7600/TacoService.svc/{thesaurus} where {thesaurus} are codes corresponding to the currently supported thesauri: `CpxTree, Chemistry, Geobase, Math, MathVoc, MeSH, NAL, NASA, Gesis, Humanities, Economics`
Method	POST
Payload	Plain text
Response	XML containing annotations of the types: Concept, Term, POSTag, Word, Token

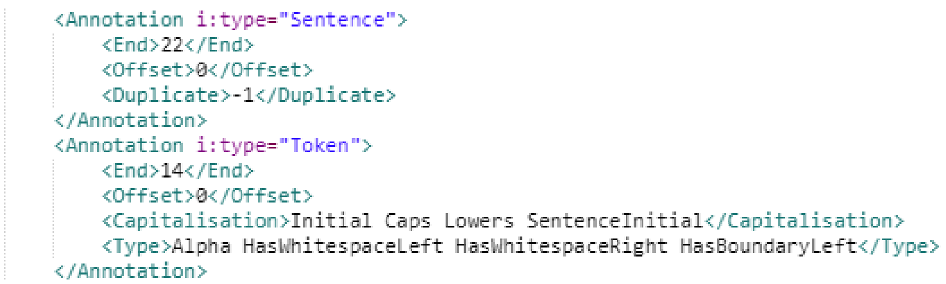
Example (see `FPE_Example1_NAL-Photosynthesis.xml`)

<table>
<tr><td>URI</td><td>https://fingerprintengine.scivalcontent.com/TACO7600/TacoService.svc/NAL</td></tr>
<tr><td>Method</td><td>POST</td></tr>
<tr><td>Payload</td><td>Photosynthesis involves the production of glucose from carbon dioxide and water.</td></tr>
<tr><td>Response</td><td>Starts with Concept and Term annotations:



```
<TextAnalysis xmlns="http://www.collexis.com/annotations/" xmlns:i="http://www.w3.org/2001/XMLSchema-instanc
    <Annotations>
        <Annotation i:type="ConceptAnnotation">
            <AFreq>1</AFreq>
            <ConceptID>14059</ConceptID>
            <Name>photosynthesis</Name>
            <Rank>1</Rank>
            <Thesaurus>NAL</Thesaurus>
        </Annotation>
        <Annotation i:type="ConceptAnnotation">
            <AFreq>1</AFreq>
            <ConceptID>6220</ConceptID>
            <Name>glucose</Name>
            <Rank>1</Rank>
            <Thesaurus>NAL</Thesaurus>
        </Annotation>
        <Annotation i:type="ConceptAnnotation">...</Annotation>
        <Annotation i:type="ConceptAnnotation">...</Annotation>
        <Annotation i:type="TermAnnotation">
            <ConceptID>14059</ConceptID>
            <Flags xmlns:a="http://schemas.microsoft.com/2003/10/Serialization/Arrays">...</Flags>
            <MatchScore>0.9</MatchScore>
            <OrderingQuality>1</OrderingQuality>
            <TermID>T101940</TermID>
            <Thesaurus>NAL</Thesaurus>
            <Tokens xmlns:a="http://schemas.microsoft.com/2003/10/Serialization/Arrays">
                <a:int>0</a:int>
            </Tokens>
        </Annotation>
```

</td></tr>
</table>

	Includes other annotation types, for instance that indicate text character offsets: 

```
<Annotation i:type="Sentence">
    <End>22</End>
    <Offset>0</Offset>
    <Duplicate>-1</Duplicate>
</Annotation>
<Annotation i:type="Token">
    <End>14</End>
    <Offset>0</Offset>
    <Capitalisation>Initial Caps Lowers SentenceInitial</Capitalisation>
    <Type>Alpha HasWhitespaceLeft HasWhitespaceRight HasBoundaryLeft</Type>
</Annotation>
```

Most important fields:

1. `ConceptAnnotation/Afreq` : absolute frequency (count) of concept in the text
2. `ConceptAnnotation/ConceptID` : ID of concept in original source thesaurus/vocab
3. `ConceptAnnotation/Rank` : Rank of concept in case different section types are involved (title, abstract, etc.)
4. `ConceptAnnotation/Thesaurus` : thesaurus that was used for indexing
5. `TermAnnotation/ConceptID` : Concept the term belongs to
6. `TermAnnotation/TermID`: ID of term, as in the original thesaurus or controlled vocabulary if present, otherwise generated
7. `TermAnnotation/Thesaurus` : thesaurus that was used for indexing
8. `TermAnnotation/Tokens` : array element(s) of token annotation. 0 refers to first token, 1 refers to second token, n-1 refers to nth token
9. `Token/Offset` : between character position of token start (0 refers to position before 1st character)
10. `Token/End` : between character position of token end

Bulk indexing:

All endpoints listed so far expect the payload to be plain text to be indexed. It is possible to have multiple documents indexed in a single call. The maximum payload size to send in this way is 25MB, even though a smaller payload may actually result in better throughput per unit of time.

URI	**`https://fingerprintengine.scivalcontent.com/Taco7600/Bulk.svc/{thesaurus}Xml/Document/ID`** where {thesaurus} are codes corresponding to specific thesauri: `CpxTree, Chemistry, Geobase, Math, MathVoc, MeSH, NAL, NASA, Gesis`
Method	POST
Payload	XML
Response	XML

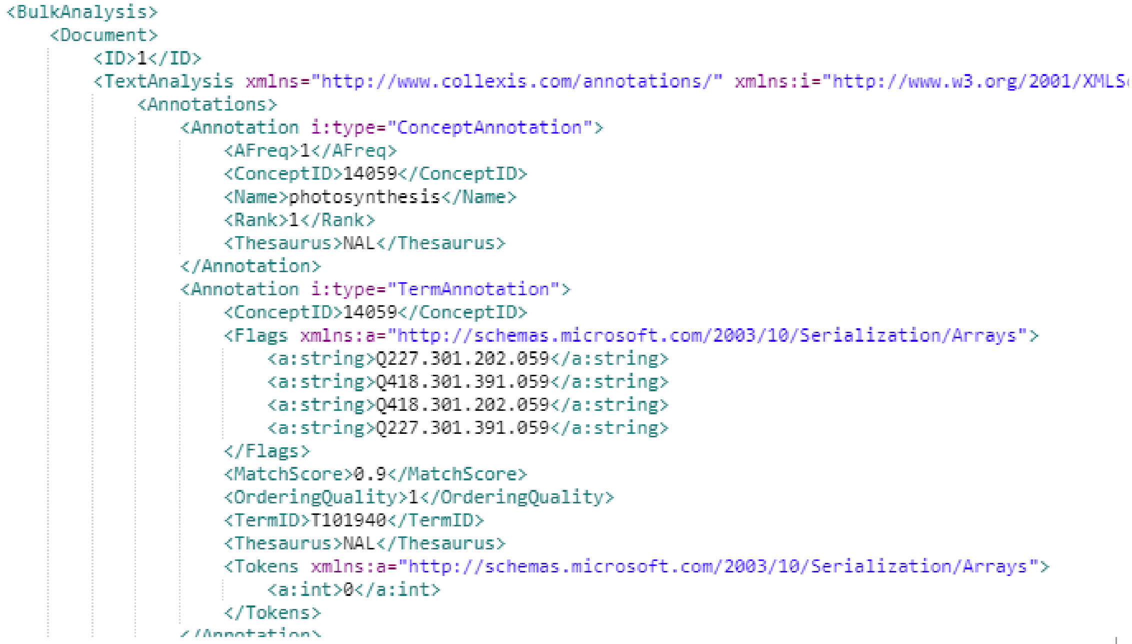
The texts to be indexed are wrapped with <Document> tags which go with an ID attribute:

```
<BulkAnalysis>
      <Document ID="1">…</Document>
      <Document ID="2">…</Document>
</BulkAnalysis>
```

Example (see `FPE_Example1_NAL-Bulk.xml`**)**

URI	https://fingerprintengine.scivalcontent.com/Taco7600/Bulk.svc/NALXml/Document/ID
Method	POST

Payload	`<BulkAnalysis>`
	`        <Document ID="1">Photosynthesis</Document>`
	`        <Document ID="2">cells</Document>`
	`</BulkAnalysis>`

Response	

```
<BulkAnalysis>
    <Document>
        <ID>1</ID>
        <TextAnalysis xmlns="http://www.collexis.com/annotations/" xmlns:i="http://www.w3.org/2001/XMLS
            <Annotations>
                <Annotation i:type="ConceptAnnotation">
                    <AFreq>1</AFreq>
                    <ConceptID>14059</ConceptID>
                    <Name>photosynthesis</Name>
                    <Rank>1</Rank>
                    <Thesaurus>NAL</Thesaurus>
                </Annotation>
                <Annotation i:type="TermAnnotation">
                    <ConceptID>14059</ConceptID>
                    <Flags xmlns:a="http://schemas.microsoft.com/2003/10/Serialization/Arrays">
                        <a:string>Q227.301.202.059</a:string>
                        <a:string>Q418.301.391.059</a:string>
                        <a:string>Q418.301.202.059</a:string>
                        <a:string>Q227.301.391.059</a:string>
                    </Flags>
                    <MatchScore>0.9</MatchScore>
                    <OrderingQuality>1</OrderingQuality>
                    <TermID>T101940</TermID>
                    <Thesaurus>NAL</Thesaurus>
                    <Tokens xmlns:a="http://schemas.microsoft.com/2003/10/Serialization/Arrays">
                        <a:int>0</a:int>
                    </Tokens>
                </Annotation>
```

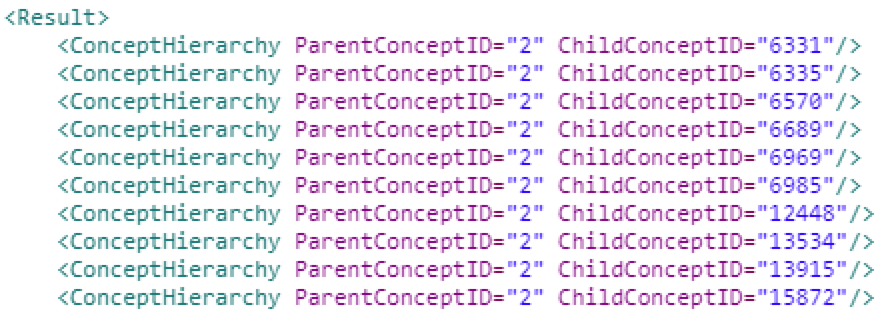
3.3 Thesaurus structure and elements

Vocabularies generally do not contain hierarchical structure (broader/narrower relationships), but most thesauri do. This structure can be obtained with the following API call:

URI	https://fingerprintengine.scivalcontent.com/TACO7600/Query.svc/Hierarchy/{thesaurus} where {thesaurus} are codes corresponding to specific thesauri: `CpxTree, Chemistry, Geobase, Math, MathVoc, MeSH, NAL, NASA, Gesis`
Method	POST
Payload	<empty>
Response	XML containing hierarchical stucture

Example (see `FPE_Example2_NAL-HierarchyFragment.xml`**)**

URI	https://fingerprintengine.scivalcontent.com/taco7600/Query.svc/Hierarchy/NAL
Method	POST
Payload	<empty>

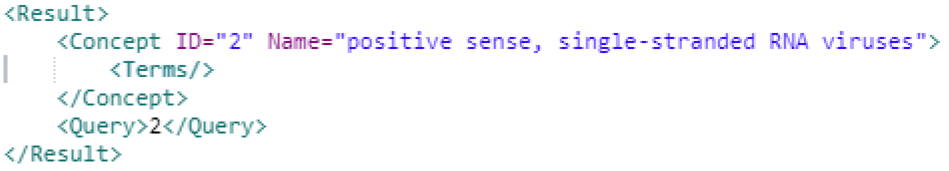
Response	ConceptHierarchy elements with Parent – Child relations by ConceptIDs:
	
	<Result>
	    <ConceptHierarchy ParentConceptID="2" ChildConceptID="6331"/>
	    <ConceptHierarchy ParentConceptID="2" ChildConceptID="6335"/>
	    <ConceptHierarchy ParentConceptID="2" ChildConceptID="6570"/>
	    <ConceptHierarchy ParentConceptID="2" ChildConceptID="6689"/>
	    <ConceptHierarchy ParentConceptID="2" ChildConceptID="6969"/>
	    <ConceptHierarchy ParentConceptID="2" ChildConceptID="6985"/>
	    <ConceptHierarchy ParentConceptID="2" ChildConceptID="12448"/>
	    <ConceptHierarchy ParentConceptID="2" ChildConceptID="13534"/>
	    <ConceptHierarchy ParentConceptID="2" ChildConceptID="13915"/>
	    <ConceptHierarchy ParentConceptID="2" ChildConceptID="15872"/>

Information on the concepts involved can then be retrieved with the following call:

URI	`https://fingerprintengine.scivalcontent.com/TACO7600/Query.svc/Concepts/{thesaurus}` where {thesaurus} are codes corresponding to the currently supported thesauri: `CpxTree, Chemistry, Geobase, Math, MathVoc, MeSH, NAL, NASA, Gesis, Humanities, Economics`
Method	POST
Payload	<conceptnumber>
Response	XML containing concept name and possible terms

Example (see `FPE_Example2_NAL-Concept2.xml`**)**

URI	`https://fingerprintengine.scivalcontent.com/TACO7600/Query.svc/Concepts/NAL`
Method	POST
Payload	2

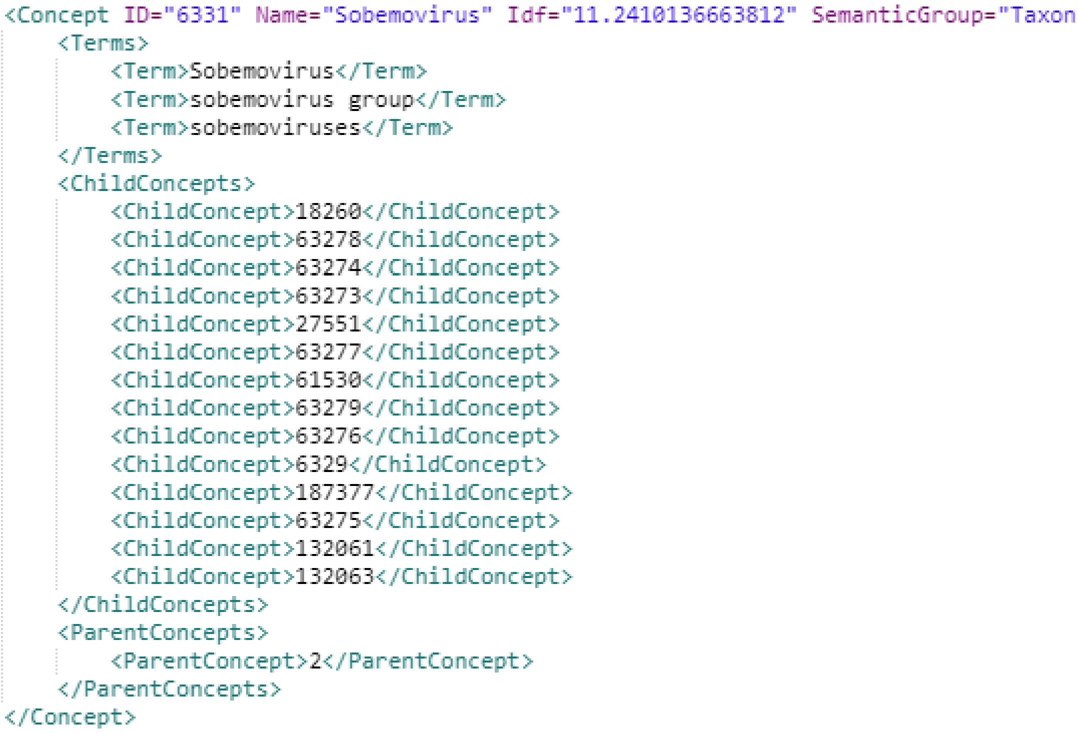
Response	
	<Result>
	    <Concept ID="2" Name="positive sense, single-stranded RNA viruses">
	        <Terms/>
	    </Concept>
	    <Query>2</Query>
	</Result>

Finally, the full concept list can be retrieved with the following query *(warning: the response can be large depending on thesaurus size)*

URI	`https://fingerprintengine.scivalcontent.com/TACO7600/Query.svc/ConceptList/{thesaurus}` where {thesaurus} are codes corresponding to the currently supported thesauri: `CpxTree, Chemistry, Geobase, Math, MathVoc, MeSH, NAL, NASA, Gesis, Humanities, Economics`
Method	POST
Payload	<empty>
Response	XML containing all concepts in the thesaurus and their child concepts

Example (see `FPE_Example2_NAL-ConceptListFragment.xml`**)**

URI	`https://fingerprintengine.scivalcontent.com/TACO7600/Query.svc/ConceptList/NAL`
Method	POST
Payload	<empty>

Response	

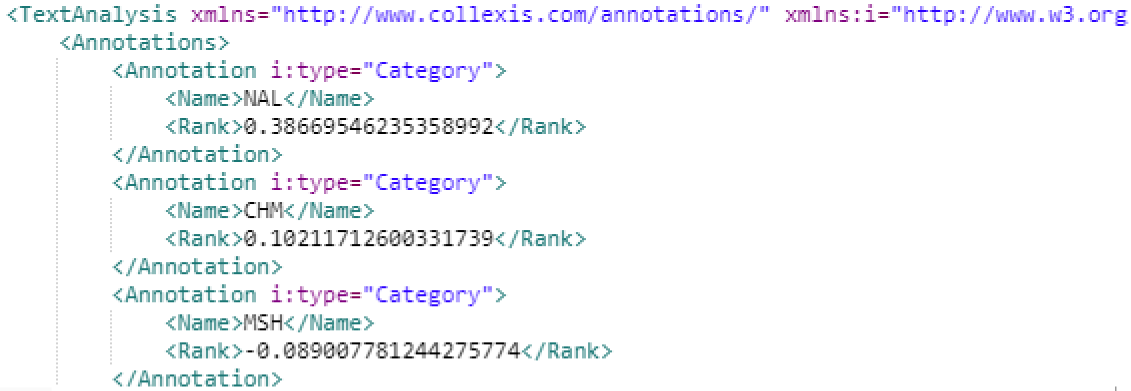
```
<Concept ID="6331" Name="Sobemovirus" Idf="11.2410136663812" SemanticGroup="Taxon
    <Terms>
        <Term>Sobemovirus</Term>
        <Term>sobemovirus group</Term>
        <Term>sobemoviruses</Term>
    </Terms>
    <ChildConcepts>
        <ChildConcept>18260</ChildConcept>
        <ChildConcept>63278</ChildConcept>
        <ChildConcept>63274</ChildConcept>
        <ChildConcept>63273</ChildConcept>
        <ChildConcept>27551</ChildConcept>
        <ChildConcept>63277</ChildConcept>
        <ChildConcept>61530</ChildConcept>
        <ChildConcept>63279</ChildConcept>
        <ChildConcept>63276</ChildConcept>
        <ChildConcept>6329</ChildConcept>
        <ChildConcept>187377</ChildConcept>
        <ChildConcept>63275</ChildConcept>
        <ChildConcept>132061</ChildConcept>
        <ChildConcept>132063</ChildConcept>
    </ChildConcepts>
    <ParentConcepts>
        <ParentConcept>2</ParentConcept>
    </ParentConcepts>
</Concept>
```

3.4 Domain classification

The Fingerprint Engine comes with a text classifier that ingests plain text, classifies it into the domains / thesauri supported, and returns a ranking that reveals the most relevant domain(s) for the text input.

URI	https://fingerprintengine.scivalcontent.com/Taco7600/TacoService.svc/DomainClassifier
Method	POST
Payload	<plaintext>
Response	XML containing domain ranking

Example (see `FPE_Example2_DomainClassifier1.xml`**)**

URI	https://fingerprintengine.scivalcontent.com/Taco7600/TacoService.svc/DomainClassifier
Method	POST
Payload	Photosynthesis involves the production of glucose from carbon dioxide and water.
Response	

```
<TextAnalysis xmlns="http://www.collexis.com/annotations/" xmlns:i="http://www.w3.org
    <Annotations>
        <Annotation i:type="Category">
            <Name>NAL</Name>
            <Rank>0.38669546235358992</Rank>
        </Annotation>
        <Annotation i:type="Category">
            <Name>CHM</Name>
            <Rank>0.10211712600331739</Rank>
        </Annotation>
        <Annotation i:type="Category">
            <Name>MSH</Name>
            <Rank>-0.089007781244275774</Rank>
        </Annotation>
```

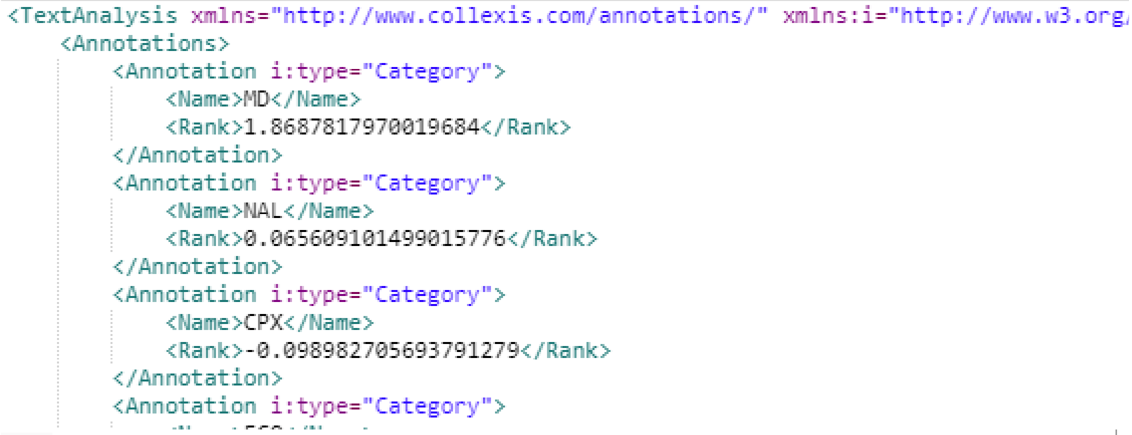
Etc.

The domain classifier relies on Support Vector Machine based classification, and ranks all available categories according to descending goodness-of-fit: the first assigned category is the best fit, the second assigned category the one but best, et cetera. Rank scores function as an ordering measure and are not intended as probabilities, but thresholding can still be useful to isolate those classes that stand out: in the example discussed, Agriculture (NAL) and Chemistry (CHM) are the top ranked classes with a rank score that is substantially higher than those of the classes that follow.

Finally, the domain classifier puts an optional class MD at the top of the ranking if it considers text to be Multi-Disciplinary, as illustrated in the following example:

Example (see `FPE_Example2_DomainClassifier2.xml`)

URI	https://fingerprintengine.scivalcontent.com/Taco7600/TacoService.svc/DomainClassifier
Method	POST
Payload	World Crude Oil Production is at a current level of 81.90M, down from 82.23M last month and up from 81.44M one year ago.

Response	
	`<TextAnalysis xmlns="http://www.collexis.com/annotations/" xmlns:i="http://www.w3.org`
	`    <Annotations>`
	`        <Annotation i:type="Category">`
	`            <Name>MD</Name>`
	`            <Rank>1.8687817970019684</Rank>`
	`        </Annotation>`
	`        <Annotation i:type="Category">`
	`            <Name>NAL</Name>`
	`            <Rank>0.065609101499015776</Rank>`
	`        </Annotation>`
	`        <Annotation i:type="Category">`
	`            <Name>CPX</Name>`
	`            <Rank>-0.098982705693791279</Rank>`
	`        </Annotation>`
	`        <Annotation i:type="Category">`
	Etc.

Appendix 1: Supported thesauri and controlled vocabularies.

Name	Domain	Short description	URL	Elements
Chemistry	Chemistry	Composed of MeSH and Elsevier's Compendex chemistry concepts	See MeSH/Compendex	See MeSH/Compendex
Compendex	Engineering	Elsevier's Compendex (COMPuterized ENgineering inDEX) thesaurus for Engineering & Materials Science, Chemistry	https://www.elsevier.com/solutions/engineering-village/content/compendex	Concepts, Terms, Hierarchy
Economics	Economics	Elsevier's FPE-generated vocabulary	None	Terms
Geobase	Earth Sciences	Elsevier's Earth Sciences thesaurus + GEMET - GEneral Multilingual Environmental Thesaurus	None	Concepts, Terms, Hierarchy
Gesis	Social Sciences	Gesis Thesaurus for the Social Sciences (licenced)	http://www.gesis.org/en/services/research/tools/social-science-thesaurus/	Concept, Terms, Hierarchy, Synonyms, Definitions
Humanities	Humanities	Elsevier's FPE-generated vocabulary	none	Terms
Math	Mathematics	University of Cambridge Connecting Mathematics Thesaurus + Elsevier's FPE-generated vocabulary	none	Terms
MeSH	Medical Sciences	National Library of Medicine's MeSH Medical Subject Headings incl. MeSH Supplemental terms	https://www.nlm.nih.gov/mesh/	Concept, Terms, Hierarchy, Synonyms, Definitions
NAL	Agronomics	NALT National Agricultural Library Agricultural Thesaurus	https://agclass.nal.usda.gov/	Concept, Terms, Hierarchy, Synonyms, Definitions
NASA	Physics	National Aeronautics and Space Administration (NASA) Thesaurus	https://www.sti.nasa.gov/sti-tools/	Concept, Terms, Hierarchy, Synonyms, Definitions