Response to NOT-CA-15-019, Request for Information (RFI): Input on National Cancer Institute Metadata Repository and Services

Elsevier values its multi-faceted and synergistic relationship with the NIH and is appreciative for the opportunity to provide a response to NOT-CA-15-019, Request for Information (RFI): Input on National Cancer Institute Metadata Repository and Services. Submitted on behalf of Elsevier by:

Holly J Falk-Krzesinski, PhD
Vice President, Global Academic & Research Relations
h.falk-krzesinski@elsevier.com
Elsevier
New York, NY, USA

Elsevier has a long track record of data and metadata standards, dating back to the 1990s when we led the TULIP project. The Elsevier XML specifications for journal articles and book chapters are widely known and in use for 3000+ propriety and society journals and the metadata for 20,000+ journals. Content, including 12M journal articles, resides in a content repository that is accessible through restful APIs. Its metadata model is described using RDF serialized as JSON-LD. The API payloads and responses in JSON-LD are treated in the same way as our main content standards.

Our content is stored in multiple content-type-specific "warehouses." Through a metadata repository, this is made in to a virtual whole, called our Virtual Total Warehouse. Our content model and metadata standards are especially focused on content versioning. "Generations" of content assets keep various files together that together constitute a version. This Virtual Total Warehouse (VTW) plays a role in acquisition, editing and curating content (in our case, journal articles, book chapters, drug monographs, patents, patient education, and much more) and a Content Enrichment Framework takes this content and can, in principle, run any semantic process on the content, depositing the results back in VTW.

Elsevier also has a linked data repository adhering to the standards of linked data and linked open data.

Elsevier's approach to unstructured information: The vast majority of information exists as an unstructured text which makes it unsuitable for efficient analysis by humans. The area of computational assistance to analysis of large volumes of textual information is traditionally split into two (somewhat overlapping) approaches - information retrieval and information extraction.

Information Retrieval (IR) systems concentrate on finding documents containing information deemed relevant to a particular topic of interest. Usually this is done by analyzing the word content of the documents using statistical methods based on keywords or word co-occurrence. IR methods are by their nature generic and to a large degree language-independent; the output of IR systems is *intended for human readers*.

Unlike IR, Information Extraction (IE) focuses on extracting information contained within the documents in a form *suitable for automatic processing*. IE systems use an *ontology* (or knowledge representation schema) as a model of a particular domain, and thus are domain-specific. The simplest form of an ontology is a list (or, even better, a hierarchical tree) of concepts relevant to the domain. More advanced forms of ontology also specify possible semantic types of relationships between the concepts. Extracting information with high precision involves deep understanding of the actual meaning of the text; as a result, IE systems are language-specific.

In developing solutions for vertical markets, Elsevier takes the IE road. Instead of building one generic, language- and domain-independent system that deals with large number of topics but provides little

depth when it comes to the subject matter, we focus on extracting structured information specific for a particular domain from English text.

I. Elsevier's Information Extraction (IE) technology: MedScan

Elsevier has developed a proprietary natural language processing (NLP)-based technology called MedScan for extraction of structured information from unstructured text, which is suitable for use in vertical markets. It is a good fit for automatic indexing of NCI's content as the MedScan Thesaurus/Taxonomy was built mostly based on NCI thesauri and has all the NIH identifiers integrated (MeSH Headings, NCI Metathesaurus IDs, Entrez Gene IDs, Organism Tax IDs, etc.). The technology works by first recognizing domain-specific named entities (concepts) in the input text, and then uses natural language processing techniques to extract *attributed, directional semantic relationships* between them. The relationships can be of any complexity from simplest binary (X affects Y) to n-ary (X protects Y from Z) and complex multi-level nested ones (effect of X on Y depends on Z).

Elsevier IE technology has modular architecture. Each module performs its specific function and has well-defined and documented input/output format. Modules with compatible interfaces can be combined into different text processing pipelines, as required by the application. All modules are written from scratch to achieve our flexibility/precision/performance goals. The modules are portable C/C++ applications interacting via files and pipes.

MedScan Technical Description: MedScan is a proprietary natural language processing (NLP) -based technology for extraction of structured information from unstructured text. Structured information is captured and formally represented using a conceptual model (ontology) of the domain. The ontology consists of a set of conceptual named entities (e.g. Proteins, Small molecules, Cellular processes, Diseases, etc) and a set of categorized relationships (Binding, Protein Modification, Expression regulation, Molecular Transport, etc) between them.

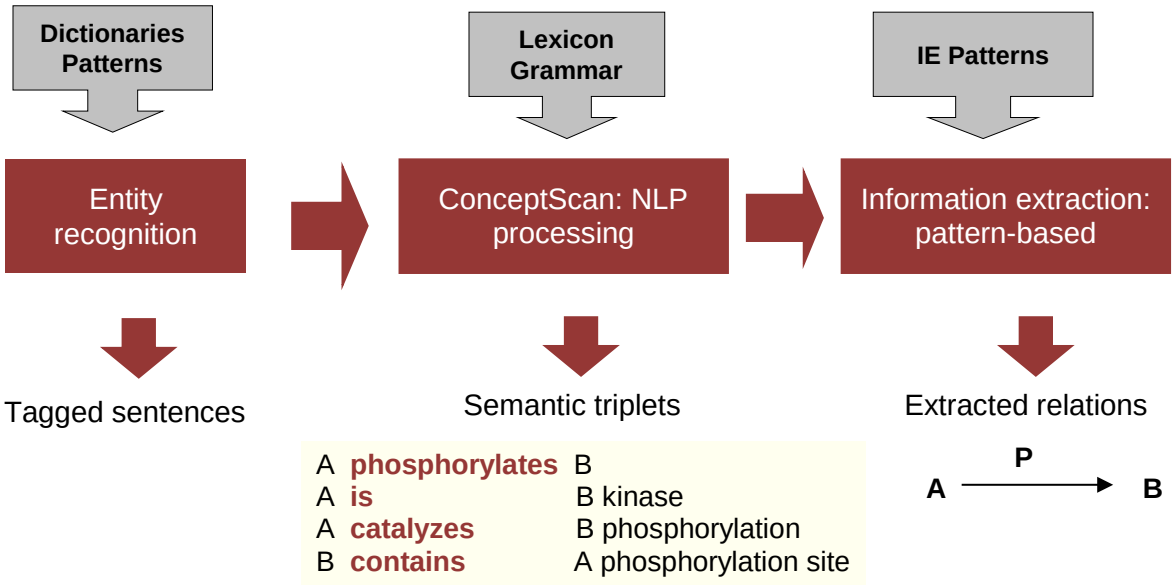


Figure 1. An overview of MedScan Architecture

MedScan first recognizes different domain-specific named entities (gene/protein names, cellular processes, cellular components, diseases, tissues, organs, etc.) in the input text, and then extracts functional relations (binding, regulation, association, molecular transport, etc.) between them. Figure 1 shows an overview of MedScan architecture.

The Entity Recognizer module utilizes hand-crafted dictionaries of domain-specific entities in combination with an advanced matching algorithm to detect them in input text.

To extract entity relationships from the text, MedScan utilizes two modules. The natural language processing module, ConceptScan, analyzes the sentence structure and decomposes each sentence into a deterministic set of Subject-Verb-Object triplets, each representing a single semantic relationship between two singular noun phrases. Next, Pattern Matcher matches carefully designed linguistic patterns over the triplets to extract and encode the entity relationships.

MedScan has been field-tested and is proven to be fast, efficient, and accurate information extraction technology. It is currently used to process the content of the entire Medline database along with more than 40 freely available full-text journals in order to extract more than 3.5 million individual facts (relations) about functions of proteins with an overall accuracy of 90% and recall of 70%. The entire processing cycle can be completed in less than 24 hours on a regular PC.

Dictionaries and Named Entity Recognition:

Table 1. MedScan dictionaries

Entity type	Number	Main sources
Proteins	136,000	Entrez Gene
Prot. Classes	7,500	GO, Enzymes, PubMed
Cell components	740	GO, PubMed
Cell processes	5,200	GO, PubMed
Diseases	6,300	MESH, PubMed
Small Molecules	270,000	MESH, PubChem, PubMed
Tissues	100	MESH, UMLS, NCI, EVoc
Cell types	360	MESH, UMLS, NCI, EVoc
Organs	2,875	MESH, UMLS, NCI, EVoc
Clinical parameters	1,786	Pubmed, ClinicalTrials.gov
Cell lines	2,500	PubMed

The Entity recognition module of MedScan utilizes hand-curated dictionaries of biomedical entities to detect them in the input text. Dictionaries are manually compiled and curated from the number of various public-domain resources (EntrezGene and SwissProt for protein names, PubChem and MESH for small molecules, GO for cell processes and components, MESH for diseases, NCI thesaurus for organs, tissues and cells, etc). Whenever possible the entities are hyperlinked to those outside resources for reference. Many additional aliases and terms are also added directly from the literature resources,

e.g. PubMed. Table 1 shows the content of MedScan dictionaries. MedScan uses number of different algorithms to achieve accurate detection of entities in text. It can also use rule- and regular expression-based approaches to detect specific types of entities (abbreviations, numbers, dates, etc). The dictionaries are in a simple tab-delimited format so they can be easily extended or modified.
The input text can be in various formats (plain text, Microsoft Office, HTML, reasonable forms of PDF, zip/tar/gzip archives of the above, etc.) The output of the entity recognition step consists of individual sentences labeled to preserve their origin with identified named entities marked up with entity IDs, using **ID{***number***=….}** format (shown in red):

15986412:5 Enzyme assay, Western blot and ID{4000000,4106278=reverse-transcription} polymerase chain reaction (RT-PCR) results demonstrated that protein and mRNA expressions of human simple **ID{445329=phenol sulfotransferase}** (**ID{6799=P-PST}**), human ID{6818=monoamine sulfotransferase} (ID{6818=M-PST}), human ID{6822=dehydroepiandrosterone sulfotransferase} (ID{6822=DHEA-ST}) and human ID{6783=estrogen sulfotransferase} (ID{6783=EST}) were induced in ID{10000000,11012376=Hep G2 cells}; ID{6818=M-PST} and ID{6822=DHEA-ST} were induced in ID{10000000,11010382=Caco-2 cells}.

The type of entity is encoded in its numerical range.

Natural Language Processing:

The central idea of Elsevier's NLP algorithm (called ConceptScan) is decomposing natural language sentences into semantic relationships (which we will also call semantic triplets). Each triplet is designed to represent a single semantic relationship between two singular noun phrases (NPs). An example below illustrates this paradigm using a complex artificially constructed sentence.

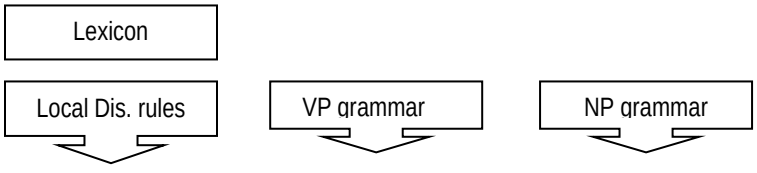
11940574:7 Because **Axin2** has been shown to associate with and inhibit **beta-catenin** abundance and function, we hypothesized that **Axin2**, which is affecting proliferation of MEF cells can work in a negative feedback pathway, regulating **Wnt** signaling and thus controlling apoptotic process.

Triplets:

Axin2	associate	beta-catenin abundance
Axin2	inhibit	beta-catenin function
Axin2	associate	beta-catenin abundance
Axin2	inhibit	beta-catenin function
Axin2	affect	MEF cell line proliferation
Axin2	work	negative feedback pathway
Axin2	regulate	Wnt signaling
Axin2	control	apoptotic process

The extracted triplets capture the main facts expressed in a sentence. The ConceptScan is used in conjunction with named entity detection algorithm to index relationships between biomedical entities and to extract entity relationships.

ConceptScan parses sentences in several sequential algorithmic steps (See figure below)



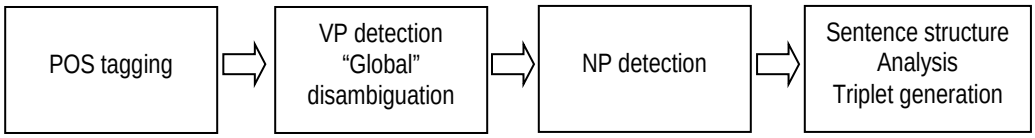


Figure 2. ConceptScan algorithm

The first step of NLP is part-of-speech tagging and local disambiguation. During this step, the words in a sentence are reduced to all possible uninflected forms, looked up in the lexicon and annotated with the respective syntactic categories.

After initial POS tagging, the local disambiguation algorithm, encoded by a set of contextual regular expression-like rules, is applied. Notably, not all ambiguities can be resolved locally. The unresolved ambiguities are preserved for subsequent processing steps.

The next step is identification of verbal phrases. Verbal phrase (VP) grammar is encoded in a single but complex deterministic finite-state automaton (DFA), with more than 25,000 states. It is matched over the sequence of syntactic categories assigned to sentence words at the POS-tagging step.

NP grammar is matched after detection of verbal phrases is complete. Similarly to VP grammar, it is encoded by a DFA. The structure of NP grammar covers prepositional attachment, conjunctions, relational constructs, appositions and exemplifications.

Once VPs and NPs have been identified, ConceptScan analyzes the structure of the entire sentence.

Information extraction:
The specific relationships between entities are extracted using separate module - Pattern Matcher. It utilizes a formalism closely resembling regular expressions to detect specific linguistic constructs expressing entity relations and to capture the expressed relations. It is specifically tailored to deal with linguistic input; it operates on the level of individual words rather than symbols and supports advanced

```
CONTROL
{
  ControlType = "ProtModification"
  in  = %Protein1(Protein)
  out = %Protein2(Protein)
}
:
%Protein1 $MODAL? $ADV* phosphorylate~ %Protein2        |
%Protein2 $MODAL? $BE $ADV* phosphorylated by %Protein1 |
Phosphorylation of %Protein2 by %Protein1               |
;
```

Figure 3. An example of the information extraction pattern. The head template encodes the name of the output frame and templates for the values of its slots, which can be literals or other frames. Named entity variables (%Protein1 and %Protein2) are distinguished by the leading '%'. The head template can restrict the named entity variables to take values of specific semantic type(s) by providing the list of types in parentheses. Named word sets are distinguished by the leading '$'. They can be defined anywhere in the pattern file and can be used in multiple patterns. In the above example $MODAL is the set of modal verbs (can, may, might, etc). The '~' postfix indicates that the preceding word can be matched in any grammatical form. Multiple patterns extracting identical information are separated by the '|' separator.

linguistic features like matching all word forms and multi-word lexemes. Pattern matching also supports all regular expression features: wildcards, sets, negation, etc. The figure below shows a sample information extraction pattern.

MedScan output: The output of MedScan is in an XML-based format describing entities and relation between them (see an example below):

```
<resnet mref="16377759:4" msrc="The catalytic domain of ID{820019=S6K1} could be
phosphorylated by Arabidopsis ID{841259=3-phosphoinositide-dependent protein kinase-1}
(ID{830330=PDK1}), indicating the involvement of ID{830330=PDK1} in the regulation of
ID{820019=S6K}.">
  <nodes>
   <node local_id="N1" urn="urn:agi-llid:841259">
    <attr name="NodeType" value="Protein" />
    <attr name="Name" value="at1g48390" />
   </node>
   <node local_id="N2" urn="urn:agi-llid:820019">
    <attr name="NodeType" value="Protein" />
    <attr name="Name" value="AT3G08720" />
   </node>
  </nodes>
  <controls>
   <control local_id="L1">
    <link type="in" ref="N1" />
    <link type="out" ref="N2" />
    <attr name="ControlType" value="ProtModification" />
    <attr name="ModificationType" value="phosphorylation" />
   </control>
  </controls>
 </resnet>
```

Figure 4. An example of a MedScan output.

MedScan Ontology of Relationships: Elsevier has developed ontology of different types of relations between biological entities. Each type of relation has a very specific semantic definition and is typically attributed with additional information, e.g. sign of relations (e.g. positive, negative or unknown) or mechanism (e.g. phosphorylation, methylation, etc). There are three set of patters currently used by MedScan to extract biological relations – patterns focused on extraction of different aspects of protein functions, small molecule functions and disease biomarkers. The Table 2 below shows the scope of biological relationships currently extracted by MedScan.

The current scope of the information extracted by MedScan can be extended by developing new dictionaries covering other aspects of biomedical domain (e.g. focused more on medical or clinical entities) and/or by developing novel information extraction patterns to capture other types of entity relationships.

The Pattern Matcher is extremely fast: it runs through more than 16,000,000 entity-tagged sentences from the entirety of Medline in less than 20 minutes.

- Protein -> Protein
  - Binding
  - Protein modification
  - Expression (positive/negative/unknown)
  - Promoter regulation/Binding
  - Regulation (positive/negative/unknown)
- Protein -> Small Molecules
  - Synthesis/Degradation
  - Mol. Transport
- Protein -> Cell processes
- Protein -> Disease
  - Positive/negative regulation
- Disease -> Protein/Small molecules
  - Changed concentration/expression (positive/negative/unknown)
  - Mutations
  - Activity (positive/negative/unknown)
- Small molecules -> Protein
  - Binding
  - Direct regulation
  - Expression
  - Indirect regulation (positive/negative)
- Small molecules -> Disease/Cell processes (positive/negative/unknown)

Table 2. Relationships currently extracted by MedScan

MedScan Customizations:
MedScan is flexible platform open for two types of end-user modifications. First, MedScan taxonomy and dictionaries can be extended to include new concepts and even new concept classes. Dictionaries are provided in a simple text-based tabular format and new concepts and concept aliases can be added to the files. Second, the scope of extracted information can be extended to include new relationships by modifying information extraction rules. The rules are recorded in a well-documented textual format and new rules can be created and added to MedScan.

MedScan Features and competitive advantages:
Elsevier's IE engine has been designed and implemented from scratch to address flexibility, precision/recall and performance problems of the off-the-shelf NLP tools. Our design efforts focused on issues specific for texts in vertical application domains characterized by complex sentence and relationship structure, highly specialized entity notation, proliferation of abbreviations and synonyms. As a result of this focus, we have surpassed the 90% precision / 60% coverage mark on technical texts in our current application domains (biology and medicine). Our engine has an unmatched performance – it can process up to 1000 sentences per second on a regular PC, which is 2-3 orders of magnitude faster than prevailing NLP technologies. High performance allowed us to achieve clean separation between modules where traditional approaches intertwine distinct functions like parsing and ontology-based information extraction to cut down on the amount of information exchanged between modules. Also,

much attention has been paid to keep domain-specific information in dictionaries and rule files, to simplify maintenance and extending the coverage to other domains.

The engine achieved production quality in 2003 and since then has been installed on many sites, including both individual and corporate-wide licenses.

II. Elsevier's Information Extraction (IE) technology: Fingerprint Engine

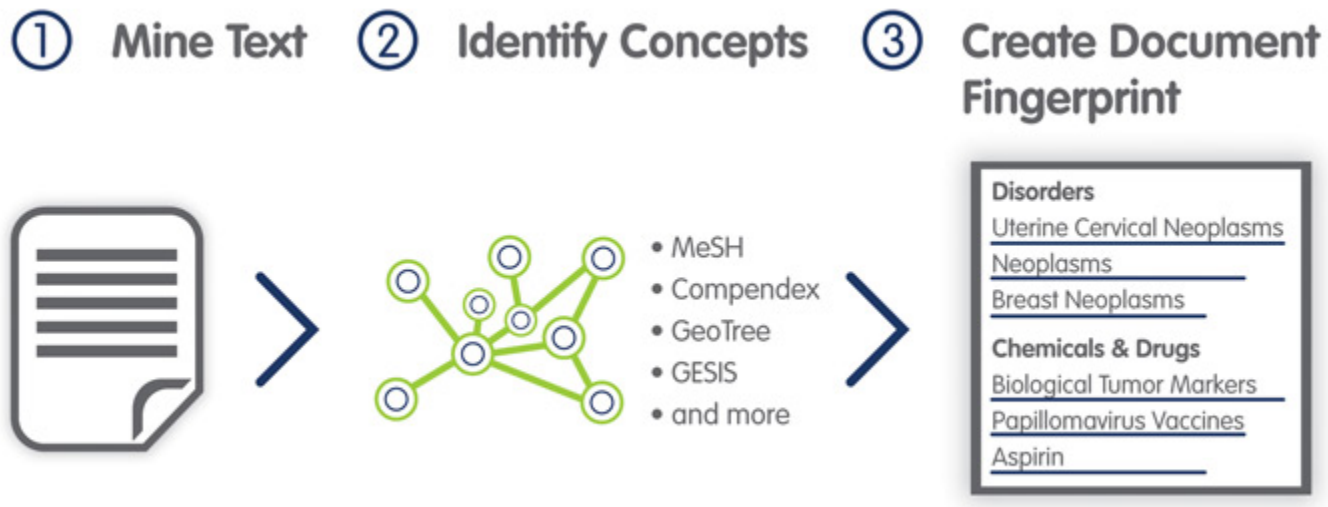
A back-end software system, the Elsevier Fingerprint Engine mines the text of scientific documents – publication abstracts, funding announcements and awards, project summaries, patents, proposals/applications, and other sources – to create an index of weighted terms which defines the text, known as a Fingerprint™ visualization.

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

The Elsevier Fingerprint Engine powers many solutions including Pure, comprehensive information management system, and Reviewer Finder, Elsevier's tool for finding reviewers.

The Elsevier Fingerprint Engine uses a variety of thesauri to support applications pertaining to different subject areas. By applying a wide range of thesauri, Elsevier can develop solutions in but not limited to: the life sciences, engineering, earth and environmental sciences, arts and humanities, social sciences, mathematics and agriculture. Thesauri provided by an institution or specific research domain can also be incorporated.

The Elsevier Fingerprint Engine creates Fingerprints via a three-step process:



1. The Elsevier Fingerprint Engine applies a variety of Natural Language Processing (NLP) techniques to mine the text of scientific documents including publication abstracts, funding announcements and awards, project summaries, patents, proposals, applications and other sources

2. Key concepts that define the text are identified in thesauri spanning all the major disciplines

3. The Elsevier Fingerprint Engine creates an index of weighted terms that defines the text, known as a Fingerprint.

Applying Fingerprints to inform decision making: By aggregating and comparing Fingerprints of people, publications, funding opportunities and ideas, the Elsevier Fingerprint Engine can reveal insightful connections with practical applications. Here are some examples of how Fingerprints are currently used to bring scholarly business intelligence to institutional data.

Pure aggregates the Fingerprints of individual documents to create unique Fingerprints that reveal your researchers' distinctive expertise. Pure also matches the Fingerprints of funding opportunities in SciVal® Funding to researchers' Fingerprints, recommending appropriate funding opportunities and suggested collaborators.

Reviewer Finder compares document Fingerprints with researcher Fingerprints, making it easier to identify reviewers and raise awareness about potential conflicts of interest.

Elsevier Journal Finder helps researchers find journals that could be best suited for publishing their articles. Journal Finder matches abstracts to Elsevier journals, scanning Elsevier's 2,200+ titles in the Health Sciences, Life Sciences, Physical Sciences and Social Sciences.