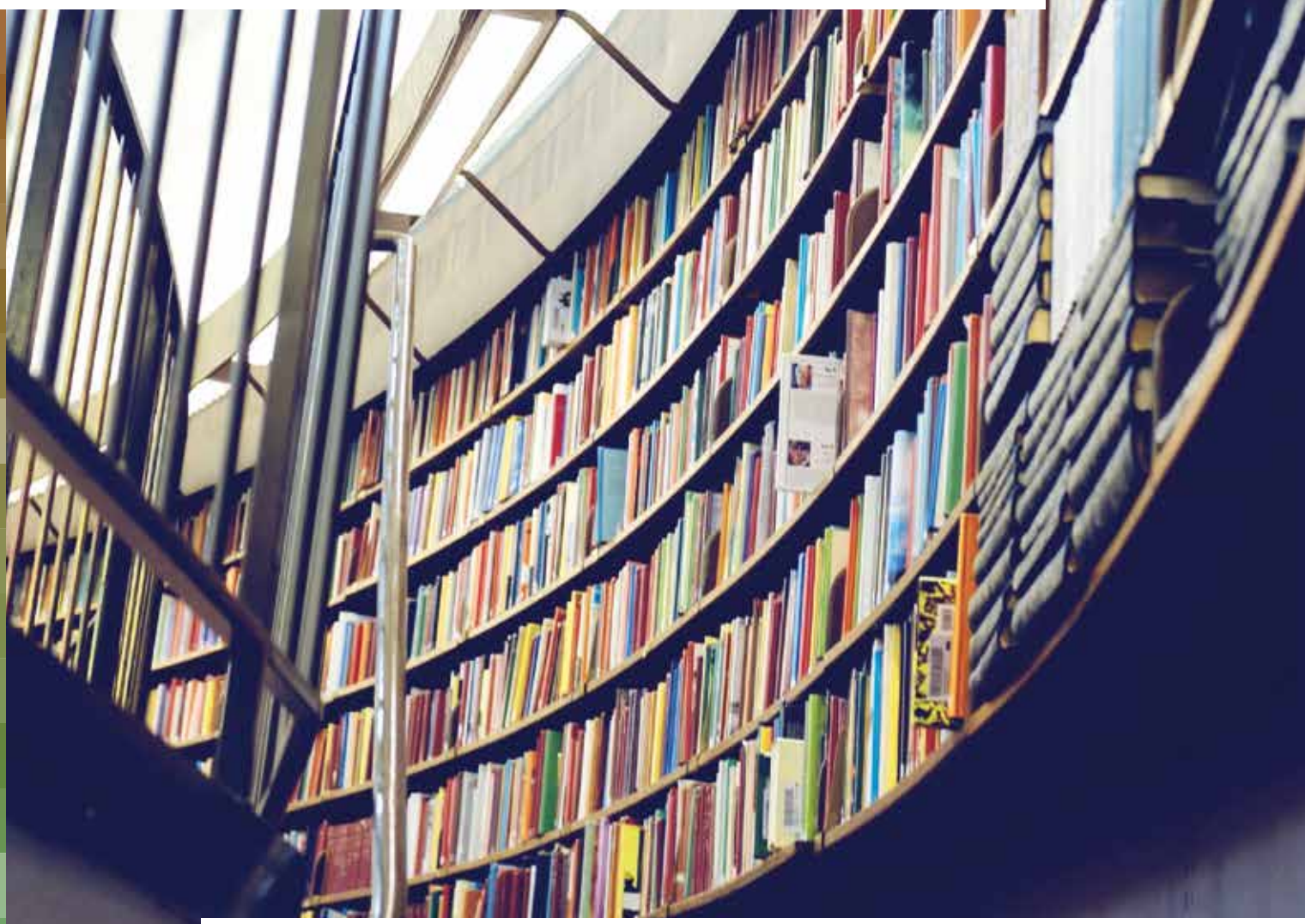
Data Sheet

Linguistics Plugin

Linguistics Plugin for Elasticsearch and OpenSearch

Semantic-linguistic analysis for better search results

With its Linguistics Plugin, IntraFind enables users of Elasticsearch and OpenSearch to get more complete and more relevant search results. At the same time, the plugin is easy to integrate and configure – even for content in multiple languages. The Linguistics Plugin is an integral part of the Elasticsearch-based search solution iFinder.



INTRAFIND

Your advantages:

- + Better search results: higher recall & higher precision
- + Don't miss anything: complete hit list by using lemmatization and compound decomposition at the highest quality level
- + Lexical-morphological normalization: very high degree of lexicon coverage by combining manually curated and procedural lexicons
- + Support for more than 30 languages
- + Easy to install, configure and use: no complex mappings / settings required even in multilingual environments thanks to integrated language recognition
- + Automatic language segmentation of mixed-language content
- + Optional integration of thesauri, also for cross-lingual search
- + Normalization of numerical values (e.g. dates, prices) and scientific-technical units (approx. 50 different units, e.g. length, area, volume, temperature)
- + Optional recognition of entities such as people, organizations, and places
- + Powerful query operators for precise formulation of search queries
  - Different search modes - from exact search to search for phonetically similar words / names
  - NEAR operators, e.g. for „search within a sentence" or within a definable word distance
  - Operators for numerical search, unit (measurement) search and entity search

Perfect search results with premium linguistics

The IntraFind Linguistics Plugin for Elasticsearch will also return hits containing the words „basketball" or „football" when searching for „ball", but not with „ballet". The user does not have to be an expert in formulating wildcard queries for this. A search for the word „landowner" will return hits for „owner of the land" based solely on morphological normalization and word compound decomposition.

For the user, search becomes much more comfortable, as he no longer has to worry about word variants and spellings. The use of expert queries with e.g. wildcards like *boat to search for all kinds of boats is a thing of the past. Inflections in any of the more than 30 supported languages such as the French beau -> belle, are taken care of.

In addition to high-quality morphological normalization, additional splitting of complex tokens ensures that product searches for „iphone12" and „iphone 12", for example, return the same results.

The plugin provides various search modes (e.g. an exact search in addition to the search in the morphologically normalized index) and powerful NEAR operators (e.g. search within a sentence). As part of these NEAR queries, arbitrarily complex queries are allowed, e.g. also wildcard queries. An extended query syntax is used for this purpose. Of course, the complete query syntax familiar from Elasticsearch is supported, including simple queries such as the match query.

An advantage of the linguistics plugin that should not be underestimated is its simple configuration, especially in multilingual environments. The typical effort of having to configure a separate field and analyzer in the schema for each language, along with controlling the indexing using a language recognition component, is completely eliminated. This is due to the fact that the IntraFind Analyzers already include language recognition and even in the case of mixed-language content, parts written in different languages are recognized and automatically handled correctly.

Lemmatization and composite decomposition

The Linguistics Plugin combines different methods of linguistic text mining to normalize inflected words to their base forms **(lemmatization)** and to decompose compound terms into their basic components **(compound decomposition)**. This is particularly important for the numerous German and Dutch multi-word terms, but also for simple cases in other languages. In addition, simple word categories (noun, verb, adjective, functional) are also provided.

Examples:
Lemmatization:
mice → mouse
journal → des journaux (F)
aficionada → aficionado → aficionar (ES)

Compound Decomposition:
sailboat → sail, boat
moonlight → moon, light
landowner → land, owner

The linguistic procedures of the plugin are based on extensive lexicon resources and are additionally characterized by a strong procedural orientation in the lexicon analysis. Linguistic results, especially for word decomposition, are therefore, in contrast to comparable procedures, „computed" by procedures and not „looked up" in the lexicon. The extensive lexicons are prepared and compiled in a high-quality and efficient way, so that the analysis is carried out quickly and saves memory. Since languages develop dynamically and have a great creative potential to create new words or word combinations, it would be inefficient to rely only on rigid lexicons.

Against this background, the Linguistics Plugin works with the basic morphological building blocks of a language on the basis of basic lexicons (full form-base form mapping). These building blocks, as well as the combinatorial rules for their analysis, can be determined and applied with a high degree of completeness. In this way, new word combinations, especially compound words, are successfully analyzed.

The plugin offers a comprehensive set of lexicons and procedures for analyzing new word creations. Moreover, the lexicons are regularly extended by neologisms (new words). The search based on lemmatization and compound decomposition is much more precise than the search usually used in search engines. The latter is based on simple algorithmic procedures for word stem formation, which in many cases tend to overgeneralize and trace words with completely different meanings back to the same artificial stem.

Compound decomposition not only causes a search for „moon" to also return hits for „moonlight". If the search query contains a compound word such as „moonlight", a near-query with the components is automatically generated, which also returns a hit for „light of the moon".

Use of thesauri

In numerous search projects, the quality of the search is significantly enhanced using thesauri. Many of these improvements are already provided automatically by lemmatization and compound decomposition of the linguistics plugin. Using thesauri, the linguistics plugin can furthermore expand original search queries with real synonyms and abbreviations (such as car, automobile, motor vehicle), hyponyms (more specific terms) or hyperonyms (more generic terms).

The plugin therefore allows the use of arbitrary thesauri and supports common thesaurus formats, such as SKOS. By using lemmatization and compound decomposition also for the lookup in the thesaurus, no inflection forms need to be maintained in the thesaurus resources. This greatly simplifies the maintenance of customer-specific thesauri and the use of existing thesauri.

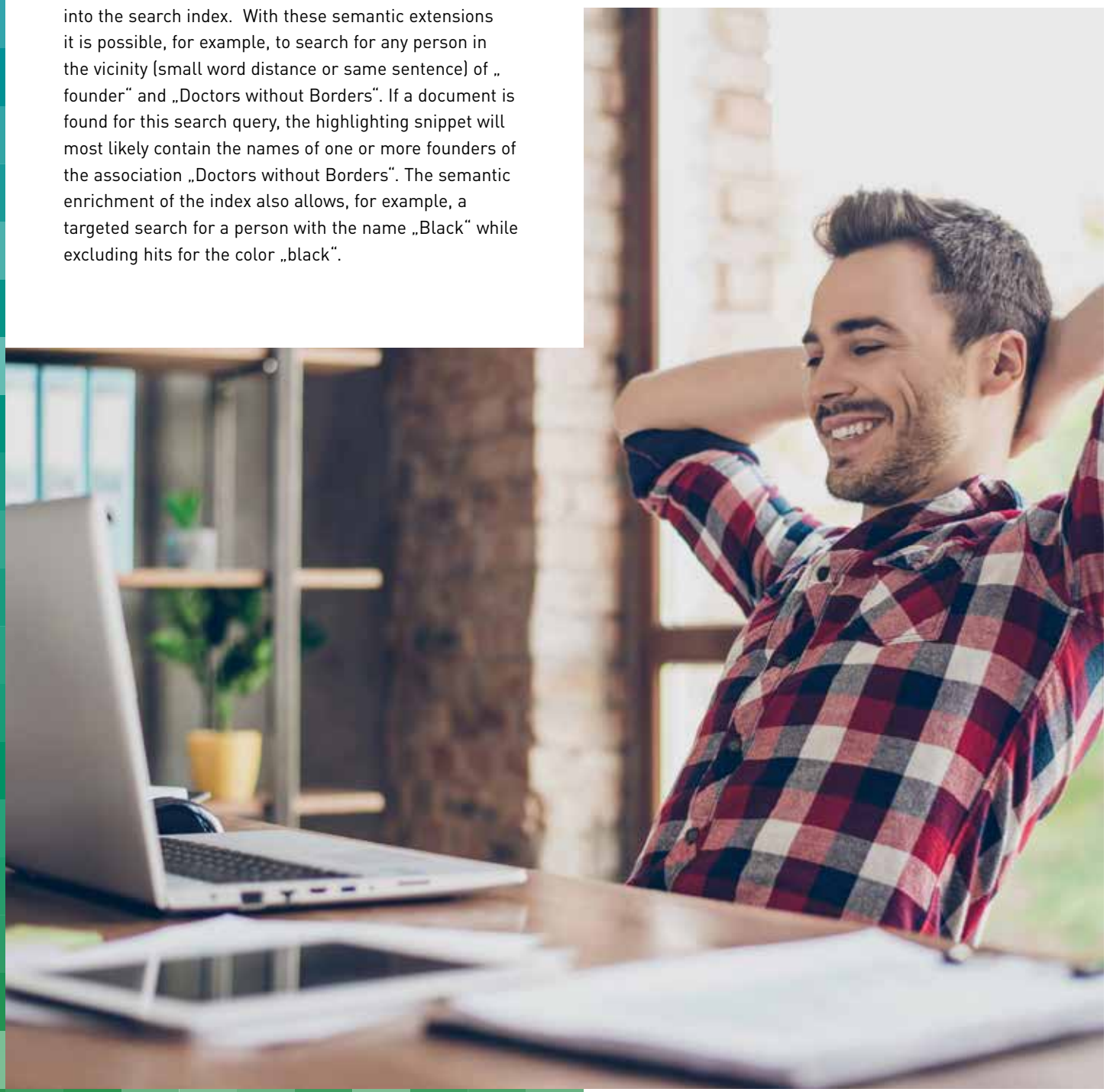
Multi-word terms are fully supported, also for the lookup. We optionally offer general language thesauri for German, English, French and a cross-lingual German-English thesaurus.

Search for entities, numerical search and unit search

Names, e.g. of persons, organizations and places (addresses), often play a special role in search. This also applies to numerical data such as prices or dates or to technical-scientific units such as area specifications, speeds or temperatures. The intention behind many search queries is often a factual question or W-question (Who?, When?, Where?, How much?).

The linguistics plugin offers the recognition of numbers and units (optionally also of entities) and their integration into the search index. With these semantic extensions it is possible, for example, to search for any person in the vicinity (small word distance or same sentence) of „ founder“ and „Doctors without Borders“. If a document is found for this search query, the highlighting snippet will most likely contain the names of one or more founders of the association „Doctors without Borders“. The semantic enrichment of the index also allows, for example, a targeted search for a person with the name „Black“ while excluding hits for the color „black“.

Searching for any amount of money in a text close to rent and address components as a broader context could provide the answer to the rent amount of an object. The numerical search can not only provide answers to questions like „How much?“. It is also possible to search for amounts in a specified interval. In addition, recognized units are automatically normalized. The search for the words „boiling point“ and a „temperature of 90 - 110 °C“ in the same sentence also delivers hits for the sentence „The boiling point of water is 212 °F“ thanks to the unit normalization. Searching for „Display 5 inch“ will also return results for „Display diagonal 4.8 inch“.



Highlights of the Linguistics Plugin:

- Currently over 700,000 single word lexemes (simplex lexemes) in the German lexicon and more than 100 procedural rules, e.g., for the handling of street names like Berliner Straße or Berlinerstraße. 50,000 subject terms and proper names for German
- Similar high-quality resources for 15 other European languages
- High quality statistical tokenization for Asian languages such as Chinese and Japanese
- Supports improved search for more than 30 languages
- Optional integration of 3rd party Lucene / Elasticsearch analyzers to support even more languages
- Optional search for proper names (persons, organizations, places, addresses, etc.) for German, English, and Spanish
- Optional integration of 3rd party components for named entity recognition
- Annual update with new word formations (neologisms)
- Numerical search and search for units (measurement)
- Optional: general language synonym thesaurus German, English, French
- Optional: German-English dictionary for cross lingual search
- Integration of arbitrary thesaurus resources
- Multilingualism: very easy configuration in multi-lingual environment (content in multiple languages)
- Mixed Language Documents: are processed correctly, detection of language change in the text
- Built-In sentence boundary detection for even better relevance and search within a sentence
- Various search modes (exact, standard, phonetics)
- Arbitrary nested NEAR operators (Phrase with wildcard queries: NEAR/1(N* York))

In addition to simple full-text queries, powerful search operators allow the formulation of very precise search queries:	
MODE/E&MAN	exact search for MAN (capitalized) - (MAN Truck & Bus as company) -> no hits for man (lower case)
MODE/P(Tschaikowsky)	phonetic search for similar sounding names
NEAR/0(func* AND NOT function)	Search for words beginning with func; ignore hits for the word function
NEAR/0(Black AND ENTIY/PERS)	Search for a person named Black (not the color)
UNIT/(300 thousand km/s; 10.000 km/s)	Search for a velocity near the speed of light (+/- 10.000 km/h)
NEAR/S(UNIT/*(liter) AND BEER)	Search for a phrase that contains beer and an arbitrary volume
NEAR/S(rent AND UNIT/(1.000 €; 100€))	Search for the word rent and an amount of 1000 € +/- € 100 in the same sentence.
NEAR/20((ENTITY/PERS OR ENTITY/ EMAIL) AND "SAP Account")	Search for a person name or email address near the multi-word term "SAP Account". This might answer the question of "who you can get an SAP Account from."

Languages available

German
English
French
Italian
Spanish
Dutch
Polish
Croatian

Greek
Czech
Portuguese
Hungarian
Romanian
Slovak
Slovenian
Arabic

Turkish
Ukrainian
Norwegian
Danish
Swedish
Finnish
Russian
Korean

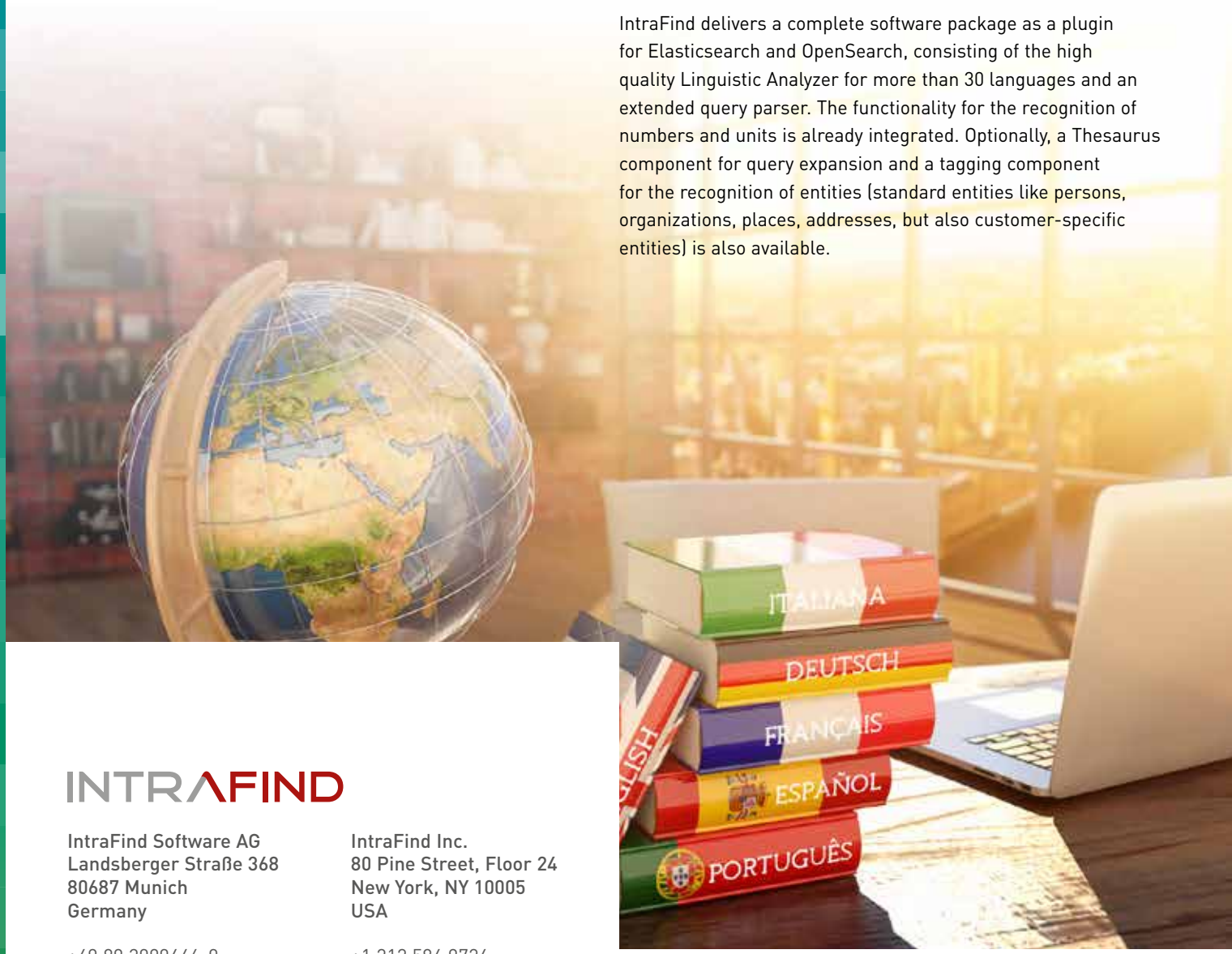
Japanese
Thai
Indonesian
Malay
Hindi
Chinese*

* (Mainland and traditional / Taiwanese including normalization)

Are you missing a language? Talk to us!

Scope of delivery

IntraFind delivers a complete software package as a plugin for Elasticsearch and OpenSearch, consisting of the high quality Linguistic Analyzer for more than 30 languages and an extended query parser. The functionality for the recognition of numbers and units is already integrated. Optionally, a Thesaurus component for query expansion and a tagging component for the recognition of entities (standard entities like persons, organizations, places, addresses, but also customer-specific entities) is also available.



INTRAFIND

IntraFind Software AG
Landsberger Straße 368
80687 Munich
Germany

+49 89 3090446-0
sales@intrafind.com
www.intrafind.com

IntraFind Inc.
80 Pine Street, Floor 24
New York, NY 10005
USA

+1 212 584 9724
sales@intrafind.com
www.intrafind.com