WHITE PAPER

Enriching Chemistry Database Content with Human-Like Indexing Of Documents

APPLYING THE POWER OF A TAXONOMY WITH 450 MILLION TERMS

Taxonomies are used to index content in databases, making facts and literature discoverable. The size of some taxonomies can make this task daunting. A comprehensive taxonomy to describe chemistry uses roughly 450 million terms, making it ~450 times the size of the English language and ~150 times the size of the next nearest discipline-specific taxonomy. To ensure that Reaxys® fully facilitates discoverability in the far-reaching and complex discipline of chemistry, two processes occur alongside each other: the familiar manual indexing and excerption method; and a novel automatic but human-like indexing process.

ELSEVIER

How can the content in a large and diverse database be made highly discoverable?

The English language has over 1 million words. How many do you think the average person can recognize? The online survey TestYourVocab.com enables visitors to test the size of their English vocabulary. The results reveal interesting trends. Data on roughly 450,000 native English speakers shows that most adult survey participants have a vocabulary size of 20,000 to 35,000 words. These are words that they can recognize and define. Topping the charts were individuals who came near 40,000 words. While these figures do not represent the average English-speaking population—it takes a certain type of person to respond to such surveys—and may be biased by the data collection methodology, the survey results hint at an upper limit to the word and concept retention capacity of the human brain.

Some jobs rely on this capacity. An indexer's job, for instance, is to recognize terms and concepts in a document and create tags or *annotations* that:

- Mirror the content of a document and its relevance to readers
- Describe how concepts and terms are connected in the document
- Make finding the information contained in the document easy

If the full spectrum of words that could appear in a document being indexed is roughly on par with an indexer's vocabulary, then the task is quite straightforward. The job becomes increasingly challenging as the number and range of terms an indexer might encounter in a document expands. But what if the complete spectrum of words and concepts that could appear in any one document outstrips an indexer's vocabulary by several orders of magnitude? How can an indexer accomplish their job in light of such a disparity between what a document might contain and what they can recognize?

THE AFTERMATH OF KNOWLEDGE

Though oversimplified, the above scenario illustrates the situation of structuring and then augmenting chemistry knowledge in a repository like Reaxys. The core competence of Reaxys is the high-quality, high-value granular data on substances and their properties and reactions. These data are manually extracted by domain experts and then organized for rapid retrieval (high fact discoverability) and immediate use (high actionability). However, Reaxys also includes a massive database of chemistry-related literature with unparalleled coverage of the application of chemistry in disciplines ranging from material and environmental sciences through engineering and geosciences to biomedicine and pharmacology. How do you make the literature in such a large and diverse database as discoverable as the facts?

Chemistry is a far-reaching and complex discipline. With approximately 110 million known unique compounds, many with several names, a comprehensive taxonomy describing chemistry knowledge—including chemical entities and concepts related to chemistry—uses roughly 450 million terms that are organized into complex relationships of synonymy, homonymy, inclusion and semantics. Granted, chemists do not need to know every name of a compound. Clustering and grouping based on structure facilitates handling information about chemical entities and classes. Nevertheless, the vocabulary of chemistry dwarfs the next largest domain taxonomies (Figure 1).

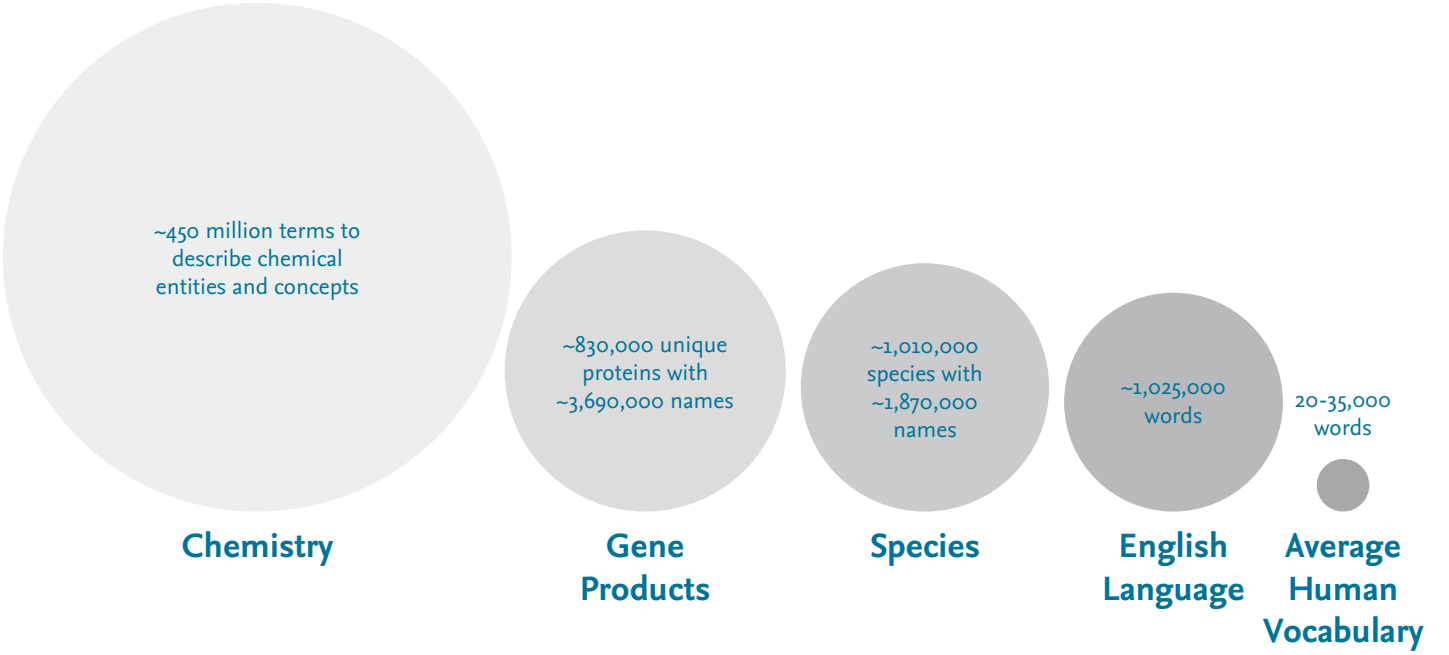


Figure 1. The number of terms used to fully capture the knowledge of chemistry is two orders of magnitude larger than the next comprehensive taxonomy describing all known gene products (450 compared to 3 million terms and synonyms). By comparison, the English language has just over 1 million words. Size of circles represents number of terms on a logarithmic scale.

Any terms of this massive vocabulary may appear in the large volume of journal titles that feed into Reaxys. In order to make those publications discoverable, each document record must include indexing keywords that reflect the full depth and breadth of chemistry concepts included in the publication, as well as the compounds discussed. Accomplishing this indexing manually requires an army of indexers with different specialties and is a very time-consuming process. In order to keep abreast with current publication rates, Reaxys has concentrated manual indexing and excerption on the 450 journals identified as having the highest-impact chemistry content.

To bring into focus the content of the remaining literature body, the Reaxys production team had to find another solution to index publications while meeting quality standards. A content enrichment production stream was constructed to leverage intelligent software and algorithms that mimic the way in which humans assess language. Using comprehensive domain-specific dictionaries, analysis engines in the stream look up word for word of the text of a document in a fraction of a second and determine the relevance of each term matched in a dictionary by assessing its relationship to other terms in the text surrounding it—in essence, by reading sentences and paragraphs the way a human does, but with the distinct advantage that computers have no vocabulary size limits.

THE TWO PRODUCTION STREAMS OF REAXYS

The content enrichment production stream is separate but complementary to the high-quality manual data excerption that produces the granular content of Reaxys. In a parallel production stream, analysis engines systematically examine hundreds of thousands of source documents per month and their output strengthens the substance database and dramatically enriches documents records in Reaxys (Figure 2).

The core, excerpted chemistry content of Reaxys comes from 450 high-impact journals and from chemistry-related patents of seven major patent offices. An additional 24,000 articles are excepted every year to generate the detailed bioactivity content of Reaxys Medicinal Chemistry. These documents enter the **manual excerption production stream**, where they are pre-annotated by a reading machine and then expertly excerpted to enhance the pre-annotation and load detailed facts about substances, reactions and properties into Reaxys.

Upwards of 16,000 auxiliary journals and patents covering an enormous breadth of chemistry-relevant disciplines feed into the literature database. Documents from roughly 12,000 of these feed into the **content enrichment production stream**. Here, document readers transform files into a common, normalized format that is then processed by a series of analysis engines called annotators. Each annotator is dedicated to a particular knowledge domain in chemistry and extracts structured knowledge from the unstructured information residing in full text.

Chemical entity recognition annotators identify compounds by name in the full text. Common names are looked up in the supporting annotation dictionaries while sophisticated algorithmic annotators translate systematic names into structures. Concept recognition annotators identify and annotate terms in text by looking up every word of the full text in annotation dictionaries constructed from ReaxysTree, the comprehensive polyhierarchical taxonomy of Reaxys. Words that are in the dictionaries are annotated according to a set of rules and these annotations translate to the standardized keywords defined by ReaxysTree.

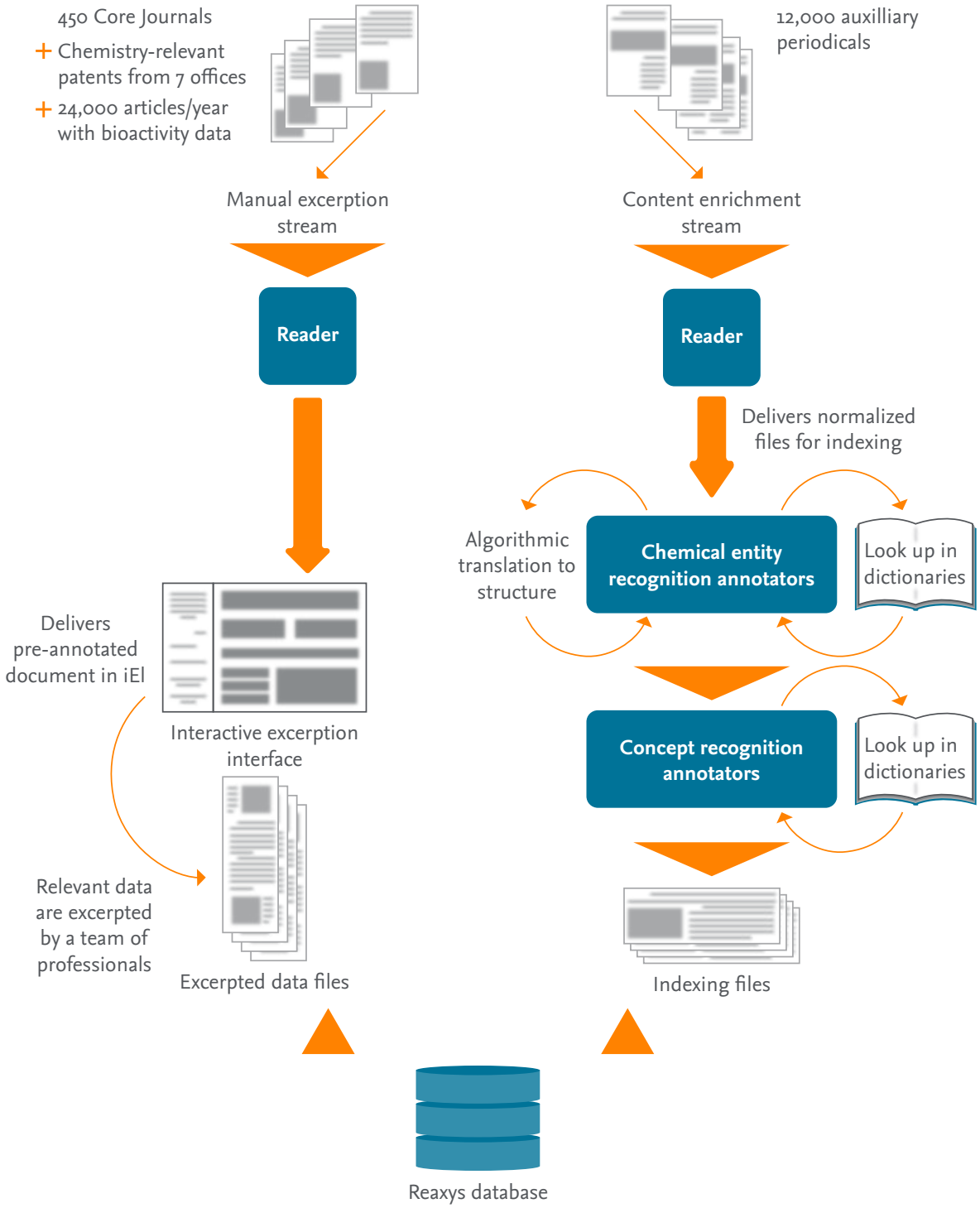


Figure 2. Two parallel production streams generate complementary input for Reaxys. One stream delivers high-quality excerpted data that feed into the highly granular database structure organizing substance properties, reaction conditions, experimental details and more. The second depplivers magnified and specific indexing that makes it faster and easier to discover and assess documents in the literature database.

As a result, information added to Reaxys from each document processed along the enrichment production stream includes not only bibliographic data, but also specific and detailed index keywords reflecting the compounds and concepts identified in the abstract, title and text body of the document. The record of this document in Reaxys is a normalized, condensed overview of the document content that is easily retrieved whenever relevant to a substance or keyword search.

The content enrichment stream complements the high-quality manual data excerption that produces the granular content of Reaxys.

CHALLENGES IN A RAPIDLY EVOLVING AND COMPLEX DISCIPLINE

The advantages of an automatic, technology-based indexing are clear. First, dictionaries supporting the annotation of documents can be as large as necessary because a computer does not have the same recall and retention limits as the human brain. These means that the heterogeneity of chemistry terminology is easily encompassed in a dictionary and a computer can, for example, very easily connect the diverse names for a compound to a single structure. Additionally, the number of indexing keywords associated with any one source document is greater, providing a more detailed and broader perspective of the content.

Second, and specific to chemical names, IUPAC nomenclature can be highly complex for novel compounds, sugars and modified peptides. The underlying rules, however, can be programmed into an algorithmic annotator that then interprets and annotates these with high precision. Third, as chemistry knowledge expands and changes, new concepts and terms can be introduced into dictionaries, so that indexing remains relevant and content is discoverable even with evolving terminology. Last but not least, it is prohibitively expensive and very difficult to find the expertise to manually index the complete disciplinary breadth of the 12,000 periodicals processed in the content enrichment production stream.

However, these advantages are not sufficient to guarantee the quality of indexing needed to preserve the value of Reaxys. The team charged with constructing the content enrichment production stream faced a number of challenges as they pieced together the right software elements to generate usable annotations. They had to find innovative but mature technology to eliminate technical barriers while ensuring consistency of processing. They had to create dictionaries to support the annotators and rules that address annotation conflicts. They had to build a framework that allows the stream to adapt over time.

These challenges arise from the complexity of chemistry as a knowledge domain and from the speed with which this domain expands and evolves. At the simplest level, speed of processing is a critical consideration. With well over 250,000 individual documents processed per month, annotators must be incredibly fast as they look up every word in the text of these documents against the annotation dictionaries. The proprietary technology used in the content enrichment stream enables a standard computer to process on average five documents per second.

Above and beyond speed, however, indexing in Reaxys must remain current with the literature and patent landscape relevant to chemistry. Thus, indexing new information must include a measure of the confidence and relevance of each annotation and must reflect up-to-date concepts and terminology.

ASSESSING CONCEPTS BASED ON SURROUNDINGS

Concept annotators scan the text of a document and compare every word to annotation dictionaries. In case of a match, an annotation is made pertaining to the recognized term. Furthermore, annotators look at words directly preceding or following the annotated term as these can significantly impact the meaning of the term. Consider, for example, the term *glucose*. On its own, a known chemical entity. Follow *glucose* with *oxidase*, however, and the meaning changes to an enzyme that breaks down glucose, which conversely has a great impact on the relevant indexing for the analyzed publication. The annotators are also designed to handle variations in term use, such as plural versions of terms, recognizing that *pentasodium tripolyphosphate* may also be written as *penta-sodium tripolyphosphate* or that *vitamins A, B and C* should be annotated as *vitamin A*, *vitamin B* and *vitamin C*.

Once a term is identified, annotators compare it to lists that determine if the annotation should be saved, discarded or altered. *Water*, for example, is a chemical which may be the topic of a research article or it may be mentioned several times because it was used as a solvent. In the first instance, water is a meaningful and context-accurate indexing term. In the latter instance, the solvent used is not necessarily relevant to the topic of the article and should not be indexed. Thus, annotators evaluate where the term *water* appears in the article. If it appears in the title or in the abstract, there is a very higher chance that *water* is a relevant keyword to index, than if it appears in the methods section.

Each annotated term is assigned a confidence and a relevance value. The confidence value reflects the certainty that the annotation is correct. This value is calculated based on a number of parameters. First, length of the term is important. Generally, the longer the term, the less likely it is that the term has a homonym. *Dieckmann condensation* is much more specific than just *condensation*, which could refer to a reaction or a physical descriptor. Second, context of the term is important. If the term appears in close proximity to words that have been annotated according to one or more of the other taxonomy facets encompassed in ReaxysTree, the confidence of a correct annotation is higher. If the term appears with others annotated according to the same taxonomy facet, then the presence of ontologically related terms (e.g., glucose oxidase with oxido-reductase and enzyme) increases the confidence value of the annotated term.

The relevance value is a ranking of how well the annotated term fits to the type of information Reaxys users want. This relevance score is determined, on one hand, by the location in the document where the term appears (terms that appear in the title and abstract have a higher relevance value) and the frequency with which it appears in the publication. On the other hand, again here ontology relationships of surrounding words contribute to the relevance value. The more children or ancestor terms in proximity of the annotated word, the higher its relevance rank. Ultimately, the confidence and relevance values determine whether a given annotation is retained and used to index the corresponding source document in Reaxys.

A CONTINUOUS LEARNING PROCESS

To remain a powerful tool, information in Reaxys must reflect the current state of knowledge in chemistry and be accessible according to preferred terminology and established concept relationships. This means that the content enrichment production stream must always use the most accurate and up-to-date dictionaries, problem-free annotation technology, and state-of-the-art document reader technology to normalize file formats of documents to be annotated.

The content enrichment product stream resides in a framework that enables regular updates to software annotators, readers and analysis systems. Thus, as new updates to document readers enable normalizing more of the highly diverse file formats used by publishers, a greater number of documents can be processed by this production stream. And as ReaxysTree evolves and expands based on input from manual excerptors and content specialists at Elsevier, annotation dictionaries are augmented with new terms or variations of a term to improve the indexing in Reaxys.

With this flexible framework that allows the production stream to adapt over time, quality assessment of annotations is an essential component of the workflow. During development, a large set of complete and partial documents was created with very detailed and carefully curated annotations. These "Gold files" serves as a standard for quality control assessments at each update made to the software to ensure that annotations are accurate and complete before adding the resulting information to the Reaxys database. Currently, indexing of chemical entities by the content enrichment stream has a precision of 90%, but the content in the Reaxys database is enhanced with information from the manual excerption production stream, which has a precision of 97%. Indexing of concepts is 94% and will improve as underlying annotation dictionaries grow.

TANGIBLE IMPACT OF FULL TEXT INDEXING

The indexing content generated by the content enrichment production stream augments the value of Reaxys by bringing the expansive body of literature in the database into focus for both substance and keyword searches.

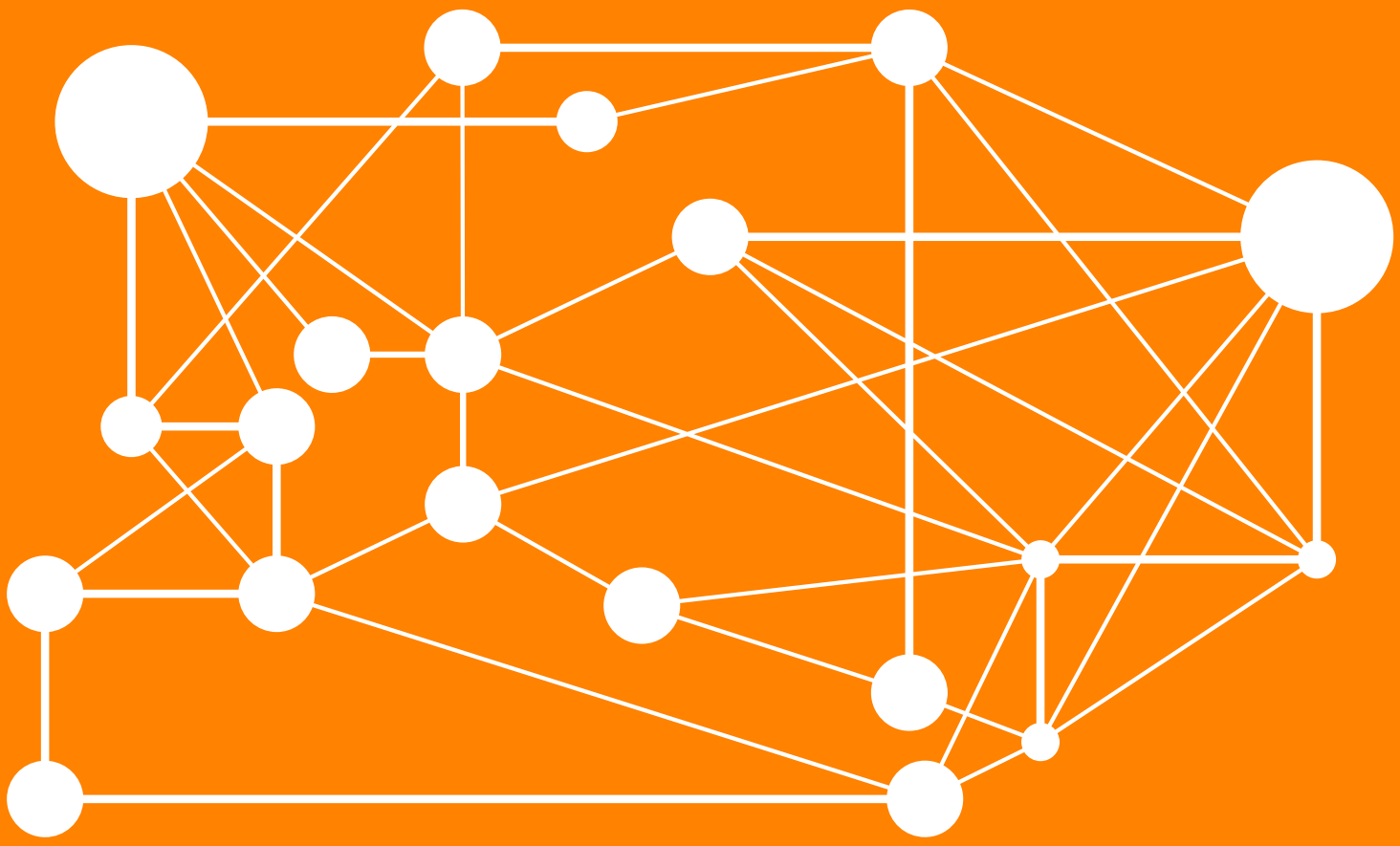
1. Wherever possible, Document records are enhanced with detailed and specific information about the actual content of the source publication, not just the keywords of the authors or those that can be extracted from the title and abstract.
2. The more granular indexing increases the visibility of any one record for a search because the keywords serve as tags to match records to query criteria. Furthermore, the depth of the indexing ensures that retrieved records are relevant to the question driving a query.
3. These index keywords also fuel more fine-grained analysis and processing of a hitset. The expanded keywords list can be the basis for filtering results, facilitate rapidly assessing the relevance of a reference to a research question, and can serve as the starting point for exploring new connections and insights.
4. The speed of processing ensures an always up-to-date and detailed perspective on the very broad landscape of chemistry-related literature.
5. The flexibility of the stream to accept and implement updates to dictionaries and software ensure that the resulting indexing is a comprehensive and in-depth description of the publication landscape that captures the full vocabulary heterogeneity of chemistry.

AT THE FOREFRONT WITH A VISION

It has taken several years to develop and implement the components of the content enrichment production stream and the framework in which it operates. Every individual element—from annotators and readers, to dictionaries and updating systems—must function correctly and fit into the overall process to guarantee the right outcome. The Reaxys content production team is tapping into computational developments as far as allowed without compromising the data quality and functional excellence at the core of Reaxys.

The potential to accomplish more is latent within the rapid advances made in artificial intelligence and the team has identified a number of technological and content areas that will receive their attention over the next years. However, the philosophy of Reaxys—to be a solution that provides answers—remains at the center of every innovation, and the question "can we continue to guarantee high quality content?" drives every development decision. The new content enrichment production stream makes information in Reaxys more discoverable and facilitates high resolution processing of results. At the same time, it allows Reaxys to keep up with the ever-growing depth and breadth of the chemistry information landscape.

And that is exactly what Reaxys is meant to do.



LEARN MORE

To request information or a product demonstration, please visit elsevier.com/reaxys or email us at reaxys@elsevier.com.

ASIA AND AUSTRALIA
Tel: +65 6349 0222
Email: sginfo@elsevier.com

JAPAN
Tel: +81 3 5561 5034
Email: jpinfo@elsevier.com

KOREA AND TAIWAN
Tel: +82 2 6714 3000
Email: krinfo.corp@elsevier.com

EUROPE, MIDDLE EAST AND AFRICA
Tel: +31 20 485 3767
Email: nlinfo@elsevier.com

NORTH AMERICA, CENTRAL AMERICA AND CANADA
Tel: +1 888 615 4500
Email: usinfo@elsevier.com

SOUTH AMERICA
Tel: +55 21 3970 9300
Email: brinfo@elsevier.com

June 2016