Linguamatics

Datasheet

I2E for enterprise scale data transformation

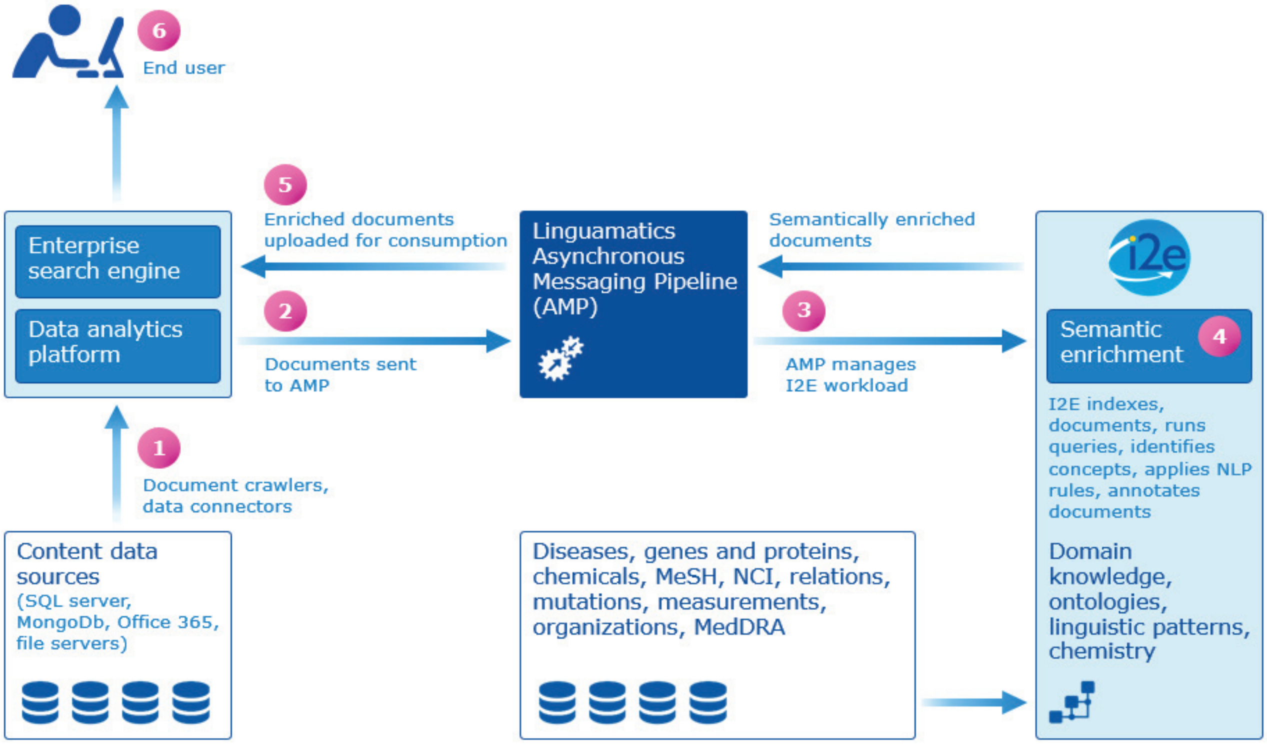
Structuring the unstructured universe

Linguamatics I2E and AMP offer unparalleled capabilities within the life science and healthcare domains to realize timely insights through semantically enhanced document annotation, fact extraction, and ongoing data transformation.

I2E provides an enterprise-scale technology for identifying and annotating concepts within unstructured and semi-structured text (including scientific literature, patents, internal documents, safety reports, regulatory documents, electronic health records). I2E comes with a toolbox of methods that can be applied within enterprise search systems, data integration workflows, ETL/ELT processes, and more. AMP (Asynchronous Messaging Pipeline) supplies workflow management for high-throughput, fault tolerant, real-time document processing.

In combination, Linguamatics I2E and AMP scale to scan millions of documents to identify and annotate semantic entities such as genes, drugs, diseases, biomarkers, organizations, authors, and other relevant concepts and relationships, including complex constructs that require linguistic processing (e.g. smoking status). This can then be consumed by enterprise applications for search and data analytics.

Figure 1: I2E data transformation workflow for enterprise search or data analytics: (1) enterprise applications crawl data sources and repositories to find documents; (2) file-level metadata and documents passed to AMP; (3) AMP manages workload and distributes load across I2E servers; (4) advanced I2E NLP engine identifies and annotates key concepts in documents, using powerful, domain-relevant vocabularies and linguistic rules; (5) normalized concepts are fed back into enterprise applications for (6) the end user to access.



6
End user
5
Enriched documents uploaded for consumption
Enterprise search engine
Data analytics platform
2
Documents sent to AMP
Linguamatics Asynchronous Messaging Pipeline (AMP)
Semantically enriched documents
3
AMP manages I2E workload
i2e
Semantic enrichment
4
I2E indexes, documents, runs queries, identifies concepts, applies NLP rules, annotates documents
Domain knowledge, ontologies, linguistic patterns, chemistry
1
Document crawlers, data connectors
Content data sources
(SQL server, MongoDb, Office 365, file servers)
Diseases, genes and proteins, chemicals, MeSH, NCI, relations, mutations, measurements, organizations, MedDRA

I2E 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 natural language processing (NLP) and disambiguation to discern the context;
- domain-specific, rule-based concept identification, annotation, and transformation; and
- integration of internal vocabularies to enable focused and bespoke annotation.

I2E enables precise, comprehensive, reliable data transformation, identifying the key concept 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.

Biomedical terminology

- Genes, proteins
- Diseases, adverse events
- Processes and pathways
- Organs, tissues, cells
- Organisms
- Mutations

Sources: MeSH, MedDRA, NCI Thesaurus, Entrez Gene, Gene Ontology

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, companies and organizations, patient demographics, and more.

Healthcare terminology

- Medical codes
- Biomarkers
- Tumor staging (TNM)
- Patient demographics
- Patient history (e.g. smoking status)

Sources: SNOMED, ICD9-CM, LOINC

Medical terminologies can be integrated into I2E. Key healthcare categories are recognized using a combination of pattern-based approaches and linguistic rules, such as TNM staging (tumor, nodes, metastasis notation), smoking status, history of disease, and ambulatory status.

Chemistry

- Compounds by name
- Drugs, excipients, active ingredients
- Chemistry by structure

Sources: ChEBI, MeSH, NCI Thesaurus, ChemAxon

I2E Chemistry, with ChemAxon, identifies known and novel chemical structures within documents, by name, structure, substructure, or similarity.

Organizations

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

Source: Linguamatics, MeSH

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).

Numerics, metrics

- Age class
- Date class
- Drug lab codes class
- Email class
- Measurements class
- Mutations class
- Numerics class
- Person class
- Telephones class
- Units class

Source: Linguamatics

I2E includes pattern ontologies that identify categories of information, such as dates, mutations, organizations, telephone numbers, or units of measurement. Pattern ontologies are incredibly valuable for identification of concepts that can be expressed in many ways, but for which there are limited dictionaries available. These pattern ontologies extend search far beyond the ontology-matching approach, to annotate novel descriptions in text of key concepts or concept types.

Linguistic components

- Relations ontology
- Linguistic units (word, sentence, noun group, verb group)

Source: Linguamatics

NLP provides linguistic components to search and mark-up, enabling the user to focus on, for example, the relations between concepts (e.g. search for compounds causing disease rather than treating disease), or 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 five 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.

Linguamatics AMP

AMP provides high-throughput, fault tolerant, real-time document processing, for use cases such as semantically enhanced enterprise search, text ETL for data warehouses, and rapid textual processing of electronic health records for Clinical Risk Monitoring. AMP empowers workflows that process and transform sets of documents asynchronously using I2E. It provides:

- integration via a rich RESTful API or directory monitoring;
- integration in Service Oriented Architectures;
- real-time scalability, both vertical and horizontal;
- high availability;
- resilience to failure;
- data security;
- configurable workflows for multiple source document types;
- flexible pre- and post-processing stages; and
- output of highlighted or annotated document, extracted facts, or RDF triples.

AMP manages multiple I2E servers for indexing and querying, distributing resources, and buffering incoming documents, and is powerful enough to handle millions of records.