CONTEXITY

Semanteer Search

Features & Technical Details

The Semanteer platform, which forms the basis of our search offerings, provides a comprehensive framework of components for collecting, extracting, enriching, and searching information from a variety of data sources, document formats and languages.

Data collection

See below a list of search connectors and authority connectors supported, or read on to find out how access privileges are enforced in combination with the authority connectors. If you need a connector that is not listed here, feel free to contact us to discuss the possibility of custom development.

Search connectors

Email repositories

- Microsoft Exchange Server (including Exchange Online)
- Any email server with support for IMAP or POP3

File systems

- Local disks / mounts
- Windows shared folders (Distributed File System - DFS)

A complete search platform

All the way from collecting data, to semantic analysis and natural language processing

Tailored to fit your needs

Start from one of our standardized solutions, or from scratch, to arrive at exactly what you need.

Based on open source

Impressive sets of features, no lock-in, all bundled, tested and maintained by us

- Networked attached storage (SAMBA compatible)
- Remote WGET-compatible file systems
- Hadoop distributed file systems (HDFS)

Cloud storage

- Dropbox
- Google drive

Databases

- Any database for which a Java Database Connectivity (JDBC) exists, including: Microsoft SQL Server, Oracle, Sybase, Postgresql, MySQL, etc.

Content, Document, and Record management systems

- Autonomy Meridio
- EMC Documentum
- IBM FileNet
- Magnolia
- Microsoft SharePoint
- OpenText LiveLink (now part of the OpenText Enterprise Content Management platform)
- Any content repository that supports the Java Content Repository (JCR) API

Popular open source solutions for web site and online shop management

- Typo3 CMS
- Drupal
- Joomla
- Magento

Web content

- Any web site that can be crawled
- RSS / Atom feeds
- Wikis (enhanced support for MediaWiki)

Intelligent search

Build a rich search experience with modern features and the capacity to learn and automatically improve by monitoring user behaviour.

Full control

A single “cockpit” to configure, manage, and monitor your entire search infrastructure.

Deployed any way you like

On-premise for maximum security, on the cloud for maximum scalability, or hosted for maximum convenience

Popular knowledge and ticket management solutions from Atlassian

- Confluence
- Jira

Through the Content Management Interoperability Services (CMIS) API you can also index content from a long list of Content and Document Management systems: Alfresco, Apache Chemistry, Ceyoniq, Cincom ECM, Day Software CRX, dotCMS, Eyebase mediasuite, EMC Documentum, eXo Platform, HP Autonomy Interwoven Worksite, HP Trim, IBM Content Manager, IBM FileNet Content Manager, IBM Content Manager On Demand, IBM Connections Files, IBM LotusLive Files, IBM Lotus Quickr Lists, ISIS Papyrus Objects, KnowledgeTree, LogicalDOC, Maarch, Magnolia CMS, Microsoft SharePoint Server, NemakiWare, Nuxeo Platform, O3Spaces, OpenCms, OpenKM, OpenText ECM, OpenWGA, Oracle Webcenter Content , PTC Windchill, SAP HANA Cloud Document Service, Surround SCM

Authority connectors

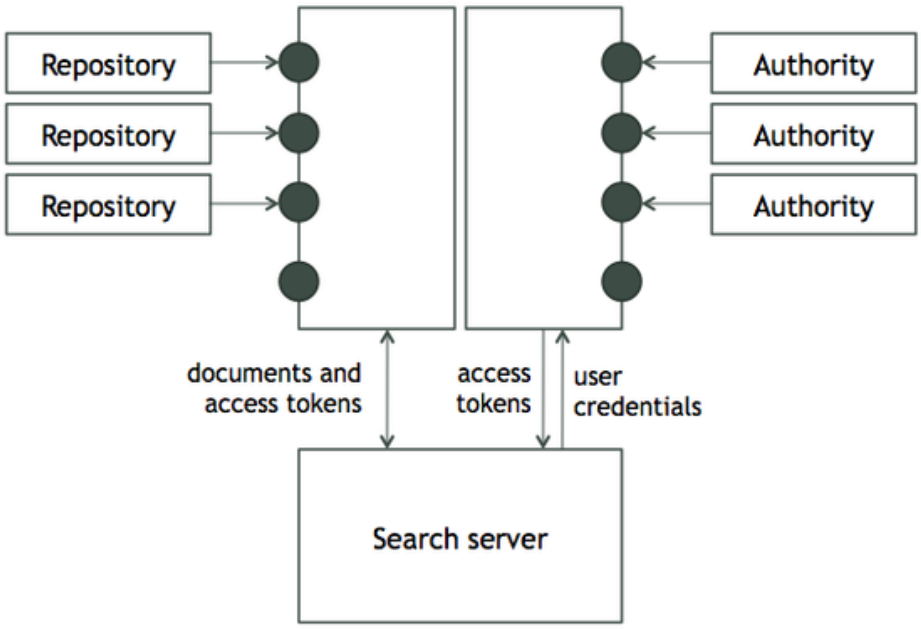
With the provided authority connectors you can retrieve authorization information for your data from:

- Active Directory
- Lightweight Directory Access Protocol (LDAP) compatible servers (such as OpenLDAP)
- Content management systems that support the Content Management Interoperability Services (CMIS) API
- OpenText LiveLink (now part of the OpenText Enterprise Content Management platform)
- Any database for which a Java Database Connectivity (JDBC) exists, including: Microsoft SQL Server, Oracle, Sybase, Postgresql, MySQL, etc.
- EMC Documentum
- Autonomy Meridio
- Microsoft SharePoint Active Directory or Native
- Custom Authority Endpoints, that provide access to your company’s own authorization scheme

Enforcing access privileges

There are two alternative approaches (that can also be combined if desired):

- Access privileges are collected from the respective authority connectors, and the users' access-related information is sent along with the search requests, and checked against the constraints embedded in the search index.
- The users' access-related information is not sent along with the requests, but instead retrieved from the original authority connectors, given the user's access credentials.



Data extraction

It is often the case that valuable knowledge and information is not only spread out in different systems and repositories (see data collection), but also captured in a variety of document formats intended for “human consumption”. To make that information findable, it first needs to be extracted and conveyed in a uniform way to the rest of the system.

Extraction of text and metadata

To achieve that goal, our platform supports extraction of text and metadata from variety of document formats, including : HTML, XHTML, XML and derived formats, Microsoft Office document formats (.doc(x), .ppt(x), .xls(x)), Open Document Format (.odt, .odp, .ods), Adobe Portable Document Format (.pdf), Rich Text Format (.rtf), Feed and Syndication formats (RSS and Atom). Metadata can also be extracted from a variety of common image, video and audio file formats.

Language sensitive text analysis

More than 50 European and Asian languages can be reliably detected in extracted text, with precision reaching levels above 99% for texts that have at least a few sentences. Multiple languages used in the same document / page can also be detected. Once the language (or languages) of text is known, it becomes possible to perform language-specific analysis on it. Examples include:

- handling of special characters in the language (e.g., characters with diacritics, such as the umlaut in German or the accents in French) so that alternative forms of writing do not cause problems with searching, “natural” sorting of text works, etc.

- word stemming (e.g., in English, “go”, “going” and “went” all refer to the same verb, and “child” and “children” to the same concept; in German the same applies for “laufen”, “lief” and “gelaufen”, and “Kind” and “Kinder”)
- removal of language-specific contractions and “stop words” (e.g., articles, helping verbs)
- expansion of terms with language- or application domain- specific synonyms, either from simple synonym lists, or from full-featured externally maintained thesauri
- etc.

Recognition of “real content” in pages

When extracting text from web pages that have been crawled, it is important for the extraction process to mimic -to the extent possible- a task that humans find almost trivial: identify the part of the page that contains the “real content”, leaving out banners, navigation areas, headers and footers, repetitive or unrelated content in sidebars, etc. To achieve this, humans use -subconsciously- a significant number of complex heuristics, that are difficult to capture and implement algorithmically. Nevertheless, this is a crucial part of the extraction process so that repetitive and insignificant content doesn’t interfere with the quality of the search results.

We support two approaches to identifying the “real content” of pages:

- In the “assisted” approach, the extractor is provided with “hints” regarding where in the page it should look for meaningful text; such ‘hints’ may range from approximate to very precise and refer to characteristics of the text, the spatial arrangement of content on the the page, or even to special “markers” that have been inserted in the page for that purpose.
- An alternative approach that can reach quite high levels of precision is entirely automatic and uses algorithmic heuristics to approximate the decision process of humans as to what is “central” to a page. While not always as effective as its “assisted” counterpart, this approach requires no human intervention, no special features embedded in the content, and can be applied successfully over pages that have entirely different structures (e.g., in a federated search scenario).

Extraction of semantic metadata

Many modern document-, content- and web site- management systems allow content creators to annotate the textual content they generate with semantic metadata, which bestow “meaning” to content elements (e.g., elements of an address, name and function of an employee in a company, etc.) or “connect” pieces of content to entities or concepts (e.g., to indicate the creator of content, express social relationships between people, specify the details of a software project, etc.)

Our platform supports the seamless extraction of such semantic metadata from content to be indexed, so that they can be used in a structured form during searching, or in later stages of analysis.

Supported formats include:

- RDF/XML, Turtle, Notation 3
- RDFa with RDFa1.1 prefix mechanism
- Microformats: Adr, Geo, hCalendar, hCard, hListing, hRecipe, hReview, License, XFN and Species
- HTML5 Microdata (such as schema.org)
- JSON for Linking Data (JSON-LD)
- Comma Separated Values (CSV) with separator autodetection.

The vocabularies that are understood (independently from the format they are expressed in) include: Dublin Core Terms, Description of a Career, Description Of A Project, Friend Of A Friend, GEO Names, CAL, Ikif-core, Open Graph Protocol, BBC Programmes Ontology, RDF Review Vocabulary, schema.org, VCard, BBC Wildlife Ontology and XHTML.

Search

At the core of any search platform are the actual search capabilities it provides, which go well beyond the character-by-character matching of search terms to words and phrases in the search engine's index. Building upon the most advanced open source search engine framework provided by Apache Lucene and Apache Solr, our platform provides a wide set of state-of-the-art search facilities:

Categorizing and filtering of results

Language is often ambiguous and words and phrases may appear in different contexts, different sources, and certainly a multitude of documents. To make it easier for users to "drill down" in the set of search results and to narrow down the search space, "facets" of the results can be exposed to, and controlled interactively by, the user. Examples include the source from which a result was retrieved, the type of document, the language used, etc.

In more structured domains, the categories and filtering capabilities may be directly related to rich "attributes" of the results. For example, in an online shop these may include product type, price range, availability, etc. Our platform also readily supports the "progressive disclosure" of facets on the basis of their relevance to the currently applied filters. For example, facets related to shoe sizes available are only shown when the user has narrowed down their search to shoes.

Auto completion

A feature that users have come to expect (and often heavily rely on) is the automatic completion of partial search terms they enter with suggestions of possible, meaningful full words and phrases. Our platform supports a wide range of algorithms and approaches to extracting such suggestions, including:

- direct extraction from the indexed content, taking into consideration how often words and phrases occur, and how distinctive they are in the indexed text; alternative approaches can be used to determine how suggested phrases are selected or built, to better match the search domain (e.g., in an online shop, autocompletion can be configured to suggest full product names or product categories rather than individual words thereof)
- utilization of a set of search terms that are used often by users, ranked by frequency and recency, so that search "themes" or topics that are "always" relevant in a search corpus are recognized, but ones that are of only temporally or situationally limited significance (e.g., regarding a subject that is currently making headlines, but is of lesser interest a couple of weeks later) are also accounted for
- employment of curated thesauri of search phrases that are maintained by human editors to match the search domain
- and, of course, combinations of the above

Fault tolerant input and spelling suggestions

When entering search terms, users sometimes misspell words, or simply don't know the correct spelling of things they are looking for. User expectations today dictate that the search engine recognizes such errors, to the extent possible, and either provides the correct results anyway, or, at least, suggests alternative spellings. To meet such expectations, we support:

- the extraction of spelling suggestions directly from the indexed text (with "stop words" automatically excluded), or, alternatively, the use of language-specific dictionaries (with combined approaches also possible)
- automatic "re-searching" which entails the automatic delivery of results for the highest ranked of the alternative suggested spellings under configurable circumstances (e.g., if the number of results is under a specified count), as well as the delivery of information on what was "wrong" with the original search term(s) and how many results they would have resulted in

Control of result ranking

In different search domains, different attributes of the results may be decisive in how these should be ranked. For example, when searching through documents, the occurrence of search terms in their titles or section headlines can be presumed to be of higher significance than the occurrence of the same terms in the document's text. If the search domain were films, then the actors' and the directors' names would be of almost equal significance as the films' names.

To fully capture the idiosyncrasies of your search domain, we provide a set of preconfigured "profiles" that determine the importance of matches of search terms against different result attributes, and are happy to work with you to further configure and fine-tune them -or create entirely new ones- for your domain.

Promotion / boosting of results

In some cases it is desirable to “boost” specific search results, so that they are listed higher than others that would be of similar levels of relevance with respect to the user’s search terms. Examples may include promotional or sponsored items, paid listings, etc. There are two mechanisms in place that support such “boosting”: the first allows you to define specific search results or specific result attributes (e.g., products with higher than average stock) that are ranked higher independently of the current search terms; the second is similar, but allows you to further specify keywords that “trigger” the boosting when they occur in the search query (e.g., help desk-related documents could be promoted whenever the word “support” is used in a search).

Monitoring of user search behaviour

What users search for, how many and which results they receive, which results they actually opt to follow up on, are all very useful information for both understanding what your users look for on your site, and how they think about the content. This information, which is collected and saved, can be directly used for search analytics reports, or as a basis for further intelligent search features (see also related documents, related searches, autocomplete, and more)

Related documents

Typical intranet and federated search solutions bring together documents and data from multiple sources. A lot of added value can then emerge from simply offering users the possibility to easily find related content across sources, such as, for instance, a report that refers to the same topic as an email listed in the search results. We offer two complementary mechanisms for establishing such relationships:

- through the analysis of document contents and the identification of important terms that are shared amongst documents
- through the users’ behavior when perusing search results (e.g., when several users that search with similar search terms open the same two documents from the results, these documents are likely to have a conceptual relationship to the topic expressed by the search terms, even if such a relationship cannot be detected through content analysis)

Alternative / related searches

Sometimes users are not aware of the best search terms to use to find what they are looking for, because they only have a fuzzy idea of their target, because they are not familiar with the terminology of a domain, etc. One form of support that can be offered to remedy that is to suggest alternative formulations of the search terms, or related / alternative searches altogether. These can be derived from searches issued by other users (e.g., when users reformulate themselves their query within a search

session to make it more specific or precise), or from the semantic analysis of the terms used (e.g., replacing words with synonyms that occur more often in the search corpus).

Language-aware & language-agnostic search

When text is extracted from the source documents or records to be indexed, it is possible to automatically determine the text's language, and analyse the text accordingly (see section Language detection and language sensitive text analysis). This type of approach works very well for mainly text-based data, such as documents, emails, web pages, etc.

In some application domains, however, this is not the ideal approach. Consider for instance texts that mix more than one languages in a single sentence, or where a lot of domain-specific terms and abbreviations are used often, or even product names that share portions of compound words. In such cases you may opt for an alternative approach to indexing and searching that is based on "N-grams", which provides efficient support for searching even "within" words for matches.

Our tools also allow you to mix the two approaches to fit your needs, with a typical setup for high end installations involving language-aware retrieval as a "first layer" mechanism, and language-agnostic, N-gram based retrieval as a "fallback" search mechanism.

Context-sensitive search results

What part of a website users search from, what pages they have seen, what they have searched for before, etc., are all indications of the user's "context" which may provide additional evidence about what kind of information the user is looking for. These indicators can be used automatically to bias the ranking of search results, effectively adding information to the user's search terms that help arrive faster at the desired results. A simple example: in a retail website that has different sections for private persons and businesses, a user that has spent time looking at the business pages, and starts a search from the page of a particular product, may well be interested in similar products for business customers, so any results that match that profile can be ranked automatically higher.

Geo-location based search

In some search domains, the physical location of things play a very significant role. Examples include finding a particular type of shop in your area, or the closest outlet of a chain, noteworthy sites in a city, the locations of interesting events, etc.

To satisfy the needs of such domains we provide advanced geo-location based search features, coupled with advanced interactive front ends, that can be based either on the popular Google Maps, or the OpenStreetMap open and free map database.

Data enhancement

Some of the most useful information hidden in data requires the kind of interpretation that only humans are capable of today. But it is not entirely beyond reach! Deep content analysis in the form of natural language processing, and data mining of users' search behavior can reveal a lot of the "hidden" pieces of information and the relations between documents and users.

Natural language processing

Although there exist facilities for structuring content semantically (see Extraction of semantic metadata), the vast majority of content available today is unstructured text. Information regarding persons, addresses, locations, etc., as well as the links between them, are often expressed in simple text, which makes it difficult to identify and utilise beyond the scope of individual documents.

To support the extraction and use of such unstructured information, our platform offers natural language processing capabilities for English, German and French that enhance simple text by:

- part-of-speech tagging of words and phrases
- extraction of topics
- extraction of named entities
- categorization of named entities (person, location, etc.)
- mapping of named entities to concepts and external "known" entities (e.g., from DBpedia)

In relatively "closed" information domains, such as, for example, a company intranet, these capabilities can be extended to include the creation of relationships between different types of entities (e.g., company employees that work in a specific department and have been involved with a particular project) and to make these searchable also.

Intelligent search support

Beyond the indexed content itself, an extremely valuable source of information that can be used to improve the search experience is the (monitored) behavior of users that interact with the search front-end (see Monitoring of user search behaviour). The information collected can include a variety of data points: where in the system the search was started from, what was searched for, what results were shown to the user, what was actually clicked, etc. These, in turn, can be used to intelligently improve results and accompanying support features, to better match the end users' needs:

- The frequency and recency with which search terms and phrases are issued by different users can serve as input for autocomplete suggestions, and "related searches" suggestions

- “Item-to-item collaborative filtering”, a technique widely used in product recommendation, can be used to discover relationships between seemingly unrelated search results, based solely on users’ selections of results for identical or similar search terms (see Alternative / related searches)
- The mapping of search terms to concepts (especially in search spaces with a structured information domain) makes it possible to create a “profile” of the user’s information needs and interests. This can, then, be used as additional input in future search sessions personalize search results by disambiguating search terms, and, more generally, promoting results that fall within the user’s sphere of interests.

Another variation of collaborative filtering, namely the “user-to-user” variant, can put such user profiles into further use: for any given user, it is possible to identify the users “closest” to him or her, based on similarities in their profile. Establishing the user’s so called “neighborhood” opens up a new set of capabilities, making it possible, for instance, to assess the likelihood that a result is of interest to an individual, on the basis of whether other similar users have found it interesting or not.

Our search solutions are built on top of mature, enterprise-quality, widely-used open source frameworks, libraries, and components.

Open Source Technologies

Apache ManifoldCF provides a framework for connecting source content repositories like file systems, DB, CMIS ... to target repositories or indexes, such as Apache Solr.

Apache Nutch is a mature, highly scalable web crawler which provides extensible interfaces for parsing (for example for Tika), indexing (for example through Solr, SolrCloud, ...), and filters for custom implementations.

Apache Tika is a library to detect and extract metadata and text content from various types of documents.

Any23 is a library to extract structured data in RDF format from a variety of web documents. Any23 supports metadata standards such as microformats and HTML5 microdata like schema.org.

Apache UIMA provides an architecture for unstructured information management by specifying component interfaces in an analytics pipeline, data representations, and a set of design patterns. Since 2009, UIMA is an OASIS standard for content analytics.

ClearTK is built on top of Apache UIMA and provides a framework for developing statistical natural language processing (NLP) components.

Apache Stanbol provides a set of reusable components for semantic content analysis, semantic enrichment, reasoning, and the definition and storage of knowledge models.

Apache Lucene provides Java-based indexing and search technology, as well as spellchecking, hit highlighting and advanced analysis/tokenization capabilities which form the core of powerful full text search solutions. http://lucene.apache.org

Apache Solr is the most popular open source enterprise search platform. Solr is based on Lucene and its major features include full-text search, facetting, geospatial search and a lot more.

OpenRDF Sesame is a de-facto standard framework for processing RDF data. This includes parsers, storage solutions, reasoning and querying, using the SPARQL query language.

Apache ZooKeeper is a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services.

MALLET is a machine learning toolkit for statistical natural language processing, document classification, clustering, topic modeling, and information extraction.

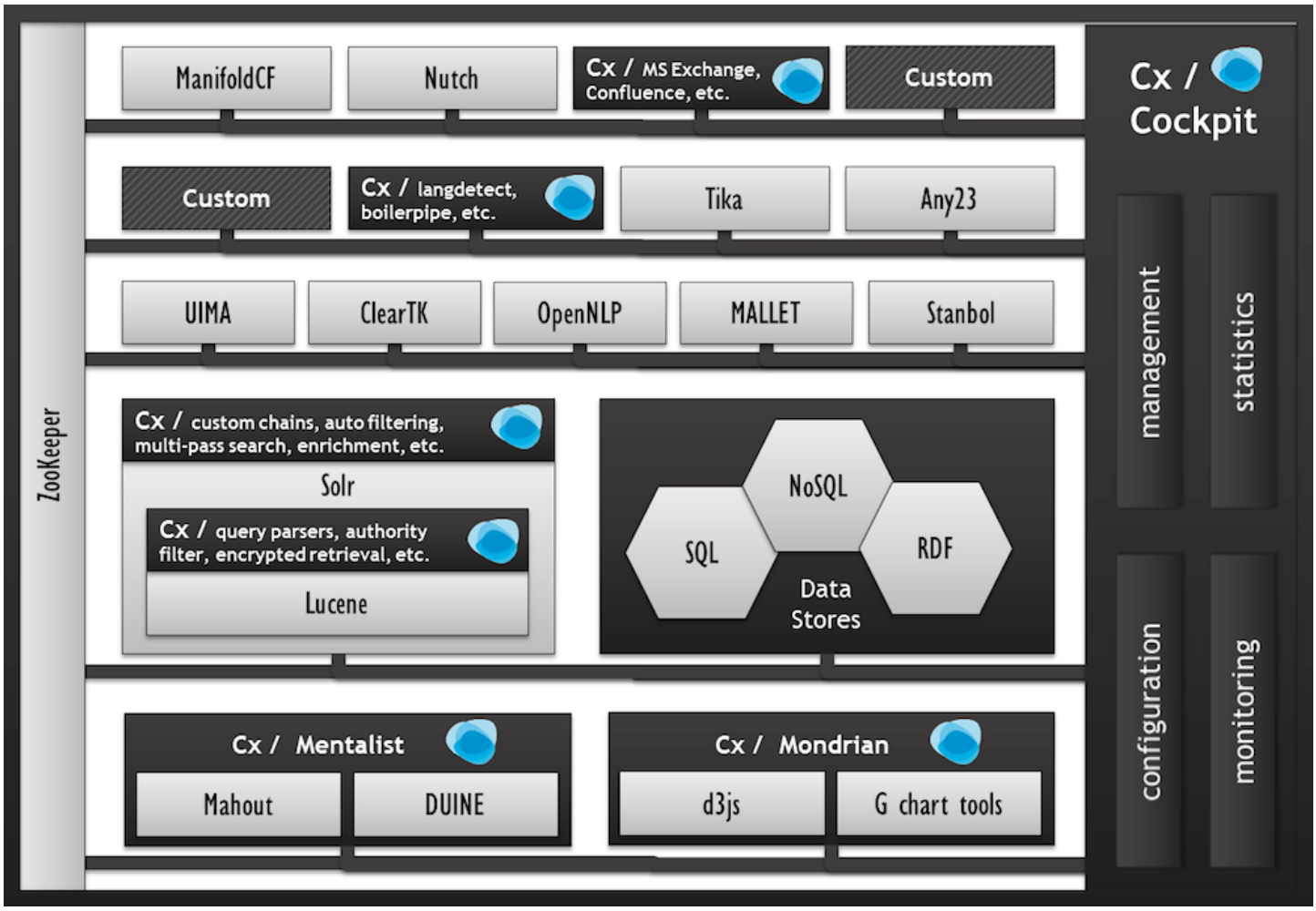
Apache Mahout is a highly scalable machine learning library providing algorithms for user- and item-based recommendations, and various types of clustering and classification.

Apache Hadoop is a framework that allows for the distributed processing of large data sets across clusters of computers.

The Semanteer Platform

The Semanteer platform provides an advanced, scalable framework for collecting, extracting, enriching and searching data, with centralized, user-friendly facilities for configuring, managing and monitoring all framework components.

The Semanteer platform is built on top of best-of-breed open source components, libraries and frameworks, and can be deployed on your premises, in the cloud, or hosted by us. The modular nature of the platform allows you to pick and choose the components that best fit your needs and organizational settings.



Recommend what matters

Combine navigation-, search-, purchase-, and multi-channel data to automatically generate meaningful recommendations.

Personalized content

Tailor your content, products, and services to fit the individual needs and preferences of your users.

CONTACT US FOR MORE INFORMATION

CONTEXITY

We collect, extract, analyse, enhance and present your information

Tel: +41 78 6116542
Email: info@contexity.ch
Web: www.contexity.ch