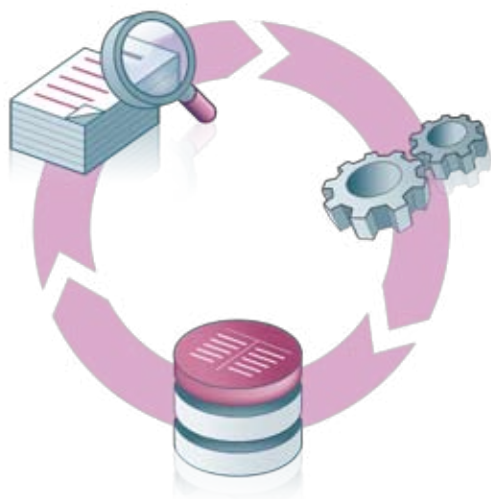


MINING FOR CHEMISTRY

The InfoChem 'Mining for Chemistry' initiative provides automated identification and extraction of chemical compounds from unstructured data such as patent documents or journal articles. The combined utilization of specialized software tools for graphical structure recognition (chemoCR™), chemical named entity extraction and name to structure conversion (ICANNOTATOR) together with leading edge chemoinformatics technologies makes our service unique and will set a new standard for information extraction from documents with chemical content.



The Concept

Valuable chemical information is often stored in scientific or patent literature as unstructured text and as images of structures/reactions. Imagine being able to abstract all this information automatically, access it via chemical structure search and have a link to the original source.

This is what InfoChem is offering.

Chemically relevant terms such as compound names, trivial or trade names, chemical fragments, or even InChI's and CAS Registry Numbers in text documents are automatically abstracted using chemical annotation software. ICANNOTATOR also assigns anchor points and highlighting information to the chemical entities found in the source document. In a second step the extracted terms are transformed into computer-readable connection tables using the name to structure conversion module ICN2S.

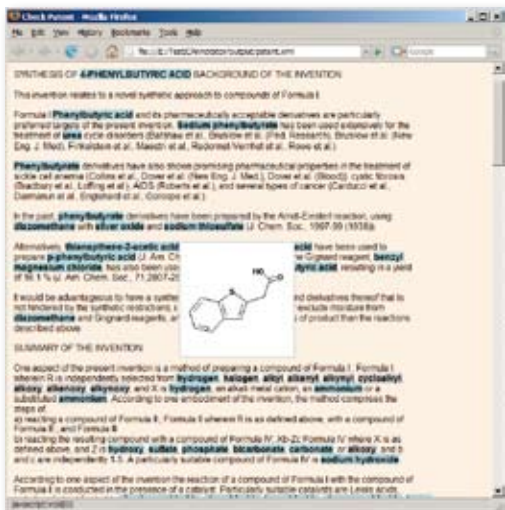
In a similar process, images containing chemical structures or reactions are recognized and converted into computer-readable formats with chemoCR™, a software tool developed by the Fraunhofer SCAI institute in collaboration with InfoChem.

Strict chemical validation of the extracted information is of utmost importance. Bad image quality, ambiguous notation or incorrect names can be the source of errors and wrong results. Consequently we apply an automatic, sophisticated chemical validation of the generated content using specific verification and checking tools.

The abstracted data undergo quality assurance and are loaded into a chemistry data cartridge such as InfoChem's ICCARTRIDGE, where they can easily be queried by structure, substructure or similarity. Combined searches of structures, facts and full text are also possible.

Chemical Named Entity Extraction and Name to Structure Conversion

For the extraction of chemical named entities, InfoChem has developed ICANNOTATOR, a powerful software tool that enables processing of English and German literature with outstanding performance and accuracy.



Name to structure conversion is performed with the InfoChem ICN2S module, a sophisticated software package that enables the automatic conversion of chemical named entities into computer-readable structures such as MOLfiles or SMILES with outstanding accuracy. Not only are conventional IUPAC or CAS names processed, but also abbreviated, semi-systematic or trivial names of common organic compounds.

Our huge dictionary containing more than 30 million unique entries supports both processes in producing accurate results.

chemoCR™

For the automated recognition of graphical chemical information, InfoChem has cooperated with the Fraunhofer Institute for Algorithms and Scientific Computing (SCAI). Using chemoCR™, a highly customizable software package, we are able to recognize images of chemical structures and reactions. The tool also re-builds the structures and converts them into a computer-readable format such as SMILES or SDfile.

What We Offer

We offer to process your legacy data, patent specifications, internal research reports or any other undiscovered, chemically relevant repositories. In close cooperation with you, the customer, in a *Mining for Chemistry* project we will identify the relevant source data, set your specific objectives, and formulate a step-by-step plan to bring out the best from your resources. Additionally, we can evaluate and accomplish the integration of the generated content into an existing chemical information system or help you to set up a completely new infrastructure, e.g., using InfoChem software components.

We place great emphasis on high quality and strict chemical validation. By applying cutting edge cheminformatics tools for the verification of the generated data, we can help you to get optimal and consistent results from your sources.