Linguamatics ontologies power scientific search and text mining

Linguamatics text-mining solutions for data transformation

Linguamatics' natural language processing-based text-mining solution, I2E, offers unparalleled capabilities within the life science and healthcare domains to realize timely insights through semantically-enhanced fact extraction, document annotation, and data transformation. Key components of the solution are the ontologies available for use with the software, which cover a broad range of biomedical and healthcare domains.

Linguamatics terminologies for pharma and healthcare domains

I2E's **biomedical terminologies** enable identification of over a million concepts, covering all the key life science domains: Diseases, genes, proteins, biomarkers, gene variants, mutations, targets, drugs, adverse events, biological processes, organs, tissues and cells, and more.

Biomedical terminology

- Anatomy (organs, tissues, cells)
- Biomarkers
- Diseases, Adverse Events
- Genes, proteins
- Experimental techniques and assays
- Mutations
- Organisms

Healthcare terminologies integrated in I2E cover key medical domains and categories, valuable for identifying key patient data from a variety of medical records. These include: Patient problem lists; history of disease; vitals (blood pressure, heart rate, pulse, respiratory rate, temperature, gender, age); lifestyle factors (smoking, pack years, drug use, alcohol use, exercise, diet, sexual activity); tumor, node, and metastasis (TNM) staging; and more. These are recognised using a combination of standard ontologies, pattern-based approaches, and linguistic rules, to enable the context around any patient variable to be taken into account (e.g. family history vs. no family history of disease).

Healthcare terminology

- Clinical biomarkers
- Clinical care, outcomes management
- Tumor staging (TNM)
- Patient demographics
- Patient history (e.g. smoking status)
- Medical codes

Key **chemical entities** can be found using ChEBI, MeSH, and NCI Thesaurus. In addition, I2E Chemistry, with ChemAxon, identifies known and novel chemical structures within documents, by name, structure, substructure, or similarity. Drug Lab Codes is a Linguamatics pattern ontology that enables the identification and extraction of many different pharmaceutical company chemical identifiers (such as LY-170053, SQ 34676, ICI 204,219).

Chemistry

- Chemistry by structure
- Compounds by name
- Drugs, excipients, active ingredients
- Drug Lab Codes

Information on **organizations and locations** can be identified and extracted by sector (e.g. named pharma companies, universities, government agencies) or by type (pattern matches to organization types, e.g. clinic, corporation, division, hospital, institute).

Organizations and locations

- By sector
- By type
- Geographical location (region, country, state, city, etc.)

In addition to the semantic, dictionary-type entity recognition, I2E provides **pattern ontologies** that identify categories of information, such as times, dates, numerics, mutations, organizations, telephone numbers, or units of measurement. Pattern ontologies are incredibly valuable for identification of concepts that can be expressed in many ways. These pattern ontologies extend search far beyond the ontology-matching approach, to annotate novel descriptions in text of key concepts or concept types.

Pattern ontologies

- Age class
- Date class
- Drug Lab Codes class
- Email class
- Measurements class
- Mutations class
- Numerics class
- Person class
- Telephones class
- Time period class
- Units class

Natural language processing (NLP) adds **linguistic components** to search and markup, enabling the user to focus on, for example, the relations between concepts (e.g. search for compounds causing disease rather than treating disease), or on making search more precise by looking for linguistic relationships, sentence co-occurrence, and using negation (e.g. "pressure" but not "blood pressure," "smoker" but not "smoker and quit 5 years ago," patients "with diabetes" but not "family history of diabetes"). Using linguistic and other wildcards enables open search that can identify concepts in context with the search terms of interest.

Linguistic components

- Relations ontology
- Linguistic units (word, preposition, noun phrase, verb phrase)

Bespoke vocabularies

I2E also provides functionality to use bespoke or custom vocabularies; these can be imported from academic or commercial sources. In-house vocabularies can also be used within I2E, whether these are dictionaries of employees from an organizational chart, or controlled vocabularies for internal drug development projects, or any other proprietary source of terms.

Terminology extraction

I2E can be used for effective terminology development directly from the target corpus of interest, to construct tailored dictionaries. This data-driven approach ensures that terminologies reflect real usage within a domain. I2E is used for faster development of new terminologies, and to expand existing terminologies with missing coverage. It can be used to extract new synonyms for existing concepts or to find new members of an existing class.

Source-specific ontologies for I2E OnDemand

I2E OnDemand is Linguamatics' SaaS offering, with key life science data sources available for immediate text mining. I2E OnDemand provides ready-to-access content from MEDLINE, ClinicalTrials.gov, FDA Drug Labels, FAERs, NIH Grants, OMIM, PubMed Central, and full-text Patents. Some of these sources need their own ontologies to enable users to mine text effectively. Linguamatics has developed specific ontologies to cover these:

- ClinicalTrials.gov: Overall status, study phase, study type, study arm type, and more;
- OMIM: Chromosome, clinical synopsis categories, inheritance, mapping method, and more;
- FAERS Fields Values: Drug characterization, drug administration route, patient age group, seriousness, and more;
- FDA Drug Labels: FDA document types (e.g. bulk ingredient, cellular therapy, human OTC drug); FDA Products (list of all products within data source);
- NIH Grants Codes: Classes for the types of grant awarded, e.g. Activity, Administering IC, Application type, Funding IC; and
- Patents: Co-operative Patent Classification (CPC) codes, e.g. A: human necessities; C: chemistry, metallurgy.

Ontology manipulation for improved text mining

From the initial sources (see Table 1), Linguamatics processes and improves each ontology to work optimally for text mining. This includes the following.

Increasing recall

- Synonym expansion for key ontologies; for example, expanding EntrezGene concepts with synonyms from UniProt. Linguamatics also uses statistically generated synonyms that are then human-curated, and adds synonyms based on syntactic patterns (e.g. liver cancer, cancer of the liver).
- Word variations, so that one synonym can match many terms in the underlying text. Variations covered include case, morphological variants (-ing, -ed), fuzzy matching (e.g. Raf II, Raf 2), glyphs, dialects (tumor vs. tumour) and accents.
- Spelling and optical character recognition (OCR) correction.

Increasing precision

- When terminologies are updated, Linguamatics checks for noisy terms using a mixture of frequency analysis and regression analysis, followed by manual review. Concept matches are weighted automatically with confidence scores. For example, acute lymphocytic leukemia has the synonym "ALL." Many systems would just skip ALL due to noise from, e.g. "All systems." I2E uses sophisticated methods to find ALL only where appropriate.

Table 1: Sources used for Linguamatics I2E ontologies or available as "ready-to-use" ontologies. Due to licensing requirements, some ready-to-use ontologies are available on request and/or are password protected (e.g. MedDRA). Note that Linguamatics has experience of helping customers use a wider range of terminologies than listed here for specific requirements.

ChEBI	MedDRA
ChemAxon Name to Structure	MeSH
Entrez Gene	NCI Thesaurus
Human Phenotype Ontology	Orphanet
ICD-9-CM	RxNorm
ICD-O	SNOMED CT
LOINC	UniProt

I2E strategies for semantic identification and annotation

I2E uses a range of strategies to identify concepts in the right context, including:

- use of thesauri, vocabularies, taxonomies, and ontologies for concepts with known terms;
- pattern-based approaches for categories such as measurements, mutations, and chemical names that can include novel, unseen terms;
- increased accuracy using NLP and disambiguation to discern the context;
- domain-specific, rule-based concept identification, annotation, and transformation;
- integration of customer vocabularies to enable focused and bespoke annotation; and
- advanced range search to enable identification of data ranges for dates, numerical values, area, concentration, percentage, duration, length, weight, volume, and many other concepts.

I2E enables precise, comprehensive, reliable data transformation, identifying the key concepts in unstructured text, and normalizing the term found in the text to a standard "primary" label. This creates a more standardized, semantic representation of data, which enhances clustering of results (e.g. for clear domain-relevant facets), and integration and loading of concepts into other databases and semantic stores.

Why wait?

I2E is a world-leading, agile, scalable, real-time NLP-based text-mining solution, powered by a wide and flexible range of ontologies, vocabularies, and dictionaries. I2E already helps top pharmaceutical companies and healthcare providers speed effective drug discovery, development, and delivery of healthcare therapeutics. To understand the power of NLP text analytics to transform your unstructured documents into actionable knowledge, contact us at **enquiries@linguamatics.com**