Building Chemical Information Systems – the ViFaChem II Project

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

[1]Forschungszentrum L3S
Universität Hannover
Appelstraße 9a
30167 Hannover
(toennies, koehncke, balke)@l3s.de

[2]Technische Informationsbibliothek

Welfengarten 1B
30167 Hannover
Oliver.Koepler@tib.uni-hannover.de

Abstract: The interdisciplinary ViFaChem II project aims at providing a chemical digital library infrastructure for creating personalized information spaces. The value added services and scientific Web 2.0 techniques actively support chemical scientists and researchers in retrieval tasks as well as in deriving new knowledge from the collected information in a highly personalized fashion. The complex requirements of a digital library for chemists are described and an overall architecture tackling these requirements is presented. Also preliminary results regarding chemical entity recognition and automatic dynamic generated document facets are presented and discussed.

1 Introduction

Today digital libraries play a major part in information provisioning. Many information providers have extended their services from the traditional catalog-based search for literature to comprehensive digital portals, like the ACM digital library in computer science, searching for information over heterogeneous document collections and databases.

However, different scientific disciplines have their own demands and the respective community has different workflows and expectations when it comes to searching for literature. Hence libraries have branched out into topically centered virtual libraries for several disciplines closely focusing on the needs of each individual science. A good example of such a portal is the chemistry portal chem.de (http://www.chem.de) and it's embedded Virtual Library of Chemistry (ViFaChem). Services of this portal include searching in bibliographic databases, chemistry databases containing comprehensive factual data about molecules and reactions, and full texts of research reports.

But for chemical literature it is not sufficient just to provide keyword-based access. Chemical information basically deals with information about molecules and their reactions. To a large degree chemical information about molecules is communicated by their structural formula instead of verbal descriptions and practitioners can very efficiently discriminate between substances based on their visual representations. For a

high quality information retrieval it is therefore to cover the information provided about chemical entities based on actual chemical workflows. In order to do that strong interdisciplinary work is mandatory.

The common text based search e.g. using *entity names* for substances cannot be easily adapted to the chemical domain. For example the standardized IUPAC name, (2S,3R,4S,5R,6R)-6-(hydroxymethyl)oxane-2,3,4,5-tetrol is rarely used for D-glucose compared to the more prominent synonyms like dextrose, corn sugar, or grape sugar. An unambiguous identifier would be the *structural formula* defining a graphical representation of a molecular structure showing atoms, bonds and their spatial arrangement. We can thus see that only interdisciplinary work will lead to a high quality information provisioning platform that is promising to be accepted by a wide range of practitioners in the field. In fact, the graphical representation of chemical structure is the natural language for the communication of chemical information.

The ViFaChem II project focuses on using knowledge about chemical workflows as a basis for creating the digital library portal. The overall vision is a personalized knowledge space for the individual practitioner in the field. Building on (automatically derived) ontologies structuring the domain, openly accessible topical databases, and specialized indexes of substances derived from a set of user-selected documents, a personalized knowledge space can be created that promises to help users combating the information flood. The rest of the paper will focus on a typical chemical workflow and show how the ViFaChem II portal addresses the problems. Since ViFaChem II is still work in progress, in this paper we will discuss the overall architecture and present examples of how the particular modules work.

2 A Use Case Scenario for Chemical Workflows

The following scenario showcases the daily tasks of a researcher in the chemical domain. Assume our scientist is interested in anti-cancer drugs, particularly the class of taxanes see e.g. [Le05], their pharmacological activities and synthesis. He may start by looking for information about Paclitaxel and related drugs. Paclitaxel (often referred to under the brand names ‘Taxol’ or ‘Abraxane’) is a terpenoid isolated from the bark of yews, with a very high activity against several tumor cell lines. Naturally our researcher is especially interested in the mode of action of Paclitaxel and maybe other compounds with similar properties. Furthermore, he is looking for experimental procedures for the synthesis of Paclitaxel-like structures or precursors.

The common information retrieval process of a text-based search as known from other domains, will fail for this scenario for many reasons: the questions of the researcher involve information about chemical entities, concepts and facts. But as stated above queries involving chemical structure information either in form of substructures or similar structures can hardly be expressed in form of keywords. Of course, searching for the chemical entity name ‘Paclitaxel’ may return some results, but as a non-proprietary name it may not be used broadly in scientific research papers. One can try the IUPAC name of Paclitaxel as generated by a large ruleset published by the IUPAC. But

especially for complex molecules there are several ways how to interpret the IUPAC guidelines for nomenclature, so one still does not have a unique identifier for the molecule.

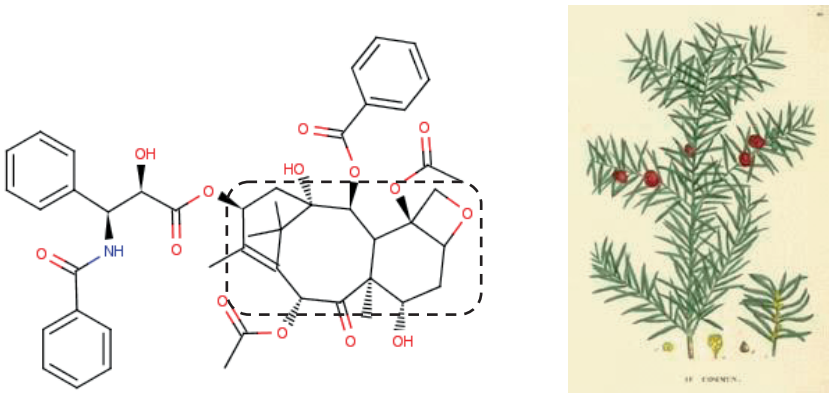

Figure 1: Structure of Paclitaxel (left) isolated from the yew tree (right)
(botanical image from: M.Grieve. 'A Modern Herbal', Harcourt, Brace & Co, 1931)

The only unambiguous representation of the entity Paclitaxel is its structural formula. Over the years several line annotations have been developed, which allow the conversion of graphical structure information into strings. Although compact strings they are not easy to handle by humans and therefore no alternative to a semantic rich drawn chemical structure. The following lines show the SMILES code for Paclitaxel.

SMILES:
CC1=C2C(C(=O)C3(C(CC4C(C3C(C(C2(C)C)(CC1OC(=O)C(C(C5=CC=CC=C5)N
C(=O)C6=CC=CC=C6)O)O)OC(=O)C7=CC=CC=C7)(CO4)OC(=O)C)O)C)OC(=O)C

Using a retrieval system with a graphical user interface our researcher can easily draw the molecular structure of Paclitaxel based on the rule sets of chemistry. Moreover, a structure based search is essential when it comes to the search for similar structures or structures which contain residues of a given lead structure.

Our scientist may find out, that taxadien-5-α-ol (cf. Highlighted C-Skeleton Fig. 1) is a central precursor in the biological synthesis of Paclitaxel. Therefore, he will search for molecules with this particular skeleton. At the latest now a text-based keyword search has become entirely useless. Such chemical structure related information retrieval can only be handled with a substructure search process. This operation is based on the information how atoms of the molecule are connected. However, the depth of structural information may vary: whereas the simplest representation contains only information about the composition of a molecule, a topological representation will also contain information about the connectivity of a molecule, showing which atoms are connected by which bond type. Moreover, the topographical representation can comprise spatial arrangements of atoms and bonds, showing the stereochemistry and conformation of a molecule, too. Therefore chemical structure databases generally contain information on

the level of topological representations. In contrast to the basic molecular formula, two-dimensional representations are the natural language of chemistry.

Thinking of integrating taxonomical information for the chemical domain, a retrieval system should also offer a navigational access over the related substance classes of the chemical entity. Thus, helpful information about structural superclasses and related substance classes is provided. Our researcher may know that Paclitaxel belongs to the taxoids, which are in turn diterpens with a taxan-like skeleton. The diterpens can be divided into acyclic, monocyclic, tri- and tetracyclic diterpens, Paclitaxel is a member of the later tetracyclic diterpens. The most prominent member of these tetracyclic diterpens is Phorbol, which interestingly is a strong carcinogen. Moreover, since our researcher is interested in drug design also information about the medical use of Paclitaxel will be used to retrieve documents and structure relevant information.

3 ViFaChem Architecture

Figure 2 gives a schematic overview of our ViFaChem II architecture. One problem when working with large document collections is the variety of file types like PDF, Word and HTML. All documents have to be processed in a specific workflow before they can be analyzed and indexed by suitable IR techniques before being stored in the ViFaChem II document repository. Section 4 will deal with the techniques used to fill the IR component of our architecture. The indexes created during the analysis are in turn offered for the personalized search functionality in the syndication step. We will discuss the various search functionalities in section 5. Finally, the user provides relevance feedback while building up a personal library from documents of the ViFaChem II collection that is then used for further personalization of the retrieval process.

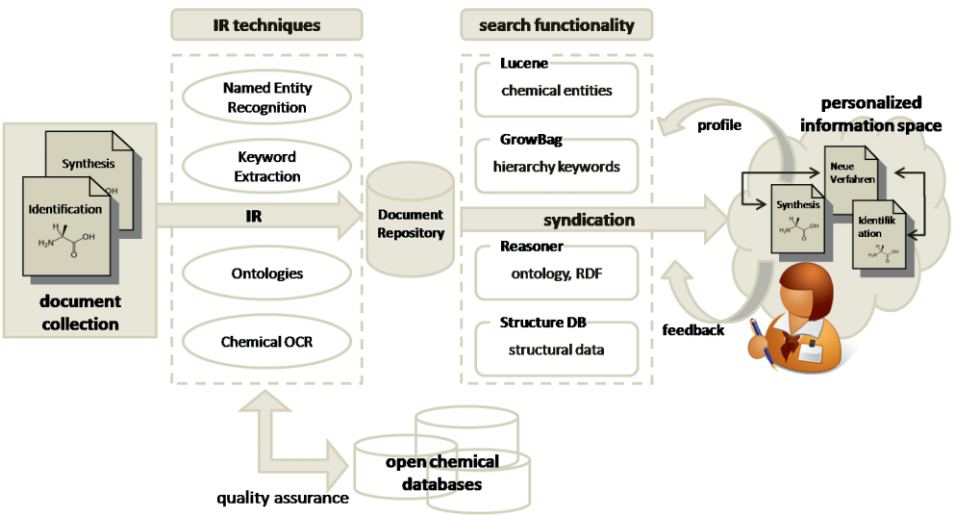

Figure 2: Basic ViFaChem II architecture

3.1 Basic Document Processing

Since we have to use many tools for deriving metadata for the use in our system, first all the different document types have to be converted into one general interface format. We rely on SciXML, which is a canonical XML format designed to represent the standard hierarchical structure of scientific articles and is originally described in [TCM99]. Its latest implementation, SciXML-CB, is based on an analysis of XML actually generated by scientific publishers in the fields of Chemistry and Biology [Ru06].

3.2 Search Engine Techniques

As described in our use case scenario researchers in chemistry are mainly interested in searching for structures. Since these structures can be quite complex, researchers often pose *graphical queries* to query a database. There is a wide collection of molecule editor software available to draw chemical structures. These programs generate a topological representation of each molecule for further processing; the data is usually handled in form of *connection tables*. Likewise a graphical query is transformed into such a format and the retrieval process takes place on the connection table representation. For the use in the ViFaChem II portal we integrated Marvin Sketch using Ajax techniques.

In addition, a *keyword based search* mechanism is needed enabling the user to make queries based on, e.g., brand names, CAS numbers, line annotations (InChI, SMILES), or bibliographic metadata. Here we used a simple inverted Lucene index.

We also introduced *facetted browsing* / *searching* as a navigational mode of access to the ViFaChem II portal. Here the Semantic GrowBag technology (c.f. 4.3) enables us to automatically generate facets from any subset (in the sense of defining particular user interest by an individual collection of documents) of our document repository. These facets can then be used for filtering the search result or just for browsing the data.

3.3 The Personalized Information Space

The actual *personal information space* relies on feedback and some organization by the user. Here it is possible to store search results together with documents, or subscribe to interesting periodicals and journals. Out of these saved documents the user profile is derived by extracting the relevant keywords from the documents. These keywords can then be used for query expansion or offered as facets for structuring the display of result sets, thus getting more relevant documents for the user. The portal is implemented with JBoss Seam using Sun JSF, RichFaces and Ajaf4JSF as main J2EE technologies.

4 IR Enabling Techniques

Chemical documents have to be extended by two types of metadata. The first type is the bibliographic metadata like authors, affiliation, publisher or year. Obviously in a library environment this metadata is readily available. The second and for our purposes more

important type is the *chemical metadata*. One can specify chemical metadata as the set of data regarding chemical entities, reactions, concepts and techniques, contained in the original document. This chemical metadata is not available out of the box and must be extracted, collected and structured. In the ViFaChem II architecture several IR techniques are used for extracting metadata, namely named entity recognition, keyword extraction, ontologies, and chemical OCR (c.f. figure 2).

4.1 Named Entity Recognition

The chemical substances considered within a document are obviously of great importance for subsequent searches. The recognition of *named entities* thus is a major step in preprocessing and indexing not only chemical documents. And indeed, natural language processing (NLP) techniques for named entity recognition in the bioinformatics domain are a highly active area. The community is assisted by a lot of publicly available resources like the well known PubMed/Medline corpus or the manually annotated PennBioIE[1] and GENIA[2] corpora. In contrast, development of NLP methodologies in the field of chemistry lags behind the biochemical world due to the lack of open access test corpora. Moreover, research in chemistry has a tradition of relying on manually edited and quite expensive commercial databases like for instance CAS Registry, the world's largest substance database with more than 15 million single- and multi-step reactions, and the respective interfaces like CAS's SciFinder. The only open source project currently available on the market is Oscar3 [CM06].

4.2 Keyword Extraction

Besides chemical metadata provided by the involved substances, also a keyword based search is necessary. Besides the typical recall-oriented full text search, here also means to aid high precision searches have to be provided. The task therefore is to tag documents with only the *highly relevant keywords* that best describe the focus of each document. For example interesting keywords may be the use of a substance for pharmaceutical purposes, or synthesis of naturally occurring substances, etc. To provide this feature all documents have to be analyzed regarding the so-called top-k keywords, i.e. the k most relevant keywords for discriminating documents with respect to a collection. Here different algorithms, e.g. TF/IDF [SG83] or Word Co-Occurrence [MI04], can be used to remove too general or often occurring phrases. After identifying the most relevant keywords for each document they are stored in an inverted file index using Apache Lucene.

4.3 Ontologies / Taxonomies

Controlled vocabularies or *taxonomies* have been found to be very useful, in particular for navigation in large result sets. The need for building up some ontology results from

[1] http://bioie.ldc.upenn.edu
[2] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA

the fact that chemical data can be represented in over 80 different file formats, but none of them, including the Chemical Markup Language [MR01], are capable of encoding knowledge in such a way that the meaning is completely preserved. Semantic Web ontologies aim to describe and relate objects using formal, logic-based representations that a machine can understand and process. So by leveraging Semantic Web technologies, it becomes possible to integrate chemical information at differing levels of details and granularity. Unlike the bio-medical domain with the MeSH taxonomy the chemical domain has access to just one single highly specialized controlled vocabulary openly available, called ChEBI [De08]. This also reflects a general problem of the chemical domain, in that the amount of data is split into a large number of sub domains making it extremely unrealistic to build a taxonomy for the entire area of chemistry.

One solution for this problem would be manually building up taxonomies for each sub domain leading to an enormous effort and needed man power. Since public libraries like the TIB have to cater to a vast variety of customers from industry and academia, however, the document collection is already highly diverse and constantly evolving. Thus the ViFaChem II portal cannot be restricted to a single taxonomy, but needs strong means of personalization without having to maintain hundreds of taxonomies and vocabularies. What is needed is an automatic way of generating some, at least light-weight, taxonomies for individual (groups of) users. Here ViFaChem II relies on the Semantic GrowBag technique [DBT07] with quite promising preliminary results.

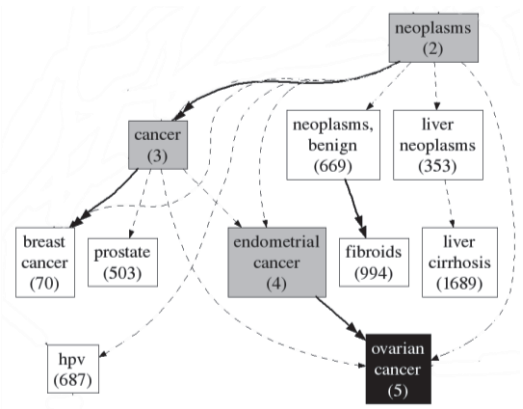

Figure 3: A sub-region of the Taxol graph generated by the Semantic GrowBag

Basically the Semantic GrowBag hierarchically organizes keywords provided in metadata annotations of digital objects based on the actual usage. It analyses the first and higher order co-occurrence of the keywords using a biased PageRank algorithm and generates a structure of nodes (keywords) and their relationships. By defining a set of resources as a starting point for the Semantic GrowBag (and thus defining their broad area of interest), users support the personalization task. Figure 3 shows a sub-region of the complete GrowBag graph automatically generated for the keyword 'Taxol' over the Medline document collection. Since Medline is focused on medical uses, keywords on the pharmaceutical action are prevalent in the GrowBag graph. Remember that Taxol is

often used in the area of cancer and especially for ovarian cancer, which indeed is a form of endometrial cancer and highly related to cancer of the breast and prostate.

Our ongoing research deals with question, whether it is possible to build up a lightweight ontology based on the GrowBag by qualifying relationships using additional domain knowledge gathered from open access data sources like PubChem. There are some examples of small chemical ontologies, e.g., [KLD08] and CO [Fe05], but up to now they still suffer from not representing enough knowledge to be useful for chemical information systems with a heterogeneous user community.

5 Search functionalities

After extracting all available metadata from the documents, this metadata is stored within a document repository which then serves as the basis of our search functionalities.

5.1 Chemical databases

Chemical structure databases are queried using a graphical interface, where a researcher can draw his/her query as a molecular structure or substructure, thus using the natural, semantically rich language of chemistry. The mouse drawn molecule query is further processed by the search engine, starting with a preliminary screening against the database keys followed by a pattern matching of the query across the returned subset of the data from the screening. Chemists differentiate between exact structure searches, substructure searches, and similarity searches. While the exact search will return only exact matches, the substructure search will return all structures including a given substructure. A similarity search will return structures based on the calculation of a similarity match, which can vary from database to database.

As already mentioned in the introduction chemical databases can either be used for retrieving chemical entities of interest or as a specialized index for scientific literature, where chemical structures represent some kind of abstract of an article. In ViFaChem II the chemical database is used as the later. The collection is structured according to an ER diagram, where the document is the central point and all other information are linked to this entity. Each entry in the document table is linked to an electronic version of the document, a link to its representation in SciXML, and a link to the annotated SciXML file. Besides, of course also the bibliographic metadata is stored in the database.

In addition we also store the chemical metadata: each chemical entity, reaction, technique and concept is stored in a related table always linked to all documents where it occurs. All this factual data and metadata can easily be stored in some arbitrary relational database system like MySQL. However, for efficiency and improved handling the structural data is stored in the specialized chemical structure database JChem Base.

5.2 Facetted Browsing and Searching

As already stated in section 4.3, taxonomies have been found very useful for navigating in large result sets, see e.g. Faceted DBLP. Hence, the ViFaChem II portal also includes a faceted search and browsing interface. The facets ViFaChem provides are adapted to the current search result. This means that there is no pre-categorization of any document relying on the expensive maintenance of a suitable category system.

When a user queries the document repository he/she will get a result set of documents related to the query. Using this result set the individual top-k keywords are calculated for personalizing the result set. Based on those top k keywords the nodes and edges of the respective pre-computed GrowBag graphs are retrieved and the most relevant keywords of the first order co-occurrence are shown as a starting point for result organization. These facets can then be selected for filtering the search results. For browsing the whole document collection the overall top k keywords are used as a basis for filtering.

5.3 Index based Search

Our bibliographic and chemical metadata, as well as the entity structures are stored in relational databases. But working with databases in the field of digital libraries usually lacks ranking features. Therefore ViFaChem II also uses another search structure. A wide spread technique is the duplication of the data in an inverted index. This works fine with text based data like the bibliographic metadata. But in the domain of chemical information systems, we also have to query the structure database which data cannot simply be put inside some inverted index. One solution for this problem would be the extension of a search engine framework such as Apache Lucene to be able to support graphical structure and substructure search. A more simple approach is using an object identifier for each structure inside the database as surrogate key connecting the structure to the actual representation of each entity. In that way the inverted index can be queried for the entity and all related documents can be retrieved ranking the result set.

6 Summary and Outlook

In this paper we have outlined the goals and discussed preliminary results and techniques of the joint interdisciplinary ViFaChem II project carried out at the Research Center L3S and the German National Library of Science and Technology. Throughout the paper we argued that chemical information systems for digital libraries with advanced content-based query methods and powerful means of personalization have to be developed in a strongly interdisciplinary fashion. ViFaChem II therefore works in tight cooperation with the Chemistry Information Centre (FIZ CHEMIE), Georg Thieme Publishers and the German Chemical Society (GdCh), analyzing and considering the specific demands and workflows for information retrieval in chemistry.

The ViFaChem II prototype offers modules for the extraction, indexing and searching of new (chemical) metadata in addition to the traditional bibliographic metadata for large

document collection. While the processing and segmentation of PDF documents is still challenging, the results with HTML and XML documents are most promising and the resulting annotated documents already offer semantically rich and –what is even more important for a library as information provisioner- correct metadata. The preliminary results of the Semantic GrowBag algorithm for the chemical domain also show interesting results for building up a chemical taxonomy, and will be further investigated with respect to automatically building lightweight ontologies.

Our future work will focus on the integration of the different modules of ViFaChem II into a single user interface to provide the semantic rich metadata not only for traditional keyword based searches, but also for navigational browsing based on taxonomies and ontologies. Combined with profile-based filter mechanisms, the ViFaChem II information retrieval process will then allow a fast and guided access to a large collection of the most relevant documents for practitioners in the field of chemical research in both industry and academia. We will also conduct user studies about the portal's usability aspects and foster the integration into the *chem.de* portal.

7 References

[CM06] Corbett, P.; Murray-Rust, P.: High-Throughput Identification of Chemistry in Life Science Texts,. In Proc. CompLife, 2006

[DBT07] Diederich, J.; Balke, W-T; Thaden, U.: Demonstrating the Semantic GrowBag: Automatically Creating Topic Facets for Faceted DBLP. In Proceedings of the ACM IEEE Joint Conference on Digital Libraries, Vancouver, BC, Canada, 2007

[De08] Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M. and Ashburner, M.: ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Res. 36, D344–D350, 2008

[Fe05] Feldman, HJ.; Dumontier, M.; Ling, S.; Hogue, C.W.: CO: A Chemical Ontology for Identification of Functional Groups and Semantic Comparison of Small Molecules. FEBS Letters. 579, 4685-4691, 2005

[KLD08] Konyk, M.; De Leon, A.; Dumontier, M.: Chemical Knowledge for the Semantic Web. In Proceedings of Data Integration in the Life Sciences (DILS2008), Evry, France

[Le05] Leistner, E.: Die Biologie der Taxane: Arzneimittel aus der Natur. In Pharmazie in unserer Zeit, 34 (2), 98-103, 2005

[MI04] Matsuo, Y.; Ishizuka, M.: Keyword extraction from a single document using word co-ocurrence statistical information. International Journal of AI Tools, 13 (1), 157-170, 2004

[MR01] Murray-Rust, P.; Rzepa, H.S.: Chemical markup, XML and the World Wide Web. 2. Information Objects and the CMLDOM, Journal of Chemical Information and Computer Sciences, 41 (5), 1113 -1123, 2001

[Ru06] Rupp, C.J.; Copestake, A.; Teufel, S.; Waldron B.: Flexible Interfaces in the Application of Language Technology to an eScience Corpus. Proc. 6th E-Science All Hands Meeting (AHM2006), Nottingham2006

[SG83] Salton, G.; McGill, MJ.: Introduction to Modern Retrieval, 1983

[TCM99] Teufel, S.; Carletta, J.; Moens, M.: An annotation scheme for discourse-level argumentation in research articles. In Proc. EACL, 1999