CHEMISTRY SPECIAL ISSUE

INFORMATION RETRIEVAL FACILITY NEWSLETTER

The Science and Industry Platform for a Sustainable Innovation Cycle

Issue #2 | October 2009

EDITORIAL

*by **Stephen R. Adams**, MCLIP MRSC CChem., Information Retrieval Facility, IP Expert Committee*

Welcome to this special issue of the IRF Newsletter, dedicated to the topic of chemical information and chemical searching.

As a former searcher in the agrochemical industry, I have always felt that the chemical patent searcher had a built-in advantage over their colleagues specialising in the electrical or mechanical fields. Patent documents in chemistry are written in two languages, not one – the human language of the text, and the symbolic language of chemical structure. Since the earliest days of computer-based storage and retrieval of chemical information, it has been possible to use this second, much more precise, language for a large proportion of patentability searching, and thus avoid the shortcomings of pure text retrieval. Structure-based searching has the added advantage of being independent of the human language of the original document (English, Russian, Japanese, Korean), which is a particularly useful spin-off effect in these increasingly multi-lingual times.

Traditionally, the 'Achilles heel' of chemical information systems has always been their cost. Creating databases of codified chemical structures has required an army of skilled document analysts to extract, reformat and store the associated meta-data, a huge technical investment to maintain and distribute, and consequently high royalty fees for access to the information. Any moves which reduce the cost of creating new quality-controlled databases should be welcomed, but we should bear in mind that although new authoring methods may hold out hope for the future, they can usually only be applied to newly-created documents, and provide no solution to the enormous back-files in the patent searcher's literature, which never go out of date. Substantial barriers remain to back-converting or re-parsing complete collections, without which it will be difficult for new search tools to earn their place in the industrial searcher's canon. Nonetheless, it is clear that commercial producers of chemical databases will increasingly have to justify their cost by providing clear evidence that they can deliver high-quality retrieval – and that does not always mean simply 'more hits', but better ones.

The article on free online databases for chemical searching illustrates the problem of synonyms (or more generally, alternative ways of expressing the same idea) for the chemical name searcher. Development in systematic nomenclature have never removed the ambiguity in naming of a single chemical compound – even assuming that the patent applicant uses a systematic name at all! This variability makes it inherently difficult to be comprehensive in retrieving all records which refer to that compound, if we are solely dependent upon text-string searching. The wide range of results obtained in this short study serves to illustrate the shortcomings of such strategies.

The interview with Peter Murray-Rust outlines some of his group's contributions to open source developments in the construction of chemical databases. As Peter describes, technology has advanced such that the mechanisms for extracting chemical structural data from patents are now capable of competing against some of the work of a human document analyst. Indeed, even the premium database producers are using some automatic tools to prepare a 'first draft' of a database record, to be refined by eye. Coupled with the public internet as a distribution method of choice, it is becoming possible for the first time to create and distribute new structure-based databases at much lower costs, or even free of charge.

Attendees at the first IRF Symposium may recall the demonstration of TempRanger, a prototype tool to retrieve references to specific temperatures or ranges of temperatures from the body of a patent document. A similar challenge underlies the work of the University of Sheffield's Natural Language Processing group. Initial work onextracting and testing retrieval for numeric quantities is outlined in a short report.

Finally, chemical searchers who are used to the 'usual suspects' when it comes to patent searching will find food for thought in the survey of open source initiatives, grouped under the Blue Obelisk group. It is not always clear how developments in cheminformatics or computational chemistry could feed into retrieval of bibliographic chemical information, such as the patent literature, but the competent chemical searcher certainly needs to be aware of background developments; they can be sure that their search customers will be!

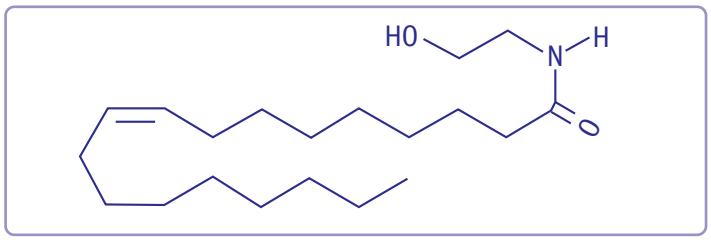
I hope that this makes for a stimulating newsletter.

Topics in this issue:

CHEMICAL PATENT SEARCHES IN FREE ONLINE DATABASES

How reliable are free online tools for an exhaustive patent search?
A small case study illustrates the shortcomings faced by chemical patent searchers.

What is a good strategy to search for chemical compounds? We compared various free online databases and search engines for N-oleoylethanolamine (OEA, CAS Registry Number 111-58-0), a very interesting compound and the Nature "lipid of the month June 2009". OEA has been found to induce satiety and decrease meal frequency, and is therefore a potential therapeutic target for treatment of obesity, diabetes and eating disorders. OEA is also used in the treatment of psoriasis, due to its ceramidase inhibiting effects.



The basic search was performed in **Espacenet**, where search was limited to title and abstracts. One patent for use of N-oleoylethanolamine (OEA) for treating psoriasis was identified, as well as 8 "simple Espacenet" patent families for use OEA (spelled as oleoylethanolamide just another name of the same compound) related to obesity.

Because the Espacenet search was carried out with title and abstracts only, no doubt, the patents from the same patent families should be present in ALL other databases (only WO or US patent families, if searched at WIPO or USPTO website).

We played with different spellings of chemical names in Google and found that variations are limited to oleoylethanolamine, oleoyl ethanolamine, oleoyl ethanolamide, and oleoylethanolamide. Search with abbreviations OEA or NOA is possible but should be restricted by IPC/USCL patent classification codes or keywords. It was possible to get through all abstracts in Espacenet with OEA, and found only one unique patent which was not retrievable using oleoylethanolamide or oleoylethanolamine.

Our next steps were:

1. Chemical names research: We took the original set of names from **Medline** and searched in the **USPTO database** with some variations. Then we made a search in **Surechem** for N-Oleoyl-ethanolamine, using SMILES (CCCCCCCCC=CCCCCCCCC(=O)NCCO, see PubChem CID: 5283454) and identified names which we had missed.
2. We extensively searched **USPTO databases** for patents and applications, and identified which names are used only in chemical context and which are used in biological context (more "chemistry-biased" names are used in biological context in US applications as well now). We compared this data with "per-term" search in **FreePatentsOnline** which used the same syntax and retrieved similar results (see table below).
3. We then identified unique biological references in US patent and patent applications (about 60 and 190 corresponding)
4. Next, we searched **FreePatentsOnline** to check if any of databases (USPTO and FreePatentsOnline) have unique references and why. We received almost identical results from both databases (FreePatentsOnline gave additional hits from cited references and provided a unique reference for [3 H] oleylethanolamide).
5. We then searched title, abstracts, claims, exported data, and excluded obviously irrelevant results (like AU3891150). The rest

		USPTO				FreePatentsOnline	
		US-patents	Type of ref	US-applications	Type of ref	US-patents	US-applications
C02	Oleoylethanolamine	26	B	89	B	27	89
C11	Oleoylethanolamide	4	B	35	B	5	35
C05	Oleoyl ethanolamine	16	B	23	B	17	23
C03	N-oleoyl-ethanolamine	8	B	13	B	14	20
C13	oleoyl ethanolamide	4	B	16	B	4	16
C12	N-oleoyl ethanolamide	2	B	13	B	2	13
C04	N-oleoyl ethanolamine	6	B	7	B	14	20
C18	N-Oleoyl-2-aminoethanol	1	B	1	B	1	1
C01	N-Oleoylethanolamine	25	B/C	72	B	26	72
C15	Oleylethanolamide	10	B/C	31	B	14	32
C16	Oleyl ethanolamide	4	B/C	2	B	5	2
C17	oleic acid ethanolamide	41	C	4	B	50	3
C07	Oleylethanolamine	6	C	2	B	7	2
C06	N-Oleylethanolamine	6	C	1	B	7	1
C09	Oleyl ethanolamine	2	C	1	B	3	0
C08	N-oleyolethanolamine	0	#N/A	1	B	0	1
C10	oleic acid ethanolamine	12	C	11	B/C	6	4
C25	N-(2-hydroxyethyl) oleylamine	5	C	3	C	5	3
C19	N-(2-hydroxyethyl) oleoylamide	1	B	0	#N/A	1	0
C24	N-(2-hydroxyethyl)oleamide	3	C	0	#N/A	6	0
C14	oleic acid ethanol amide	2	C	0	#N/A	1	0
C21	N-(2-hydroxyethyl)oleic acid amide	2	C	0	#N/A	2	0
C23	2-hydroxy-ethyl-oleic acid amide	1	C	0	#N/A	1	0
C20	2-hydroxyethyloleamide	1	C	0	#N/A	1	0
C22	.beta.-hydroxyethyl-oleylamine	1	C	0	#N/A	5	0

B- Biological application; C - Chemical application

of the patents are relevant to treatment of obesity (or special food supplements).

6. We conducted some initial searching in Boliven and Cambia's Patent Lens, and considered Google Patents without searching it. Boliven and Patent Lens both allow searching claims with complex Boolean queries. Boliven has US patents and applications, PCT applications and EP patent and applications. Patent Lens is a free patent database with focus on biomedical research, which covers only US patents and applications and EP patents. Boliven and Patent Lens retrieved large numbers of references, but more analysis would be required to determine how their relevance compared to the other sources.

The final strategy for FreePatentsOnline:

TTL/(Oleoylethanolamine OR Oleoylethanolamide OR "Oleoyl ethanolamine" OR "N-oleoyl-ethanolamine" OR "oleoyl ethanolamide" OR "N-oleoyl ethanolamide" OR "N-oleoyl ethanolamine" OR "N-Oleoyl-2-aminoethanol" OR "N-Oleoylethanolamine" OR Oleylethanolamide OR "Oleyl ethanolamide" OR "oleic acid ethanolamide" OR Oleylethanolamine OR "N-Oleylethanolamine" OR "Oleyl ethanolamine" OR "N-oleyolethanolamine" OR "oleic acid ethanolamine") OR **ABST/**(Oleoylethanolamine OR Oleoylethanolamide OR "Oleoyl ethanolamine" OR "N-oleoyl-ethanolamine" OR "oleoyl ethanolamide" OR "N-oleoyl ethanolamide" OR "N-oleoyl ethanolamine" OR "N-Oleoyl-2-aminoethanol" OR "N-Oleoylethanolamine" OR Oleylethanolamide OR "Oleyl ethanolamide" OR "oleic acid ethanolamide" OR Oleylethanolamine OR "N-Oleylethanolamine" OR "Oleyl ethanolamine" OR "N-oleyolethanolamine" OR "oleic acid ethanolamine") OR **ACLM/**(Oleoylethanolamine OR Oleoylethanolamide OR "Oleoyl ethanolamine" OR "N-oleoyl-ethanolamine" OR "oleoyl ethanolamide" OR "N-oleoyl ethanolamide" OR "N-oleoyl ethanolamine" OR "N-Oleoyl-2-aminoethanol" OR "N-Oleoylethanolamine" OR Oleylethanolamide OR "Oleyl ethanolamide" OR "oleic acid ethanolamide" OR Oleylethanolamine OR "N-Oleylethanolamine" OR "Oleyl ethanolamine" OR "N-oleyolethanolamine" OR "oleic acid ethanolamine")

This case study does not include a generic search – a mandatory procedure in a chemical patent search – using oleoylalkanolamine and like terms, or patent classification search, or "Markush" search. Therefore, patents which claim OEA-like compounds and patents describing OEA generically, as a representative of a class of chemical compounds, were out of scope of the search. Possibly missing patents may be important either for novelty, or freedom-to-operate evaluation.

Our conclusions:

- Different names are used for the same compound in patents; surprisingly, these different names are used in different context: one-word names (Oleoylethanolamine, Oleoylethanolamide and Oleylethanolamide) are preferably used by biologists or biochemists, compound names (Oleoyl ethanolamine, oleic acid ethanolamine, Oleyl ethanolamide, oleic acid ethanolamide) are preferably used by chemists. One needs to search all variations to get the complete picture.

- Compound chemical names are searchable with different syntax in different systems: for example, SPEC/"Oleyl ethanolamide" on **USPTO** and SPEC/Oleyl-ethanolamide in **Patents.com**, and one needs to know the difference to find the answer.

- **Surechem**, with its structure search capability, is missing "Oleylethanolamide" from the list of synonyms, so that name was not correctly converted to the structure and not retrievable by structure search, but still retrievable by keywords. Oleoylethanolamine and Oleoylethanolamide are retrievable by structure by one structure query.

- Obviously if one gets the chemical name correct, one can combine it with other keywords, but in this case browsing results is perfectly fine to identify relevant patents. If the correct name is not used, results will be lost.

- "Correct names" can be learned by reading non-patent publications. We made an extensive study of **PubMed** on the subject, revealing all pertinent names of the compounds used in biomedical context. They are used in patents, too.

- If all names of the compound are known and relatively simple, then **FreePatentsOnline** gives reliable data. It has a robust search interface, complete database of US patent and patent applications, and tools for exporting data. The challenge here is how to find all pertinent names. For more complicated compounds, the chemical name would be difficult to search as phrase because FreePatentsOnline does not provide tools for proximity searching. This is why chemists prefer not to search by names when possible and use specialised databases, indexed or structure databases. For this reason, FreePatentsOnline recently added structure searching with SMILES, which would make compounds searching in FreePatentsOnline more comprehensive.

- As for specific database comparisons, it is not enough to get statistics, we need to understand why we are missing patents in our searches. Some of the differences in retrieval levels could be ascribed to database coverage differences, currency of databases, using improper search syntax because of inconsistencies among search systems, and different data fields available or searched by default. These same concerns would be considered in comparative studies of commercial databases.

The main problem is how to get chemical names correct. A solution consists in looking for "SureChem"- like databases with structure search (instead of keyword search). Chemical Abstracts, which assigns unique CAS Registry numbers for any recognisable variations of chemical compounds which are indexed, is a potential solution for the time being, even though they do not have all patents with OLE listed, only the important ones.

Summary: The analysed free online tools can give a good overview, but are not reliable for an exhaustive patent search. Professional patent searchers know that one must always use multiple sources and search tools, e.g. database indexing, structure searching and free-text searching, to make searches comprehensive.

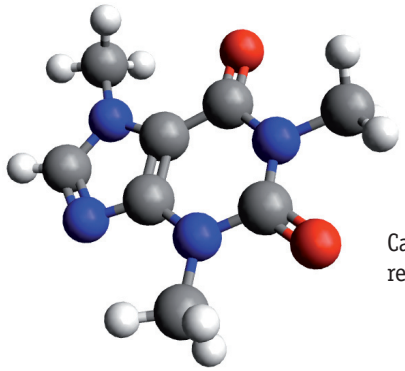
With special thanks to Aleksandr Belinskiy, whose contribution made this case study possible, and Tom Wolff for his advice.

COMMERCIAL RESTRICTIONS ARE AN IMPEDIMENT TO POWERFUL TECHNOLOGIES

An interview with Peter Murray-Rust, Reader in Molecular Informatics in the Unilever Centre at the Department of Chemistry of the University of Cambridge. Peter leads a research group that created the Chemical Markup Language and is a well-known advocate of open source and open data.

At the moment we see so many new initiatives in optical structure recognition, chemical name and reaction recognition. Which are the most promising developments?

The most important development is the rapid increase in high quality open source software. In optical structure recognition the new tool OSRA* from the National Institutes of Health has made a promising start and should increase in value and scope. The standard of optical structure presentation has increased considerably over the last 10 years because of the increased use of high quality fonts and of vectors rather than bitmaps. Although the ideal position is to have a fully semantic image, it is often possible for single structures to get complete recognition of the structure directly from a PDF document or equivalent.



Caffeine molecule, rendered in Avogadro

For chemical name recognition we have developed the tool OPSIN*, distributed with the OSCAR*-package, which promises a very high precision. OPSIN has been informally tested against a number of the commercial tools and has a lower error rate with a recall just comparable with most of them. Because the algorithms in these tools are open and classes of compounds which are not covered can be easily added, we expect that over a very short period, perhaps as little as 6 months, these tools could become emerging de-facto standards in the community.

In terms of reaction recognition, the recognition of simple reactions in structure diagrams is becoming straightforward. The quality of reaction recognition from text depends very much on the part of the document where it is found, but we have high recall and good precision for standard paragraphs in which the description of the synthesis of the compound is made.

In which areas do you expect a major development of information retrieval tools: biology, medicine, pharmaceuticals, organic/inorganic chemistry, polymer chemistry?

The application of information retrieval will be extremely important for all of the areas in the future because it is the only way of dealing with the scale of the problem. The methods will be based on the analysis of text and of diagrams. The major problem is that many of these documents are covered by copyright and that many of the publishers expressly forbid the use of machine methods to process these documents, although this would be technically possible. So, when it is possible to use machine processing of the literature without commercial and legal restriction, then there will be a major increase in the power of the technology.

The search for chemical compounds today generally involves a lot of manual work. Is it possible to obtain the same quality without any human interaction?

It depends on how cooperative the publishers are. If the publishers collaborate in making it easy to understand the documents, then we can increase the quality beyond what humans can do. But at the moment many documents are of very poor quality because the publishers do not understand or do not wish to have machines read their documents. So that many documents are scanned, they are OCR´d, rather than being created as semantic documents.

Is it harder to extract chemical information from patents than extracting chemical information from the general scientific literature?

Definitely and for several reasons. One, patents describe generic classes of science rather than specific instances and it is more difficult to understand generic concepts than specific ones. Second, patents are very often written in such a way as to make them difficult to understand whereas scientific documents are written so that they can be easily understood. The third is that the technical quality of patent documents is often much poorer than the technical quality of scientific papers, because they are OCR´d and include bitmaps.

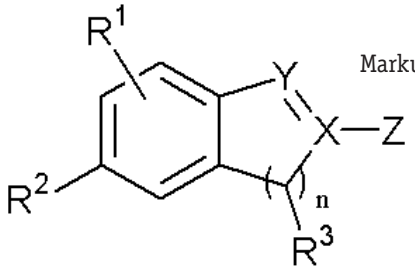
..which brings us again to the problems of authoring methods?

Exactly! Our work with Microsoft Research is worth mentioning at this point: we have developed an open source chemical authoring tool called CHEM4Word*. This tool would be available to patent offices and to the rest of the community. If the patent offices wish to create higher quality documents or to help their inventors create high quality documents, this is the type of tool required.

Because chemical information is sometimes described in a very generic way (e.g. with Markush structures), patent searchers can be confronted in their search with millions of generic structures. Do you see a solution to this problem?

We are doing research on it. It is possible with modern machines to enumerate a large number of structures described by a Markush structure and to search for them. So I think that this is partly

* See article on open source initiatives on pages 6 and 7 of this newsletter



Markush structure

solvable, simply by brute force, by putting up a large number of computers with high processing power. The second thing is that we are trying to develop a mathematical formalism for Markush structures which will allow a much more generic and therefore a cheaper way of comparing one generic structure with another. But that is a very difficult problem – we haven´t solved it yet.

Chemical patent searchers have a long list of unfulfilled wishes, like to know where a specific chemical compound is mentioned, to know if a compound is added as a reactant or as an additive, or to see a combination of chemical compounds and properties. Do you think that these wishes can be fulfilled?

First, we need to understand the structure of the patent. The more help we can get from patent experts in defining the background, the examples, the claims etc… the easier it will be to process. If the compound is in the examples, we need a section in the document which is highlighted in mark-up language, saying that these are the examples. And at the moment, this doesn´t normally happen. Also, the patent offices can do a great deal in providing more advanced documents to search. Whether a chemical compound is used as a reactant or a minor or major additive, is the sort of work that our natural language processing is addressing. For example, if two compounds are added into a preparation, we can work out whether it is a reaction or a catalyst by the language in the document. Then in terms of compounds and physical properties, it depends very much on how they are reported: it is easy to extract them from a table, but much more difficult to process from unstructured text. In good cases we can get a hundred percent and in poor cases we would be lucky to get five percent.

I am convinced that by promoting open source and open data, and at the same time developing a peer-to-peer system for publishing molecular information at source, we come closer to fulfilling our wish list.

TREC-CHEM* – EVALUATING FOR THE FUTURE

The evaluation of existing retrieval methods paves the way to the development of better technologies – TREC-CHEM provides a good example for chemical information retrieval.

Any evaluation campaign has a set of criteria that generally fall into one of two categories: **effectiveness** (does the system do what it was designed to be doing?) and **efficiency** (how fast/reliable/cheap is it?). While in principle these two categories do not conflict, in practice, because human experts have to be involved in the effectiveness category, it is hard to run one experiment that goes both sufficiently deep in the analysis to assess actual effectiveness in real user context and sufficiently large scale to give a clear image of the scalability of the different systems. This is why we divided our track into two sub-tasks.

The first sub-task of TREC-CHEM asks participating research groups to **answer 18 requests for information**. These requests have been generously provided by chemical patent experts based on their own experience. The answers are currently being evaluated manually by both students and the experts that provided them. The purpose of this task is to understand the weak points of the participating systems and specific areas where effectiveness can be improved.

The second sub-task of TREC-CHEM asks participating systems to **find relevant patents** with respect to a set of 1,000 existing patents. The results returned by the participants in this case cannot be evaluated manually, but will be assessed based on existing citations from the 1,000 patents and their family members.

The results for the first task have just finished being evaluated by students and are now being corrected by experts. For this task, 6 research groups have submitted results, using different methods of retrieval, for a total of 31 runs (a run is the application of one specific method of retrieval to the given set of documents). The results for the second sub-task are expected by September 1, 2009.

The methods applied vary substantially, **from 'basic' IR methods** (e.g. vector space models without any pre-processing of the text) to **advanced, chemistry-specific methods** using named entity recognition software and synonyms of chemical substances. It will be extremely exciting to look into the results of these methods as soon as the experts will have contributed their opinions on the results sets. Until then, partial results, based on student evaluations show that about 45% of the documents retrieved (and evaluated) have been judged relevant by the students. However, there are careful analyses to be made. For one of the topics, the relevant results were only 6%, while for another 93%. Even more, the two students that evaluated each topic did not always agree. In fact, almost 1 in 5 evaluations had conflicting results (non relevant versus relevant or highly relevant). The experts working on the topics now will have the final say in the matter, but it is interesting to understand where the disagreements arise and what can be done for a better evaluation.

The final results from the experts on the 18 manual topics as well as the results of the participating systems for the 1,000 automatically evaluated topic set will be available in the coming months and presented at the TREC event in November 2009 in Gaithersburg, MD.

* The TREC Chemistry Track (TREC-CHEM) is organised by the IRF in collaboration with University College London and York University Canada, and with the support of NIST (USA) – see more details on www.ir-facility.org.

OPEN SCIENCE, OPEN SOURCE, OPEN STANDARD, OPEN DATA IN CHEMISTRY

The open source trend which can be observed on the Web has started to spread in the Chemistry community a few years ago. Meanwhile, some interesting tools have been produced.

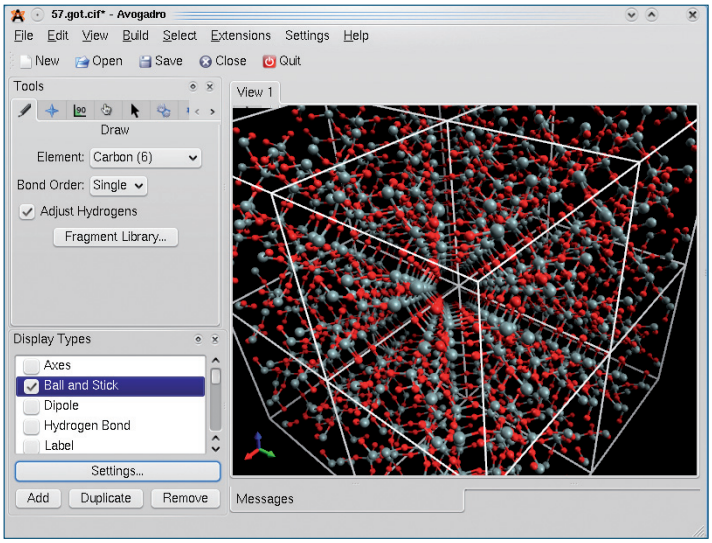
Frustrated with the closed systems that chemists currently have to work with, a group of chemists, programmers and computer scientists have met on the Internet and founded ***The Blue Obelisk Group***.

They all share a belief in the concepts of open data, open standards and open source. They express this in code, data, algorithms, specifications, tutorials, demonstrations, articles and anything that helps get the message across. They offer, for example, a collection of links to free Web services for the platform independent use of chemoinformatics programmes.

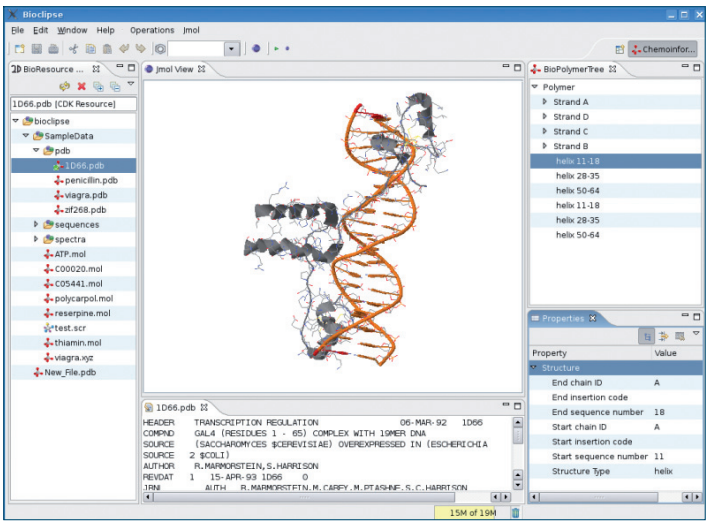
Another core Blue Obelisk project is the development of a ***shared data repository***. This repository lists many important chemoinformatics data such as elemental properties, atomic radii, etc. including references to original literature. Software developers can use this repository on online webpages or in chemistry software for free. One of the first Blue Obelisk activities was the development of an ***algorithm dictionary***. This dictionary lists many important chemoinformatics algorithms including references to original literature. Software developers can link against this list on online Webpages allowing Web search engines to find implementations of certain algorithms.

In addition, there is an increasing number of open source chemistry projects which, through the Blue Obelisk Group, maintain interoperability and promote the sharing and reuse of chemical data between projects:

Avogadro is an advanced 3D molecular editor designed for cross-platform use in computational chemistry, molecular modeling, bioinformatics, materials science, and related areas. It offers a flexible rendering engine and a powerful plugin architecture.



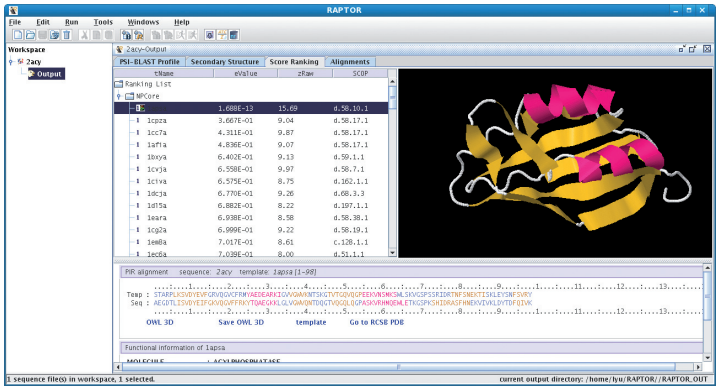
Bioclipse is a Java-based visual platform for chemo- and bioinformatics with a plugin architecture that currently includes plug-ins for the CDK and Jmol.



cclib (computational chemistry library) allows users to easily implement computational chemistry algorithms that use the results from calculations from any of a large number of popular computational chemistry packages (incl. GAMESS, GAMESS-UK, Jaguar, Gaussian, Molpro and ADF).

The ***Chemistry Development Kit (CDK)*** is a Java library for structural chemo- and bioinformatics. It is now developed by more than 50 developers all over the world and used in more than 10 different academic as well as industrial projects worldwide.

Jmol is an open-source Java viewer for chemical structures in 3D with features for chemicals, crystals, materials and biomolecules



Kalzium is an application which shows some information about the periodic system of the elements. It can be used as an information database.

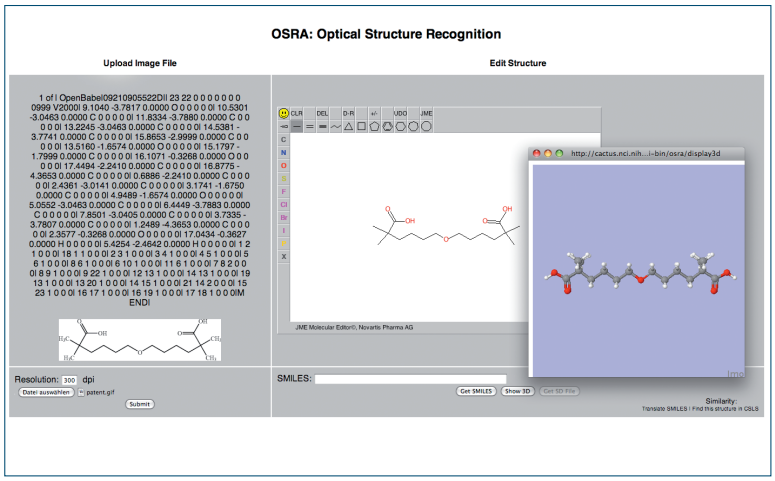
The ***NMRShiftDB*** server is open source software which can be used to maintain a local repository of the results of NMR experiments. This software was developed for the NMRShiftDB database, an open-source, open-access, open-submission, open-content Web

database for chemical structures and their associated nuclear magnetic resonance data.

Open Babel is a chemical toolbox designed to speak the many languages of chemical data. It is an open, collaborative project allowing anyone to search, convert, analyse, or store data from molecular modeling, chemistry, solid-state materials, biochemistry, or related areas. It provides a command-line interface (babel), a programming library (libopenbabel), as well as bindings to several languages such as Python, Perl, Ruby and Java.

The ***Murray-Rust Research Group*** is another advocate of open data and of openness in scientific communication. Based on the fact that most scientific data is lost during publication, they have launched several initiatives which contribute to building a global knowledge base:

OSRA is a utility designed to convert graphical representations of chemical structures as they appear in journal articles, patent documents, textbooks, trade magazines etc., into *SMILES* (Simplified Molecular Input Line Entry Specification) or SD file – a computer recognisable molecular structure format. To demonstrate the capabilities (and limitations) of OSRA the following Web interface has been created: http://cactus.nci.nih.gov/cgi-bin/osra/index.cgi.



OSCAR is a toolkit for chemical computational linguistics, chemical named entity recognition, and extraction and validation of experimental measurements from the text of journal articles. OSCAR is now seven years old and is widely used in chemistry and bioscience for the identification of chemical entities in text. Informal studies have shown it probably has the highest precision and recall of any commonly used tool. OPSIN's name2Structure has been informally tested against corpora of names, and again it is not far behind the leading commercial tool and has a smaller error rate.

CrystalEye is an automatically-extracted, highly interactive, rich repository and index of published crystallographic measurements.

SPECTRa-T is a proof-of-concept system to build a semantic data repository by text mining of chemical theses.

SPECTRa provide tools to simplify the deposition of chemistry data into repositories, in order to promote open data.

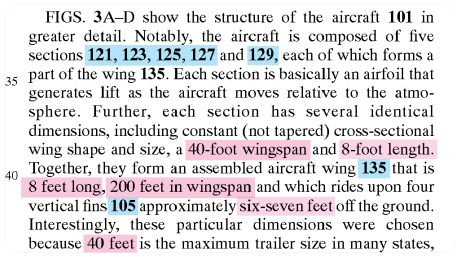
CHEM4Word is a project of Microsoft Research in partnership partnership with Dr. Peter Murray-Rust and his team at the Unilever Centre for Molecular Science Informatics to support the authoring and rendering of semantically-rich chemistry information in Word 2007 documents. The goal of the Chem4Word project is to enable similar authoring, display, and mining scenarios for chemistry-related information within Office Word.

Interesting links

> http://wiki.piug.org – PIUG blog; Free and open sources of chemical information
> http://en.wikipedia.org/wiki/Wikipedia:Chemistry – Wikipedia chemistry
> http://zusammen.metamolecular.com/2009/03/09/sixty-four-free-chemistry-databases-serialized – List of 64 free chemistry databases
> http://www.commonchemistry.org – Web resource that helps find names or CAS Registry Numbers for chemicals of general interest
> http://neurocommons.org/page/Ontologies – Ontologies including those under the OBO Foundry umbrella, including the foundational ontology BFO
> http://www.genomicglossaries.com – Genomics Glossaries & Taxonomies from Cambridge Healthtech Institute
> http://daniel.iut.univ-metz.fr/yachs/tutorials/whatsinside.php – Yet another Chemical Summarizer: an automatic text summariser specialised in Chemistry documents.
> http://www.iupac.org/inchi – InChI™ : The IUPAC International Chemical Identifier is a non-proprietary identifier for chemical substances that can be used in printed and electronic data sources thus enabling easier linking of diverse data compilations
> http://www.surechem.org – SureCHEM chemical structure searching in patent databases
> http://reccr.chem.rpi.edu/software.html – RECCR: Rensselaer Exploratory Center for Cheminformatics research - list of software
> http://www.qspr.pe.kr/my/index.php?option=com_bookmarks&Itemid=28 – Weblinks - Chem(o)informatics, Molecular Modeling: Interesting weblinks for cheminformatics and molecular modeling
> http://www.redbrick.dcu.ie/~noel/linux4chemistry – An exhaustive list of interesting links to open source, freeware, shareware and commercial software
> http://depth-first.com/articles/2007/01/24/thirty-two-free-chemistry-databases – A list of 32 free chemistry databases
> http://blueobelisk.sourceforge.net/wiki/Main_Page – See article above
> http://www.chembiogrid.org – Combination of grid computing and chemical informatics that allows convenient integration of distributed chemical tools, simulations, documents and databases

FINDING NUMBERS IN PATENTS

IP professionals have identified better search options for numeric ranges as a key priority, especially within highly complex chemical, biological, pharmaceutical and related patents that contain many references to various types of numbers, including liquid and dry measurements, temperatures, quantities, and time periods.


FIGS. 3A–D show the structure of the aircraft **101** in greater detail. Notably, the aircraft is composed of five sections **121**, **123**, **125**, **127** and **129**, each of which forms a part of the wing **135**. Each section is basically an airfoil that generates lift as the aircraft moves relative to the atmosphere. Further, each section has several identical dimensions, including constant (not tapered) cross-sectional wing shape and size, a 40-foot wingspan and 8-foot length. Together, they form an assembled aircraft wing **135** that is 8 feet long, 200 feet in wingspan and which rides upon four vertical fins **105** approximately six-seven feet off the ground. Interestingly, these particular dimensions were chosen because 40 feet is the maximum trailer size in many states,

Sifting through and sorting among all of these types of numbers requires highly sophisticated search tools that can not only distinguish between a page number and a number of pages, but also find relevant documents when a discrete value is not in the text. How, for example, would a searcher find a document relevant to the concept of "40 kilometers per second" if the document itself says "between 0 and 50 kilometers per second"?

Beside matching value ranges, another issue is one of semantic equivalence. For example, the same query above could be expressed as "40,000 meters per second" or, approximately, "90,000 mph". In all these cases, all the relevant documents need to be found, regardless of which variant was used in the original text.

This is the challenge Matrixware is working to address by funding research into numeric searches based on semantic annotation by the University of Sheffield's Natural Language Processing (NLP) Group. The NLP Group already has developed rule-based semantic annotation applications tailored to Matrixware's Alexandria patent document repository. A simplified example of such a rule, is as follows:

```
Rule: FindANumberFollowedByAUnit
(
{Number}
{MeasurementUnit}
):match
-->
:match.Measurement = {}
```

More than 30 such annotation rule sets, run sequentially from within a single bundled application, provide a basis to correctly identify and distinguish between number-related text in a patent document, and then extract that information within a relevant context.

During the summer, Matrixware experimented with the NLP Group's annotation applications using Matrixware's new MAREC (MAtrixware REsearch Collection)patent collection as a corpus. The application annotated more than 178 million measurement-related mentions in 13.5 million patents from Europe, the U.S. and Japan. The NLP Group's early work on semantic annotation for patents was presented at PaIR '08. On-going testing is refining and expanding on this work.

The enrichment of the original content through annotation is only the first half of the solution. The other half entails the development of a retrieval infrastructure that is capable of employing annotations in order to focus the search on the relevant documents, reducing the number of spurious matches. This improves search precision while keeping the recall high.

Sheffield's NLP Group, therefore, also is developing indexing and retrieval systems specifically for use with annotated text that make use of annotation semantics to identify and retrieve relevant matches regardless of how they are expressed.

The retrieval system, for example, can match measurement mentions with a related unit, such as finding kilometer values for a query based on miles. It also helps find discrete numbers implied within a range mentioned in the text. Units used in both the original document and the user's query are normalized based on the International System (SI).

These advanced retrieval functions rely on an experimental new indexing and query system being developed by the University of Sheffield and Ontotext, a semantic technology lab based in Sofia, that will be exposed through various interfaces, available through the Matrixware.net website. These capabilities might also be applied to interfaces dedicated to finding numeric ranges, or custom interfaces combining many search modalities.

If you would like further information, please contact Matrixware´s professional services team at ***MXE.ProfessionalServices@matrixware.com***

Did this newsletter meet your expectations?
Which topics would you like to read about in the next issues?
Please send your comments and suggestions to
newsletter@ir-facility.org

Imprint:
Information Retrieval Facility Society
Operngasse 20B | A-1040 Vienna | Austria
Phone: +43-1-236 94 74 | Fax +43-1-585 01 41 | www.ir-facility.org