Exposing the Hidden Web for Chemical Digital Libraries

Sascha Tönnies[1], Benjamin Köhncke[1], Oliver Koepler[2], Wolf-Tilo Balke[3]

[1] L3S Research Center, Appelstraße 9a, 30167 Hannover, Germany

[2] TIB Hannover, Welfengarten 1B, 30167 Hannover, Germany

[3] IFIS TU Braunschweig, Mühlenpfordtstraße 23, 38106 Braunschweig, Germany

{toennies, koehncke}@L3S.de, oliver.koepler@tib.uni-hannover.de, balke@ifis.cs.tu-bs.de

ABSTRACT

In recent years, the vast amount of digitally available content has lead to the creation of many topic-centered digital libraries. Also in the domain of chemistry more and more digital collections are available, but the complex query formulation still hampers their intuitive adoption. This is because information seeking in chemical documents is focused on chemical entities, for which current standard search relies on complex structures which are hard to extract from documents. Moreover, although simple keyword searches would often be sufficient, current collections simply cannot be indexed by Web search providers due to the ambiguity of chemical substance names. In this paper we present a framework for automatically generating metadata-enriched index pages for all documents in a given chemical collection. All information is then linked to the respective documents and thus provides an easy to crawl metadata repository promising to open up digital chemical libraries. Our experiments, indexing an open access journal, show that not only the documents can be found using a simple Google search via the automatically created index pages, but also that the quality of the search is much more efficient than fulltext indexing in terms of both precision/recall and performance. Finally, we compare our indexing against a classical structure search and figured out that keyword-based search can indeed solve at least some of the daily tasks in chemical workflows. To use our framework thus promises to expose a large part of the currently still hidden chemical Web, making the techniques employed interesting for chemical information providers like digital libraries and open access journals.

Categories and Subject Descriptors

H.3.1 [**INFORMATION STORAGE AND RETRIEVAL**]: Content Analysis and Indexing – *indexing methods*

H.3.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Information Search and Retrieval

H.3.7 [**INFORMATION STORAGE AND RETRIEVAL**]: Digital Libraries

General Terms

Algorithms, Experimentation, Performance.

JCDL'10, June 21–25, 2010, Gold Coast, Queensland, Australia.

Keywords

Digital libraries, information extraction, chemical digital collections, information retrieval, Web search, hidden Web.

1. INTRODUCTION

During the last years, the access to information and the information provisioning process in libraries have dramatically changed. Beside the common catalog-based searches for literature, information providers nowadays extend their services, in particular to personalizable digital portals enabling user-centered searches over heterogeneous document collections and topical databases. However, consumers have different workflows and expectations when searching for relevant literature, strongly depending on the scientific domain, the level of expertise, and the task at hand.

In the domain of chemistry information seeking is essentially centered on chemical entities. Moreover, practitioners, as well as academic researchers, are usually interested in finding *all* related documents to individual chemical entities. For both the search is basically recall-oriented because especially for synthesis procedures or production processes missing information about for instance existing patents or expected yields may lead to considerable financial losses.

The usual representation of chemical entities is based on chemical structures which are embedded (as images) into the documents. Whereas domain experts can easily identify the shown structures and classify them in the context of the document, it is currently impossible to extract this information automatically. First commercial tools like CLiDE Pro[1] or chemoCR[2] show the basic desirability. However, current recognition rates definitely do not allow for automated indexing of chemical document collections [19]. This is even more serious because the growing number of publication platforms, like open access journals or the demands of retro digitalization, calls for an automatic yet accurate way of indexing at least the documents' important chemical structures.

Actually, the problem of uniquely naming chemical structures in texts is not very new. For a long time, chemists have developed different algorithms for converting a chemical structure to unique line notations. Such a notation is, e.g., the *IUPAC name* which yields into a unique representation for small molecules (introduced around 1920). But for more complex molecules, the IUPAC rules are still ambiguous. Moreover, for the use in digital systems chemical names have been transformed into linear notations. Today, the prevalent linear notations are the *International Chemical Identifier* (InChI) and the *simplified molecular input line entry specification* (SMILES) which indeed are unique representations,

[1] www.keymodule.co.uk/CLiDE.html

[2] www.scai.fraunhofer.de/chemocr.html

but show high complexity and are almost impossible to dissect for humans. Therefore, they are not widely used in chemical documents and thus cannot be extracted for indexing purposes.

In fact, beside graphical representations, chemical documents refer to entities usually using trivial names and rely on the reader to figure out the contextual information. But also this does not help indexing: each chemical structure may have several different trivial names, often chosen with respect to the paper's context, e.g., pharmaceutical names, brand names, or terms from natural product chemistry. As always, the challenge for search engines using the entity name is to discover all related synonyms and disambiguate terms based on the document context. In particular, failing to index all entities may lead to the exclusion of highly relevant documents.

Facing these problems, chemical information service providers offer specialized indexes. These indexes are built up by *manually* identifying and indexing all chemical structures from a document collection in structure databases. The resulting structure databases then are accessed through graphical interfaces. By drawing a chemical structure a domain expert can thus formulate a query, which in turn will be parsed by the chemical query parser and matched against entities' fingerprints stored inside the structure database. The amount of manual work required for building up and maintaining such indexes results in high costs. Today, the most important provider is the Chemical Abstract Service (CAS[3]) offering high quality data at a price of about 30,000 USD/year for a single user subscription. Obviously for the growing open access movement this type of indexing documents is not a viable option.

Our aim is to make the large body of chemical knowledge stored in the Web widely searchable and accessible, however, with a minimal amount of manual indexing. Therefore, our system automatically extracts chemical entities from document collections, indexes them with synonym mark-up and disambiguation, and finally makes the documents searchable by commonly used Web search interfaces, like for instance Google or Yahoo!.

Hence, our contribution is twofold:

- Firstly, we developed an information service that automatically generates enhanced metadata representations from chemical documents. These metadata enrichments include extensive information for each entity found in the full-texts, e.g., trivial names with synonyms, InChI codes, SMILES, and basic chemical properties. By generating respective HTML pages and linking to the respective document sources, current crawlers can easily index the information in connection with each document. Our experiments clearly show the added value for chemical document retrieval.
- Secondly, by providing rich and diverse metadata our system is able to support typical, and even sophisticated chemical workflows. In contrast, previous approaches in digital libraries, like e.g., indexing entities by simple chemical formulae, see e.g., [16] are entirely useless from a chemist's point of view due to the ambiguities: for instance for the simple formula C_6H_6 there are already more than 200 different structures, each of them with different chemical properties and uses.

The rest of the paper is organized as follows: in section 2 we will give an overview of related work. A typical use case scenario for searching for chemical literature is shown in section 3 followed by a detailed description of our indexing workflow in section 4. Section 5 presents our evaluation results. Finally, we will conclude with a summary and an outlook of future work in section 6.

2. RELATED WORK

Already during the nineteenth century, inspired by the work of Jacob H. van't Hoff and August Kekulé, drawings of chemical structures became the common way of communicating chemical information about substances and their reactions. Today, we speak of chemical structure representations as the 'language of chemists' [8]. The chemical structure is a simple to understand, yet most precise way to uniquely describe a chemical entity, leaving the ambiguity of systematic, IUPAC, trivial or brand names behind. Graphical representations of chemical entities are therefore commonly used as query terms in searching for chemical information. However, although easily recognized by the human eye, graphical representations of chemical entities still cannot be easily transferred into the digital world once published in a document.

Over the last years, several projects focused on developing a chemical optical recognition for the reconstruction of chemical structure information from digitized documents. However, recognition rates always have proven to be insufficient in a production environment [12], [20], [22], [6]. That's why the most comprehensive database for chemical entities, is still manually created by the Chemical Abstracts Services (CAS) as part of the American Chemical Society. The CAS Registry, as addition to the CAS database, was already introduced in 1965 to overcome problems with identifying chemical entities based on their names. And indeed, CAS still spends a tremendous amount of funding in the manual abstracting and indexing of journal articles, conferences, patents and many other research publications in the chemical domain. For each chemical entity approximately three Euros have to be spent to fully store relevant information in the CAS registry, when extracted from literature and correctly drawn by a domain expert for a structure database. Currently CAS registry comprises over 50 millions of substances; however, access is strictly limited to subscribers.

Considering the spirit of open access journals it seems questionable to rely only on high priced commercial abstracting and indexing databases like Chemical Abstracts. Currently there are 111 chemistry journals listed in the Directory of Open Access Journals (DOAJ[4]). But opening up the knowledge of these sources to practitioners in the chemical domain requires domain specific tools for searching and (automatically) indexing information. The idea of building chemical databases poses many challenges, the most important being entity extraction, representation and matching.

The problem of entity extraction from full texts for automatic indexing is currently considered for a variety of domains. In chemistry the only open source chemical entity recognition tool currently available is the OSCAR3 framework [5], which can identify and extract multiple name variations of chemical entities. In combination with name-to-structure algorithms these entity names can be transformed into chemical structure information [18]. Of course the automated recognition of chemical entities is still dealing with the challenges of ambiguity. But, as we will see

[3] http://www.cas.org/expertise/cascontent/index.html

[4] http://www.doaj.org

later, indexing with automatically extracted phrases can already provide sufficient retrieval quality for most documents.

For the internal digital representation and exchange of structures several text-based formats have been developed. Based on the algorithms developed by Morgan [13] and Gluck [7] it is possible to store two-dimensional atom-bond structural representations of chemical entities in a tabular form, so-called connection tables. Besides, linear notations have found widespread use. The early Wiswesser line notation (WLN) [1], or the later SMILES [21], ROSDAL [3] and SYBYL line notation [2] are representations of chemical structures in the form of a linear string of alphanumeric symbols. The latest development is the InChI Code, an open standard for chemical structure description, by the IUPAC [14].

Beside exact substance matching via text strings today's databases have to store chemical structures in several other ways to enable also substructures, or similarity searches. Besides the entire chemical structure saved as a colored, undirected, cyclic graph, fragmentation codes, fragment, or substructure keys and molecular identifiers are used [4], [11]. Fragments are often stored as fingerprints coded as bit vectors. Both structure and substructure search in databases are based on graph isomorphism algorithms. Algorithms and concepts slightly differ by vendor and are mostly proprietary. Here, the general problem is that for each implemented structure database the fingerprints may severely differ. Thus, it is impossible to simply crawl the information from the Web to build up a comprehensive search index.

In current systems these efforts resulted not only in the storage and display of graphical representations of chemical entities, but also in a graphic-oriented search process. It allows a domain expert to actually draw a compound or key fragment as query input. But such specialized information retrieval interfaces are no longer limited to high priced commercial databases in a client-server environment. Recently chemical information about millions of compounds has been made available on the Web. Databases like PubChem[5], Chemspider[6], ZINC[7], ChemBank[8] or ChemDB[9] provide detailed information about some chemical structures, names and properties, also embedding graphic-oriented query interfaces for searching for chemical entities into browsers. But these platforms still require a domain specific indexing and storage of the chemical information in a structure database. A straightforward keyword-based access like provided by common search engines such as Google or Yahoo!, is still insufficiently supported for Web pages dealing with chemical information.

Recently a first few approaches trying to index digital chemical collections for keyword-style Web search have been proposed. For example, [16] and [15] both present naïve approaches to enable chemical search by indexing the empirical formulas of occurring substances. However, this type of search has very limited applications due to the ambiguity of the empirical formula itself. For instance, the simple formula C_6H_6 represents not only the well-known ring-structured compound *benzene*, but in addition more than 200 existing, but chemically entirely different structures instantiating different connectivities and topologies of 6 carbon and 6 hydrogen atoms.

The most closely related work to our approach is Harvard's QueryChem Portal[10]. It allows searching the Web based on an expanded query automatically generated from any chemical structure drawn in a graphical user interface [9]. Similar to our approach, first the chemical structure is converted into a SMILES code which in turn is used for a reference lookup in chemical Web databases like PubChem, ChemBank or Zinc. The lookup provides corresponding synonyms which are then used for a Web search via the Google API. Although such a query expansion definitely is a first step, this approach can only rely on data already correctly indexed by Google. Since most chemical documents are hidden in chemical digital libraries, they still are not retrieved, even by an expanded query. Hence the key to solve this problem lies in proper indexing.

3. USE CASE

The following scenario is typical for the daily work of a practitioner in the chemical domain. Assume our scientist is interested in the synthesis of odorous substances, e.g., as ingredients for perfumes. In particular, our chemist may be looking for building blocks usable in various synthetic pathways. Here, a simple precursor is the molecule *methoxybenzene* (see figure 1), which is a common intermediate in the production of pharmaceuticals or odorous substances. In fact, a derivate of *methoxybenzene*, *1-methoxy-4-(1-propenyl)-benzene*, is the main component of anise oil (see figure 2) which can be isolated by steam distillation from star anise (*Illicium verum*) or anise (*Pimpinella anisum*).

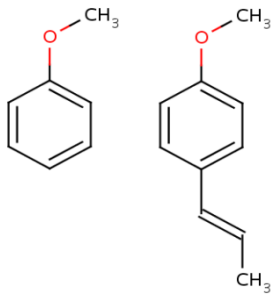

Figure 1. Methoxybenzene and 1-methoxy-4-(1-propenyl)benzene

Figure 2. Anise, from Koehler's Medicinal-Plants 1887

For the sake of open access assume that in his/her search for information our practitioner faces lacks access to commercially available chemical structure databases (due to the high prices or license limitations). Focusing on a name-based search our practitioner has to face the challenge of disambiguating chemical names (IUPAC, INN, trivial or brand name). Picking up our example entity *methoxybenzene*, one could also search for *phenoxymethane*, *phenyl methyl ether*, or even the trivial name *anisole*. All these names represent a valid verbal description of the substance.

[5] http://pubchem.ncbi.nlm.nih.gov/

[6] http://www.chemspider.com/

[7] http://zinc.docking.org/

[8] http://chembank.broadinstitute.org/

[9] http://www.chemdb.com/

[10] http://www.querychem.com

Therefore, our chemist first tries a keyword-based Web search using the query term '*methoxybenzene*', specifically on information from freely available open access journals.

For example, the ARKIVOC Journal is one of the oldest open access journals in Organic Chemistry, published since 2000, containing detailed experimental information about various compounds. But for the ARKIVOC collection a search for '*methoxybenzene*' returns zero hits. Still, only given the full texts it is impossible to distinguish whether the document collection simply does not contain any document with the entity or if our practitioner has only selected a verbal descriptor of the compound not used within the documents. In fact, a query on '*anisole*' would have retrieved 7 correct results. Thus, providing and maintaining a proper index, linking all relevant information about substances to the papers they occur in, is vital.

Moreover, keeping the risks and extremely high costs for R&D in chemical and pharmaceutical industry in mind, the index should provide rather broad information. This is because chemical searches generally demand a high recall rather than high precision: missing one important publication can compromise the whole work of a research project.

4. INDEXING HIDDEN COLLECTIONS

The basic idea of our approach is to automatically create and link enriched index pages comprising the chemical metadata for a document collection. By linking these pages to the original documents they serve as a search index over the related journal. The algorithm uses name to structure algorithms and dictionary lookups for the chemical entity recognition.

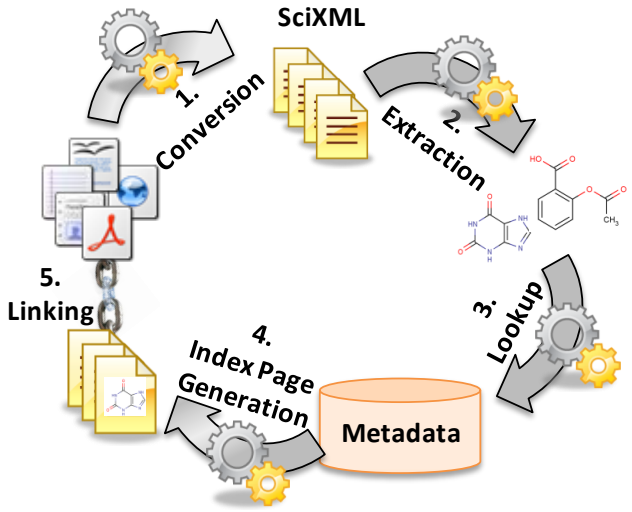

Figure 3. Workflow Overview

Our workflow (figure 3) comprises the following five steps, which are explained in more detail in the upcoming sections:

I. Convert various file formats and layouts into a single interface representation of the document.
II. Identify all referenced chemical entities by chemical entity recognition.
III. Extensive metadata enrichment for all retrieved entities.
IV. Generation of enriched index pages.
V. Linking index pages to the original sources.

4.1 Document conversion

Since open access journals show a plethora of document styles, layouts and file formats, they have to be converted into one general interface format. The standard choice is SciXML, which is a canonical XML format designed to represent the common hierarchical structure of scientific articles and is originally described in [17]. Its latest implementation, SciXML-CB, is based on an analysis of XML actually generated by scientific publishers in the fields of Chemistry and Biology [10].

Whereas it is rather trivial to convert structured document formats, e.g., XML or HTML into the respective SciXML representation, the reality is different: most open access journals have only a PDF document collection. However, PDF documents are unstructured and do not lend themselves easily to content extraction. For instance, PDF documents store all characters using the absolute position within the document and thus all paragraphs are split during OCR processes into single line paragraphs. Since entity names usually are quite long, the probability that names are split into several parts by the OCR process is rather high. Thus, entity extractors have a hard time figuring out whether different parts belong to the same entity or are entities in their own right. Imagine the chemical name *4-(aminomethyl)cyclohexamine* separated into *4-aminomethyl* and *cyclohexamine.* In addition subscript and superscript letters are important in chemical formulas and names, thus, extracting them correctly is essential. For instance, the chemical name *(1,7,7)-Trimethyl-tricyclo[2.2.1.0^{2,6}]heptan* is not a valid name without the superscript letters 2,6. As a last step, text fragments from tables and figures have to be removed.

1. */* Adjustment of algorithm parameters */*
   Given a set of PDF documents define a corresponding set of regular expressions defining layout specific parameters, e.g., position of captions and table formats.
2. */* Convert PDF documents to their respective representation in HTML.*/*
   For each document do
   2.1. Convert to HTML using *pdftohtml*; this produces a HTML file for each page. The HTML encapsulates every coherent text fragment into a <DIV> element enriched by style descriptions like font size, font family and absolute position.
   2.2. Concatenate all pages of each document to a single file.
3. */* Removing unnecessary text fragments */*
   For each HTML file do
   3.1. */* Calculate average line distance and length*/*
   Iterate over all <DIV> elements and determine the average line distance / length in paragraphs.
   3.2. */* Remove reference section **
   Identify the beginning of the reference section using the corresponding regular expression. Remove all succeeding <DIV> containers.
   3.3. */* Remove tables */*
   Identify all table captions using the corresponding regular expression. According to the general layout iterate over the succeeding (or preceding) <DIV> elements. Derive distances between each two elements using the position information. Once the distance is larger than the average calculated in 3.1 or a page break occurs, delete all <DIV> containers between the caption and the current position.

3.4. */* Remove figures */*
While figures are already removed during the OCR process, text fragments contained in certain figures (e.g. chemical reaction schemes) may still remain. Therefore identify all figure captions using the corresponding regular expression and remove all captions. Identify remaining text fragments: if the line length in any <DIV> container is shorter than the average, delete the respective element.

3.5. */* Identify abstract and keywords */*
Identify the abstract / keyword section with the corresponding regular expression. Mark up the section as abstract / keyword by adding the respective class attribute to the <DIV> element.

3.6. */* Convert SUB and SUP */*
Identify all candidates for sub- and superscript elements based on the absolute positioning and the font size. Convert the corresponding <DIV> element into a <sub> or <sup> element.

3.7. */* M*erge paragraphs **/*
Merge all remaining unclassified <DIV> elements into one single paragraph representing the document's full text.

3.8. */* Convert and save as SciXML */*
Convert the resulting HTML file into its corresponding SciXML representation using the SciXML Java Object Model[11].

Algorithm Part 1. Enhanced PDF to SciXML conversion

4.2 Chemical Entity Recognition

After the conversion of all documents into SciXML, all chemical entities contained within a document have to be recognized and annotated. In fact, the recognition of *named entities* is a major step in preprocessing and indexing not only chemical documents. Natural language processing (NLP) techniques for named entity recognition are a highly active research area. For example in the bioinformatics domain a lot of publicly available resources are already in place, e.g., the well known PubMed / Medline corpus or the manually annotated corpora generated by the PennBioIE[12] and GENIA[13] groups. In contrast, the development of NLP methodologies in the field of chemistry lags behind.

We decided to rely on the only open source project currently available on the market Oscar3 [5] which offers a range of functionalities to automatically extract chemical terms like chemical entities, reactions, concepts, and techniques. These annotations are collected in a so-called standoff annotation file (annotated SciXML) which contains pointers to the respective elements in the source text.

4. */* Entity extraction*/*
For each SciXML document do
4.1. Process all text with the Oscar3 framework. This produces an annotated SciXML file marking up chemical entities, reactions, concepts and techniques.

Algorithm Part 2. Automatic entity extraction

[11] http://www.l3s.de/vifachem/resources.html

[12] http://bioie.ldc.upenn.edu

[13] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA

4.3 Metadata Enrichment / Index Generation

The goal of our workflow is to generate enriched index pages for all documents within the collection. Thus, our next step is to collect further metadata like synonyms, SMILES and InChI for all extracted chemical entities. Generally, this information can be retrieved from topic-centered databases. The most comprehensive open access database for the area of chemistry is PubChem[14].

However, for large scale metadata generation lookups using the PubChem Web interface or Web service is far too slow. To address this problem we used the PubChem SQL dump to store all entity data in a file based hash map. By using a random access file it is now possible to directly access the relevant metadata, using the chemical name as key, without sequentially scanning the file. In fact, we measured a performance improvement of about two orders of magnitude in comparison to a Web service call: the hash map lookup needs for all kind of queries only 0.01 sec in contrast to the Web service calls needing between 1.7 and 3 sec depending on the complexity of the query. We also tried loading the PubChem dump into a relational MySQL database which, however, still resulted in around 0.2 sec response time for all queries.

5. */* Enrich chemical entity metadata.*/*
For each standoff annotation file do
5.1. Create a corresponding index page in HTML.
5.1.1. Fill the header's <TITLE> element with the journal name and paper title.
5.1.2. Adding available <META> fields out of the Dublin Core Metadata Element Set into the header container.
5.1.3. Add the paper's title within a <H1> tag.
5.1.4. Copy the paper's abstract into a paragraph.
5.1.5. Link the index page to the original URL.
5.1.6. Create an empty table for the enriched entity metadata.
5.2. For each chemical entity marked up in the standoff annotation file do
5.2.1. Use the PubChem hash map to retrieve all corresponding metadata.
5.2.2. Add a table row storing the chemical entity with all metadata.

Algorithm Part 3. Create enriched index pages

The collection of generated index pages is now ready to be used as an enriched search index over the documents collection. The beauty of our workflow is that the index pages can also be indexed and subsequently be retrieved by general purpose Web search engines, like e.g., Google or Yahoo!. We will evaluate the retrieval performance of our approach in the next section.

5. EVALUATION

For our evaluation we used a collection of 2588 chemical documents from the journal *Archive for Organic Chemistry* (ARKIVOC)[15] which is one of the most renowned open access sources for organic chemistry. This document collection has been processed by our system described in section 4 resulting in a set of enriched index pages. To assess the difference between a Web

[14] http://pubchem.ncbi.nlm.nih.gov

[15] www.arkat-usa.org

search over our semantically enriched index pages and plain full-text retrieval we used a simple Lucene whitespace analyzer to build an inverted index for the full-text documents (baseline) and the enriched index pages. For structure search the chemical entities are stored in a MySQL database in a structure table constructed by ChemAxon[16].

Basically we performed four different experiments:

- First, we evaluated the impact of our enriched index pages in terms of average result set relevance. The results of randomly chosen text queries were evaluated in a precision/recall analysis.
- To evaluate the quality in terms of ambiguity resolution we compared the retrieval results using enriched index pages to an exact structure search.
- To show the practical applicability of our approach especially over large document collections we also compared the respective retrieval times of structure and text search.
- Since our global aim is to expose chemical document collections hidden in digital libraries via commonly used Web search interfaces, like e.g., provided by Google or Yahoo!, we made our enriched index pages available online. Then we analyzed the number of pages crawled by Google and to what degree our pages are actually indexed.

5.1 Impact of Semantic Enrichment

In this experiment we evaluate the impact of our enriched index pages using a precision/recall analysis. Relevance can only be assessed manually by domain experts (in particular chemists), in what is a very expensive process. Therefore, we performed the precision/recall analysis only on a subset of documents (still about 10% of the entire collection). To choose a *representative* subset, we analyzed the number of occurrences of individual chemical entities in the document collection. Figure 4 shows the distribution of the 5000 most often occurring chemical entities.

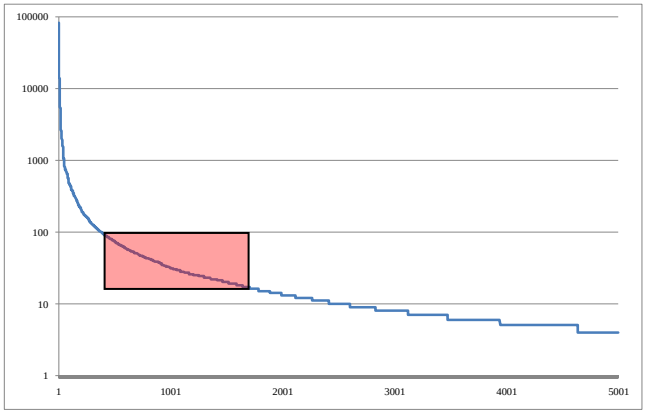

Figure 4. Distribution of entity occurrence in documents

Since it is not sensible to choose entities for evaluation that occur either in almost all documents or are extremely rare, we chose our query entities for evaluation only from entities occurring in less than 100, but more than 20 documents. We retrieved all documents matching the queries and randomly chose a subset of 10%. From these documents we randomly selected a total of 5% of the occurring entities resulting in 22 textual query terms varying from trivial entity names to InChI codes. For the evaluation domain experts in the field of chemistry considered all retrieved documents with respect to each query and judged the relevance in a binary fashion.

To determine the practical value of our textual indexing, the domain experts used a very strict relevance rating: documents are only marked as relevant, if there was an exact match for the query entity regarding both syntax *and* semantics. For example, the relevance judgment distinguished between actual substances and substance classes. Since classes are often simply given in the plural form of the respective substance this poses a difficulty for stemming in text search engines. Even worse, in some documents complex entities are described using a basic entity name as placeholder for a more detailed structure shown in some image. Since the actual structure may have totally different chemical properties also such documents have been considered as errors in the relevance analysis. Finally, sometimes an entity name can even be used as a placeholder for describing certain characteristics or functionality of other entities, i.e. although some entity name may occur in a paper, the actual entity may not be relevant. The experts also counted such documents as false retrievals in the text search.

In total from all documents retrieved as query results the domain experts marked 158 documents as relevant regarding the respective queries. Table 1 shows the resulting precision/recall values.

Table 1. Precision and relative recall values for baseline and enriched search

Search type	Retrieved	Retrieved + Relevant	Recall	Precision
Baseline	87	58	0.37	0.67
Enriched	259	150	0.95	0.58

As expected, we experienced a very low recall value of only 37% for the baseline approach. In contrast, the recall for our enriched index pages is 95%. The semantic enrichment thus yields essential benefits. For example, there will almost never be a hit in the baseline full text documents for queries on InChI codes, whereas our index pages include the InChIs and all synonyms of the query term for most of the structures. But given the strict relevance voting necessary for practical usefulness this tremendous recall benefits have to be paid for in terms of precision. Still, the precision of our approach has only slightly decreased at 58% compared to 67% for the baseline documents. Basically, due to our enrichments the result set size grows, however, this increases also the number of technically correctly found, but semantically irrelevant documents.

Table 2. F_x-Measure values for baseline and enriched search

	F_1-Measure	F_2-Measure	$F_{0.5}$-Measure
Baseline	0.47	0.40	0.57
Enriched	0.72	0.84	0.63

[16] www.chemaxon.com

To also quantify the overall benefit of our enrichment technique we computed the weighted F-Measures. Table 2 shows the different F-Measure values of the different search types. For the classic F_1-Measure we can already see a dramatic improvement of more than 0.2 over the baseline. Moreover, document retrieval in the area of chemistry is rather recall oriented: it is fatal to miss a single document related to the query. For an industrial research team missing relevant research results (e.g., with respect to patents) may lead to enormous costs for the respective company. Hence, the actually most significant measure for our scenario is the F_2-Measure weighing recall higher than precision. Here, our algorithm even scores an improvement of more than 0.4. But even when a user focuses on a precision-oriented search, our algorithm still results in a small benefit of 0.06 for the $F_{0.5}$-Measure.

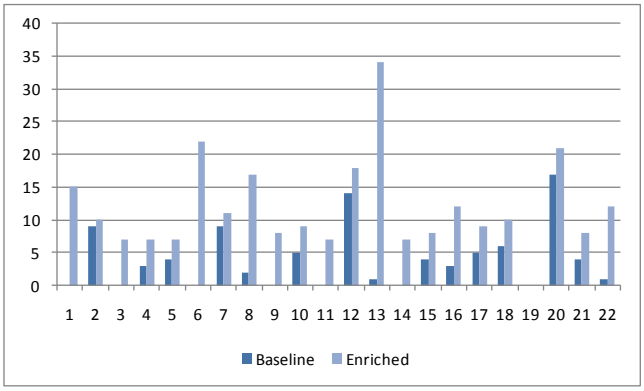

Figure 5. Retrieved documents per query: enriched versus baseline search

Investigating the search results per query more closely we found that the benefit can really be seen in all searches. Figure 5 shows a detailed overview of the number of retrieved documents per query. For all queries the enriched index pages retrieved more relevant documents than the baseline search. An exception is query 19 where no matching document was found in either approach. The respective query term *InChI=1S/C5H8O/c1-2-4-6-5-3-1/h2,4H,1,3,5H2* cannot be found because the responsible entity in the original document could not be matched uniquely to the PubChem entities. As we can see, there is still need for further improvement for metadata enrichment.

5.2 Quality of Semantic Enrichment

To measure the quality of our enriched search approach we compared the results to a chemical structure search, which currently is state of the art for chemical digital libraries. But a structure search has complex requirements: it is necessary to use specialized commercial software, e.g., ChemAxon's JChem suite, to build up a structure database. The structural data is stored in a proprietary format (varying dependent on the vendor) and also the access to the data is only possible by using appropriate graphical query interfaces where structures can be sketched.

Structure search applications offer different query types: beside an exact structure search also sub-/super-structure and similarity searches are possible. Unfortunately, these search types are not directly portable to textual searches, because e.g., substructures of an entity are not simply substrings of the entity name. Therefore, we have to focus on exact matching structures in our experiments, and leave other kinds of structure searches to future work. For each of our query terms we took the corresponding structure information of the chemical entity and retrieved all matching documents. The document and query set is the same used in section 5.1.

Table 3. Precision and relative recall values for enriched and structure search

Search type	Retrieved	Retrieved + Relevant	Recall	Precision
Enriched	259	150	0.95	0.58
Structure	262	154	0.98	0.59

Table 3 shows that the relative recall value for our enriched index pages of 95% is very similar to the respective value for the structure search. And also the precision values of 58 % for enriched and 59% for structure search are almost identical. Hence, also the F-Measures shown in Table 4 are nearly the same. Please note, that although structure search has more complex requirements, it offers only a slight advantage for exact matching queries over searching our enriched index pages.

Table 4. F_x-Measure values for enriched and structure search

	F_1-Measure	F_2-Measure	$F_{0.5}$-Measure
Enriched	0.72	0.84	0.63
Structure	0.73	0.86	0.64

Again we investigated this effect on query level. Figure 6 compares the retrieved documents for each query entity. As expected from the precision/recall analysis, in most queries enriched and structure search retrieved the same number of documents.

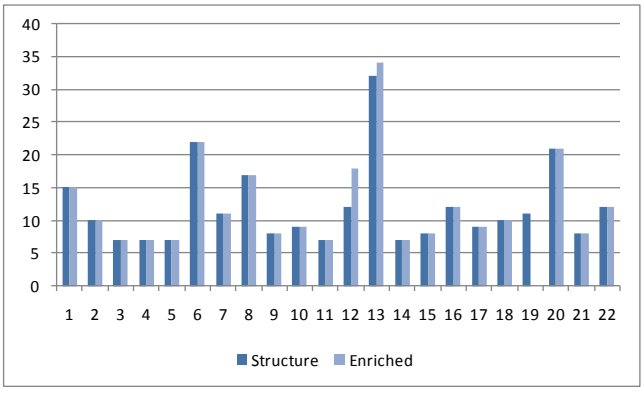

Figure 6. Retrieved documents per query: enriched versus structure search

The only exceptions occur for queries 12, 13 and 19. We already commented on the ambiguous entity term in query 19; of course a structure search can resolve this ambiguity accounting for the slightly increased recall of structure search. Moreover, for queries

12 and 13 some irrelevant documents were found in the text search, because the query entity was a substring of some more complex entity occurring in the document. For example, the query term for query 12 is *iodobenzene*. Here, also irrelevant documents containing entities like e.g. *diacetoxyiodobenzene* or *tetraiodobenzene* are retrieved. Also the abbreviated naming of entities by using their functional groups only contributes to the false retrievals.

To summarize this experiment, we can state that a text search on enriched index pages indeed yields similar results to a chemical exact structure search with respect to the retrieved documents.

5.3 Search Performance

In this experiment we compare the respective retrieval performance in terms of response times for text- and structure search. The measured time comprises query processing until all relevant documents have been retrieved. We performed experiments over several days on our digital library server to get representative average values. We did three batches, each run including 10.000 queries, varying the query terms for the text search between SMILES, names and InChI codes. The 10.000 query entities were chosen randomly from our entity database. For the structure search always the SMILES code is used which is internally converted into a unique structure representation of the respective entity. Please note, that usually also the drawing of the actual structure followed by a conversion into a SMILES code or CML would be part of the structure search. We discounted these costs by directly starting from the SMILES code. In any case, the conversion of linear notations to fingerprints is a step that has always to be performed in structure search independently of whether a SMILES code is directly given or the structure is drawn. After finding the exact matching entity for that structure all related documents are retrieved.

In text searches beside single term queries also query terms concatenated with Boolean operators are commonly used. Therefore, we simulated 'AND' and 'OR' searches. Since in structure search Boolean queries are not easy to perform, the only way here is to make two subsequent structure searches.

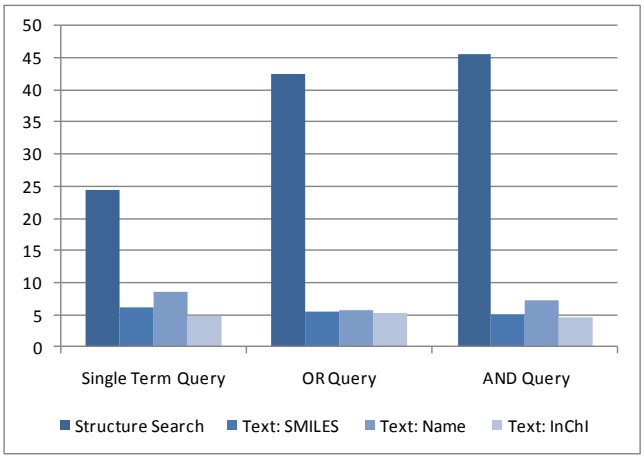

Figure 7. Retrieval times [ms] for different search types

Figure 7 shows the average retrieval times measured for the different search types. As a general trend we can see that text searches are far more efficient. For instance, in text search it makes no difference which query term is used or if more than one term is concatenated in a Boolean query. The retrieval times only vary between five and eight milliseconds (note that name search is slightly less efficient than SMILES or InChI, because of many synonyms). Using a structure search the document retrieval is always about an order of magnitude slower due to the complex matching of fingerprints. Moreover, the time for queries using Boolean operators is rather high, since here two (or more) structure searches are needed (in our experiments we only used simple queries comprising two terms).

In summary, our results show that a text search is always much faster than a structure search independently of the text search's query term. Moreover, for Boolean queries the retrieval time for text queries does not increase.

5.4 Indexing for Web Search

Our overall aim is to improve access to chemical document collections hidden in digital libraries via common Web search providers. Therefore, we simply made all enriched index pages for the ARKIVOC journal available on the Web. To have a chance of being indexed the generation and layout of our enriched pages is important. Most crawlers would mark pages within a site as spam, if they just show some index terms and do not include at least some full text or links. Therefore, our pages include, beside the actual enriched metadata table, the document's title, its abstract and a link to the full text. On the other hand, high quality open access journal will also feature high pageranks, thus crawlers will index them prominently.

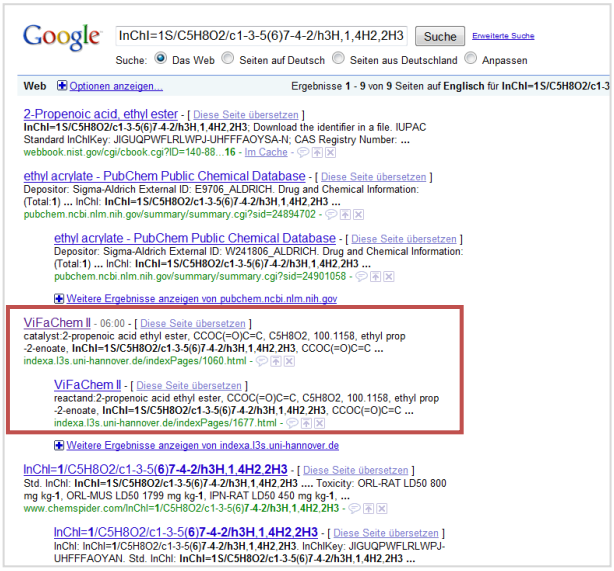

Figure 8. Google Search Example for InChI code

After three month of being online the Google index indeed contained already around 600 of our pages. However, it is not traceable how the pages are indexed and exactly why a page is indexed and some other not. Figure 8 shows a screenshot of a text search on the term '*InChI=1S/C5H8O2/c1-3-5(6)7-4-2/h3H,1,4H2,2H3'*. The enriched index page for the relevant ARKIVOC journal paper '*Effect of substituents and benzyne generating bases on the orien-*

tation to and reactivity of haloarynes' appears on third place in the Google result, directly after the respective dictionary entries of the substance from the National Institute of Standards and Technology and the PubChem substance database.

Although we did nothing to promote the index pages, i.e. our pages still have a Google pagerank of 0 (as opposed to pagerank 7 for both NIST and PubChem), they are still found and provide access to relevant documents that would not have been found otherwise (as the respective ARKIVOC journal papers do never appear in the Google search result). Please note that for investigating the indexing process we always chose '*ViFaChem II'* as title for all of our enriched pages to detect them easily in the Web search results. Of course, usually the journal name and title of the related document is used.

6. CONCLUSION AND FUTURE WORK

In recent years, the information provisioning process for topic-specific literature has essentially changed. More and more documents are directly accessible via Web search. But for the domain of chemistry the retrieval process needs to fulfill special requirements: the representation of chemical substances is always based on graphical structures. Hence, the access to document collections is usually performed by simply drawing a query entity and then performing adequate structure searches. However, on one hand the correct automatic extraction of chemical information still needs research, on the other hand the indexing is proprietary and needs specialized commercial structure databases provided by vendors, like e.g., ChemAxon. Thus, state of the art in chemical document retrieval is still annotating digital libraries with manually created and curated metadata.

The idea of our approach is to open up chemical literature hidden in digital libraries by simply enabling text queries in commonly used search interfaces, like e.g., provided by Google or Yahoo!. To facilitate this, we had to solve several problems. Chemical substances can have many different and often ambiguous textual representations, like trivial names, InChI codes or SMILES. In chemical documents besides structure images usually only trivial names are used for brevity and improved readability.

We developed a workflow allowing the automatic generation of customized index pages including all metadata information extracted from publicly accessible databases for each occurring chemical entity. Our framework can easily be used, e.g., by libraries, open access journals, or other content providers in the chemical domain. We also performed experiments to show the usefulness of our approach. The retrieval quality of our enriched index pages is almost as good as chemical exact structure searches and significantly better compared to a baseline/full text search.

However, there is still some room for improvement especially when considering the exact semantics of the substances in documents. Due to the strict relevance definition of experts in the field, it is essential to determine the exact semantics of an entity within a document. A correct text match does not always mean that a retrieved document is really relevant with respect to the semantics of the query. We already discussed the naming of substance classes as a plural form of the substance name. But also in chemical reactions substances can assume different roles and, therefore, are only more or less relevant with respect to a query. Up to a certain point, natural language techniques promise to solve this problem. We will investigate such techniques in detail in future work.

7. ACKNOWLEDGMENTS

This work was supported by the German Research Foundation (DFG) within the *ViFaChem II* project.

8. REFERENCES

[1] The Wiswesser Line-Formula Chemical Notation (WLN). Chemical Information Management, Cherry Hill, N. J., 1976.

[2] Ash, S., Cline, M., Homer, R., Hurst, T., and Smith, G. SYBYL Line Notation (SLN): A Versatile Language for Chemical Structure Representation. Journal of Chemical Information and Modeling 37, 1 (1997), 71-79.

[3] Barnard, J., Jochum, C., and Welford, S. ROSDAL: A universal structure/substructure representation for PC-host communication. Chemical Structure Information Systems: Interfaces, Communication and Standards, ACS Symposium Series No. 400, American Chemical Society (1989), 76–81.

[4] Barnard, J. Structure Representation and Searching. Ellis Horwood, Chichester, UK, 1991.

[5] Corbett, P. and Murray-Rust, P. High-throughput identification of chemistry in life science texts. Computational Life Sciences II, Springer Berlin Heidelberg (2006), 107-118.

[6] Filippov, I.V. and Nicklaus, M.C. Optical Structure Recognition Software To Recover Chemical Information: OSRA, An Open Source Solution. Journal of chemical information and modeling 49, 3 (2009), 740-3.

[7] Gluck, D.J. A Chemical Structure Storage and Search System Developed at Du Pont. Journal of Chemical Documentation 5, 1 (1965), 43-51.

[8] Hoffmann, R. and Laszlo, P. Representation in Chemistry. Angewandte Chemie International Edition in English 30, 1 (1991), 1-16.

[9] Klekota, J., Roth, F.P., and Schreiber, S.L. Query Chem: a Google-powered web search combining text and chemical structures. Bioinformatics (Oxford, England) 22, 13 (2006), 1670-3.

[10] Liakata, M., Q, C., and Soldatova, L.N. Semantic annotation of papers: interface & enrichment tool (SAPIENT). Human Language Technology Conference, (2009).

[11] Lynch, M. and Holliday, J. The Sheffield Generic Structures Project-a Retrospective Review. Journal of Chemical Information and Modeling 36, 5 (1996), 930-936.

[12] McDaniel, J.R. and Balmuth, J.R. Kekule: OCR-optical chemical (structure) recognition. Journal of Chemical Information and Modeling 32, 4 (1992), 373-378.

[13] Morgan, H.L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. Journal of Chemical Documentation 5, 2 (1965), 107-113.

[14] Stein, S.E., Heller, S.R., and Tchekhovskoi, D. An Open Standard For Chemical Structure Representation: The IUPAC Chemical Identifier. Proceedings Of The 2003 International Chemical Information Conference, Infonortics (2003), 131-143.

[15] Sun, B., Mitra, P., and Giles, C.L. Mining, indexing, and searching for textual chemical molecule information on the web. WWW, ACM (2008), 735-744.

[16] Sun, B., Tan, Q., Mitra, P., and Giles, C.L. Extraction and search of chemical formulae in text documents on the web. WWW, ACM (2007), 251-260.

[17] Teufel, S., Carletta, J., and Moens, M. An annotation scheme for discourse-level argumentation in research articles. European Chapter Meeting of the ACL, (1999).

[18] Townsend, J.A., Adams, S.E., Waudby, C.A., de Souza, V.K., Goodman, J.M., and Murray-Rust, P. Chemical documents: machine understanding and automated information extraction. Organic & Biomolecular Chemistry 2, 22 (2004), 3294--3300.

[19] Valko, A. and Johnson, P. CLiDE Pro: A chemical OCR tool. Proceedings of the 8th International Conference on Chemical Structures (ICCS), (2008).

[20] Valko, A.T. and Johnson, a.P. CLiDE Pro: the latest generation of CLiDE, a tool for optical chemical structure recognition. Journal of chemical information and modeling 49, 4 (2009)

[21] Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. Journal of Chemical Information and Modeling 28, 1 (1988)

[22] Zimmermann, M., Bui Thi, L., and Hofmann, M. Combating Illiteracy in Chemistry: Towards Computer-Based Chemical Structure Reconstruction. ERCIM News, 60 (2005), 40-41.