

# News

The Newsletter of the CDK Project

Volume 1/1, July 2004

## Editorial

by Egon Willighagen and Christoph Steinbeck

### Editorial

This is the first of, hopefully, many CDK Newsletters. The CDK community is growing and there is increasing need to exchange information about the project. This newsletter aims to inform both the user community as well as developer about the current project status, thereby supplementing CDK's JavaDoc and the developers' and users' guides. Clearly, following the dynamic development of the project as well as the sometimes crowded discussion on IRC and mailing lists, is a demanding task for someone who is not working full parttime :-)) on the project. The CDK newsletter will provide a regular update for those who lost the overview. It will contain small tutorials, describe recently published articles about or related to CDK, or about research where CDK was used, describe CDK based software (both free as well as commercial), and hopefully much more.

The intention is to publish the newsletter at least twice a year, but this will also largely depend on contribution by others. Users and developers are encouraged to submit articles of any length in which they describe the use of CDK in their own work.

The editors chose to write the CDK news in  $\text{\LaTeX}$  which is a typesetting program superior to Word :) Submitted articles must be in  $\text{\LaTeX}$  format, which sounds more complex than it actually is, as is shown in one of the articles of this newsletter. As a small exception to this, plain ASCII will be accepted too, but

is discouraged.

We would like to thank the editors of the *R News* for the approval to use their stylesheet for creating the CDK newsletter. The *R News* is the newsletter for the R project, a statistical software package, which can be found at <http://cran.r-project.org/>.

It should be clear that this new feature of the CDK project largely depends on the articles submitted by CDK users and developers. The editorial board and associated reviewers will review and edit articles with respect to spelling, grammar, etc., but not scientific novelty or merit. Research articles, of course, will better be sent to an indexed journal. Nevertheless, we expect this newsletter to be read by many people; and it is an excellent way to get your piece of code known.

We are still looking for one or two persons who would like to take place in the editorial board. We still need a reviewer, and someone who likes to write a recurrent article, e.g. an article about topics discussed on the user list.

Hopefully this has given you an idea where we plan to go with this newsletter, but this is probably best illustrated with this first issue.

Egon Willighagen

University of Nijmegen, The Netherlands

egonw@sci.kun.nl

Christoph Steinbeck

Cologne University Bioinformatics Center CUBIC, Germany

c.steinbeck@uni-koeln.de

### Contents of this issue:

Editorial . . . . .	1
NMRShiftDB . . . . .	2

What's 2004 going to bring? . . . . .	3
Literature . . . . .	4
Submitting articles to CDK News . . . . .	6
CDK ChangeLog . . . . .	7

# NMRShiftDB

A free information system for organic molecules and their spectral data

Christoph Steinbeck

## Introduction

Identification and structure elucidation of unknown natural products is an important aspect of current fields like drug discovery, metabolomics or chemical ecology. In a process known as dereplication, a scientist would record molecular fingerprint spectra and search spectral databases to check whether the compound at hand is already known (Figure 1). Only if this search is unsuccessful it is reasonable to reach for one of the more sophisticated ab-initio tools for computer-assisted structure elucidation [1, 2].

The work described here aims to use free software, the easy access provided by the World Wide Web and the collaborative potential of the Open-Source movement to build a completely transparent structure-property database "NMRShiftDB" for storage and retrieval of small organic molecules and their NMR chemical shift data [3]. The software has reached a stable state and has successfully been operating over the last few months (<http://www.nmrshiftdb.org>). It is intended to grow into a general spectroscopic information system by extending it to store other types of spectroscopic data, like mass and infra red spectra.

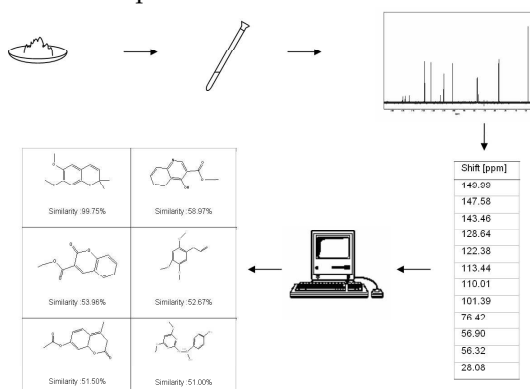


Figure 1: Dereplication as a first step in the computer assisted structure elucidation of a natural product

## Features

NMRShiftDB provides the following major functionality:

- Spectra and subspectra similarity searches
- Structure-, substructure- and structural similarity searches

- Prediction of NMR spectra based on HOSE codes and the database material
- Interface for user registration and administration (necessary for logging submissions)
- Interface for peer-reviewing submitted data for quality assurance

It also offers various other, non-spectrum related search facilities, like chemical name, formula, molecular mass etc. A simplified depiction of NMRShiftDB's datastructure is shown in Figure 2. A full entity relationship (ER) diagram can be found in [3].

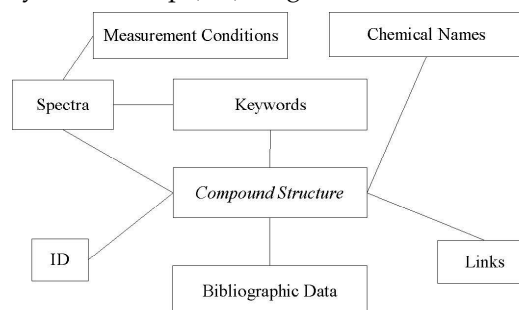


Figure 2: Simplified structure of a NMRShiftDB dataset

## Looking for Collaborations

Openness of both source code and content is a fundamental principle of the NMRShiftDB. Its software is published under GNU General Public License (GPL) [4]. Database content as well as derived data fall under Open Content License (OPL) [5]. Software and data get published regularly on <http://www.sourceforge.net> and are archived there. A replication of the database by collaborating institutions is explicitly encouraged. We are envisioning an extended mirror system to achieve a high availability of the system. Currently a system of four mirrors in three different geographic locations is at work. The mirroring system will increase availability and enable participating institutions to control responsiveness of their server directly.

Since new datasets can be added by the user community in an open submission process, NMRShiftDB needs to ensure quality of its data systematically. Each submitted dataset is subjected to an automatic quality control followed by a peer review process in order to secure a uniformly good database quality. Data in the database is also checked against itself regularly.

## Results

The first stable release of NMRShiftDB was released in November 2003 on <http://www.nmrshiftdb.org>. At the time of writing, the system's extent is characterized by 8239 structures, 8971 spectra (most of them <sup>13</sup>C spectra, some <sup>1</sup>H and a few <sup>31</sup>P) and almost 200 registered contributors. A standalone client is being developed to aid in collecting data locally and contributing them to NMRShiftDB later. This tool will allow for easier contribution and a more convenient assembly of private, specialized or in-house collections.

## Acknowledgments

We would like to thank Dr. Willy von der Lieth, Dr. Wilhelm Hull (German Cancer Research Center (Deutsches Krebsforschungszentrum, DKFZ)) and Dr. Heinz Kolshorn (University of Mainz, Germany) for generously contributing datasets from in-house databases. Thanks also go to Dr. Thomas Kämpchen, University of Marburg, for organizing the installation of the Marburg mirror of NMRShiftDB. We are also grateful to the Max-Planck-Institute of Chemical Ecology, Jena, Germany, for hosting our main server hardware, for providing the technical support, and for constantly adding data to the database. The

NMRShiftDB project is funded by the German Research Council (Deutsche Forschungsgemeinschaft, DFG)

*Dr. habil. Christoph Steinbeck*  
Cologne University Bioinformatics Center  
[c.steinbeck@uni-koeln.de](mailto:c.steinbeck@uni-koeln.de)

## Bibliography

- [1] C. Steinbeck. The automation of natural product structure elucidation. *Current Opinion in Drug Discovery and Development*, 4(3):338–342, 2001.
- [2] C. Steinbeck. Computer-assisted structure elucidation. In Johann Gasteiger, editor, *Handbook on Chemoinformatics.*, volume 2, pages 1378–1406. Wiley-VCH, Weinheim, 2003.
- [3] C. Steinbeck, S. Kuhn, and S. Krause. NMRShiftDB - Constructing a Chemical Information System with Open Source Components. *Journal of Chemical Information and Computer Sciences*, 43(6):1733 – 1739, 2003.
- [4] The Free Software Foundation. The GNU General Public License, 1991.
- [5] The Open Content Movement. The Open Content License, 1998.

# What's 2004 going to bring?

**A force field, QSAR, substructure search, and much more.**

*by Egon Willighagen*

## 2004

CDK is nearing its fourth anniversary (September 2004) and is growing faster each year: both the developer and user communities are getting larger each month, as well as the number of products based on the CDK library. This article tries to give an idea of what can be expected from the CDK project later this year.

## Force Field

One new feature that will be added this year is a force field. The group of Christoph Steinbeck has worked on a Java implementation of the MM2 force field[1]. Such force fields are used to calculate the energy of a 3D molecular structure, and in combination with an

optimization method, it can be used to optimize the 3D geometry of that molecule. It provides a faster alternative to a more accurate quantum mechanical calculation.

Independently, a CDK plugin is being written that interfaces with Ghemical [2] using a web interface. Such interoperation between CDK and other programs and libraries is expected to show up in CDK more often in the future.

## QSAR

Another field which is likely to get much more attention is quantitative structure-activity/property relationships (QSAR/QSPR). QSAR models are made by correlating molecular descriptors with activities or properties of the set of molecules being modeled. Thousands of descriptors have been proposed (see [3], and still more are proposed every day. It is expected that CDK will implement a subset of these later this year.

Very recently, a new SourceForge project has been started to address this specific field of research af-

ter a discussion on the `cdk-devel@lists.sf.net` mailing list: <http://qsar.sf.net/>. This project aims to bring together open source developers from many projects and develop a Java GUI program that interfaces with all aspects of QSAR(-like) research: setting up a data set, descriptor calculation, model building, up to model validation.

The CDK project is expected to contribute to this project by providing implementations of several of these components.

## SMARTS

Recently, the `UniversalIsomorphismTester` was adapted to allow for custom `Atom-Atom` and `Bond-Bond` matching. Prior to this change `Atoms` were matched based only on element symbol. As a result it was not possible to distinguish an  $sp^2$  and  $sp^3$  carbon. In addition, it was not possible to match an atom to any halogen. This shortcoming has been fixed now.

The next step is to write a SMARTS [4] parser and editor that can create `SMILESAtoms` that can match real atoms based on the given query. This subproject has been started recently, but the full query language is not implemented yet. A basic example has been implemented (see `cdk.test.isomorphism.SMARTSTest`). In this example the SMARTS query `'C=*` is used, thus a carbon double bonded to any atom. This is the source code that implements this:

```
SmilesParser sp = new SmilesParser();
AtomContainer atomContainer = sp.
    parseSmiles("CC(=O)OC(=O)C");
// acetic acid anhydride
QueryAtomContainer query =
    new QueryAtomContainer();
SMARTSAtom atom1 = new SMARTSAtom();
atom1.setLabel("*");
SMARTSAtom atom2 = new SMARTSAtom();
atom2.setSymbol("C");
query.addAtom(atom1);
query.addAtom(atom2);
```

## Literature

"Literature" is a recurrent column describing recently published articles that have in some way to do with CDK.

by Egon Willighagen

This column intends to give an overview of recently published articles that have some relation to CDK: they might describe algorithms implemented

```
query.addBond(
    new OrderQueryBond(atom1, atom2, 2)
);
boolean isSubstructure =
    UniversalIsomorphismTester.
    isSubgraph(atomContainer, query);
```

The `SMARTSAtom.match()` method only implements the `* atom` and much needs to be done before this fully works. Feel free to browse the source code in the `cdk.smiles.smarts` package.

## More

These three things are not the only ongoing development of CDK, but show three very interesting new features. People are encouraged to read the *CDK ChangeLog* which will appear in each issue. But here are some keywords: more reactions, tighter CML support, partial atomic charges and more CDK plugins.

Egon Willighagen  
University of Nijmegen, The Netherlands  
egonw@sci.kun.nl

## Bibliography

- [1] M.L Allinger. MM2. A Hydrocarbon Force Field Utilizing V1 and V2 Torsional Terms. *J. Am. Chem. Soc.*, 99, 1977.
- [2] Ghemical. <http://ghemical.sf.net/>, April 2004.
- [3] R. Todeschini and V. Consonni. *The Handbook of Molecular Descriptors*, volume 11 of *Methods and Principles in Medicinal Chemistry*. Wiley-VCH, Weinheim, Germany, 2000.
- [4] Daylight website. <http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html>, April 2004.

in CDK, use CDK in research, or describe software that uses CDK. Normally, this article will discuss articles published since the previous issue, but since this is the first issue, it will describe all CDK related publications that appeared so far.

The articles will be described in the order in which they appeared, but I'll take the liberty to start with the CDK article itself.

## The CDK article

Early 2003, the CDK article was published [?]. It describes the CDK project, and an important part of the architecture of the library. A must read :). Therefore, I will not further discuss it here.

## JChemPaint

JChemPaint's publication appeared in 2000 in a special issue of *Molecules* <http://www.mdpi.net/molecules/> dedicated to a virtual conference where a poster featured the program [1]. The article does not describe the CDK based JChemPaint version which was released around the same time this issue was published, but describes the older program. Note that some algorithms in CDK actually originate from this JChemPaint version, or one of its libraries.

## CML Reading

The CML reading algorithm that is used by the `cdk.io.CMLReader` was originally written for Jmol and JChemPaint and was originally published on the Chemistry Preprint Server [2], and in 2001 appeared in the *Internet Journal of Chemistry* [3]. The article describes the SAX based XML reading and describes how it deals with CML conventions.

## NMRShiftDB

Last year, another CDK-based piece of chemoinformatics software, NMRShiftDB, was released to the academic community. Technical details have been published [4]. The database is described in a separate article in this newsletter.

## Chemistry enriched RSS

An extension of Rich Site Summary (RSS) [5] with chemical information was introduced last January coined CMLRSS [6]. RSS is a system used by websites to distribute news headlines over the internet.

Currently, the headlines only contain a simple textual description, though embedding bibliographic information has been proposed by some scientific journals. CMLRSS extends RSS by embedding molecular information in CML format. The article describes a CDK plugin, the RSSViewer, which is able to download news feeds and extract chemistry (crystal/molecular structures) and display that in Jmol, JChemPaint or any other CDK plugin aware program.

Though I'll try to keep up with literature, I might oversee an interesting article that mentions, uses, extends, compares or otherwise relates to the CDK. To make sure it gets covered in this column you can send me an email with the bibliographic information for that article.

Egon Willighagen  
University of Nijmegen, The Netherlands  
egonw@sci.kun.nl

## Bibliography

- [1] Stefan Krause, Egon Willighagen, and Christoph Steinbeck. JChemPaint - Using the Collaborative Forces of the Internet to Develop a Free Editor for 2D Chemical Structures. *Molecules*, 5:93–98, 2000.
- [2] Chemical Preprint Server. <http://www.chemweb.com/preprint>.
- [3] E.L. Willighagen. Processing CML Conventions in Java. *Internet Journal of Chemistry*, 4, 2001.
- [4] C. Steinbeck, S. Kuhn, and S. Krause. NMR-ShiftDB - Constructing a Chemical Information System with Open Source Components. *J. Chem. Inf. Comput. Sci.*, 43(6):1733 – 1739, 2003.
- [5] RDF Site Summary 1.0 (RSS). <http://web.resource.org/rss/1.0/spec>.
- [6] Mark J. Williamson Peter Murray-Rust, Henry S. Rzepa and Egon L. Willighagen. Chemical Markup, XML, and the World Wide Web. 5. Applications of Chemical Metadata in RSS Aggregators. *J. Chem. Inf. Comput. Sci.*, 44:462–469, 2004.

# Submitting articles to CDK News

A tutorial for writing  $\LaTeX$  articles for the CDK News newsletter.

by Egon Willighagen

## Latex

Articles for this newsletter are written in  $\LaTeX$  for its superior layout mechanism [1]. We understand that some users prefer to use Word or OpenOffice, but for this newsletter we will stick to  $\LaTeX$ .

For those who are not familiar with  $\LaTeX$ , it is a documentation tool that separates to a large extent content from layout. This means for the editors that they do not need to bother about designing each issue;  $\LaTeX$  will do that for them. The article source for a submission, however, does not just contain the content, but also mark up about the function of some content. This allows  $\LaTeX$  to properly layout the content to create the nice design of this newsletter.

For example, the title is marked up like `\title{Some Title}`.

A complete example of a CDK News article is as follows:

```
\title{Submitting articles to CDK News}
\subtitle{A tutorial for writing \LaTeX
articles for the \CDKnews newsletter.}
\author{by Egon Willighagen}
```

```
\maketitle
```

```
\section*{Latex}
```

```
Articles for this newsletter are written
in \LaTeX for its superior layout mechanism
\cite{art2:LaTeX}. We understand that some
users prefer to use Word or OpenOffice, but
for this newsletter we will stick to \LaTeX.
```

```
For those who are not familiar with \LaTeX,
it is a documentation tool that separates to
a large extent content from layout. Etc.
```

```
\address{Egon Willighagen\
University of Nijmegen, The Netherlands\
\email{egonw@sci.kun.nl}}
```

Not that complex, is it? Note that new paragraphs are started by having an empty line in the source code. Many other things are marked up using commands, e.g. `\LaTeX`. There are many of them, and a number are special to this newsletter.

There are a number of  $\LaTeX$  commands which are very useful for writing articles. Here are a few supported by the newsletter stylesheet:

`\code` Indicate text that is a literal example of a piece of a program.

`\kbd` Indicate keyboard input.

`\file` Indicate the name of a file.

`\command` Indicate a command name, such as `ls`.

`\dfn` Indicate the introductory or defining use of a term.

`\acronym` Use for abbreviations written in all capital letters, such as 'CDK'.

`\class` Indicate a CDK class, such as **Atom**.

`\pkg` Indicate a java package, such as `java.util`.

`\module` Indicate a CDK module, such as **core**.

`\url` Indicate a URL, such as `http://cdk.sf.net/`.

For example, a command line example should be marked up as `\code{\command{cdk-view} -h}`, resulting in this output: `cdk-view -h`.

Also, the full capabilities of  $\LaTeX$  are at hand. Defining equations is, thus, done as in normal LaTeX. Many  $\LaTeX$  tutorial explain how to include them. Here are a few useful examples.

The verbatim environment is used to wrap multi line source code:

```
\begin{verbatim}
AtomContainer container = new AtomContainer();
container.addAtom(new Atom("C"));
\end{verbatim }
```

of which the output will look like:

```
AtomContainer container = new AtomContainer();
container.addAtom(new Atom("C"));
```

A enumerated list can be made with the `enumerate` environment:

```
\begin{enumerate}
\item First item.
\end{enumerate}
```

The output looks like:

1. First item.

And a simple dotted list with:

```
\begin{itemize}
\item First item.
\end{itemize}
```

The output looks like:

- First item.

The capabilities of  $\LaTeX$  go far beyond this very short introduction. Googling the web for "latex tutorial" will give a list with many sources for further information. Furthermore, the editors may be contacted for assistance.

## The stylesheet

If you would like to layout the document in the format as it will appear in this newsletter you will need to download a  $\text{\LaTeX}$  distribution. The most common is TeTeX [2] which runs on Unix systems, but also on the Windows platform using CygWin [3]. On the CDK website an example article can be downloaded, as well as the 'CDKnews.sty' stylesheet.

To see what an article will look like when formatted using the 'CDKnews.sty' stylesheet the  $\text{\LaTeX}$  source for the article can be wrapped in another file, e.g. 'wrapper.tex', which looks like:

```
\documentclass[a4paper]{report}
\usepackage{CDKnews}

\bibliographystyle{unsrt}

\begin{document}

\begin{article}
  \input{art}
\end{article}
```

## CDK ChangeLog

"CDK ChangeLog" is a series in the newsletter summarizing the changes in the CDKlibrary since the previous newsletter.

by Egon Willighagen

This series gives an overview of recent changes in the CDK library, but in this special case - this is the first newsletter - it will focus on the last few releases. For each release the important changes are given.

### The 20040120 Release

- An important change in the release made on 20 January 2004 is the addition of the **ValencyChecker** as an alternative for the older **SaturationChecker**. The difference lies in the list of atom types it uses. The new class uses a list of atom types which explicitly gives the formal charge of the atom type. The **HydrogenAdder** has been adapted to be able to use both checkers; when constructing the object the valency checker that needs to be used can be given:

```
HydrogenAdder hAdder =
  new HydrogenAdder(
    "org.openscience.cdk." +
    "tools.ValencyChecker"
  );
```

```
\end{document}
```

The stylesheet should be placed in the same directory as the two  $\text{\TeX}$  files. A PDF file can then be created with the command `pdflatex wrapper.tex`.

I hope that this tutorial helps anyone getting started with using  $\text{\LaTeX}$  for writing articles for this newsletter. Good luck and send in those articles!

Egon Willighagen

University of Nijmegen, The Netherlands

egonw@sci.kun.nl

## Bibliography

- [1] L. Lamport.  *$\text{\LaTeX}$ : A Document Preparation System*. Reading, Massachusetts, 1994.
- [2] The teTeX HomePage. <http://www.tug.org/teTeX/>, 2004.
- [3] Cygwin Information and Installation. <http://www.cygwin.com/>, 2004.

- The lazyCreation patch was applied to the **ChemObject** improving the memory usage and time to create a new **ChemObject** considerably. This patch was already available in previous releases, but was only applied when specified. The patch delays the memory allocation and initialization of a few object fields until the variable is really used.
- The coordinate system of the **Renderer2D** has changed to match a more commonly used system with (0,0) in the lower-left corner, instead of the top-left corner common in Java. It is important to note that this modification changes wedge bond based stereochemistry!

Many other bug fixes, addition and other changes are documented in the complete CHANGELOG which can be found online [1].

### The 20040202 Release

- This release was mostly a bug fix version of the 20040120 release, but also includes a reworked build process: module information is now extracted for the '.java' files using a JavaDoc do-let. This makes compiling of a specific module much easier, and makes it easier to under-

stand dependencies between classes and modules. The doclet creates '\*.javafiles' in the src/ directory which explicitly lists all classes in a specific module. And only those get grouped into the jar file for that module.

- Release 20040202 also included updates to the **io** module: a **HINReader** and **HINWriter** for HyperChem (<http://www.hyper.com/>) files were added, the **CMLWriter** can now write namespaced CML, and a stereochemistry writing bug was fixed in the **MDLWriter**.

## The 20040324 Release

- The most important change in this release is the generalization of the way in which the **UniversalIsomorphismTester** compares bonds and atoms. The comparison is now customizable with the default being the comparisons that were used so far: bond order match, and element symbol match. This makes it possible to use more complex ways to compare atoms, and the SMARTS substructure search mentioned elsewhere in this issue is based on this extension.
- Support for reactions was improved in this release. Reaction SMILES can now be parsed

and created. These one line representations extend normal SMILES by listing which molecules are reactants, products or agents, like catalysts. The following strings describes the reaction of acetic acid and ethanol under acidic conditions: CC(=O)O.OCC>[H+]>CC(=O)OCC.O.

- Finally, I would like to mention the fix of the **IterationMDLReader** which was introduced in the previous release. This reader allows parsing a SDF file molecule by molecule allowing to process files with thousands of molecules without running out of memory.

## Wrap up

A lot of changes are not mentioned in this article. As mentioned earlier, check the full changelog for a complete list of changes [1]. Especially, the API change sections are important, indicating changes which could break your CDK based software.

*Egon Willighagen*  
*University of Nijmegen, The Netherlands*  
egonw@sci.kun.nl

## Bibliography

- [1] CDK Changelog. <http://cdk.sf.net/changeLog.html>.

### Editors-in-Chief:

Egon Willighagen [egonw@users.sf.net](mailto:egonw@users.sf.net) and  
Christoph Steinbeck [steinbeck@users.sf.net](mailto:steinbeck@users.sf.net)

### Editorial Board:

Egon Willighagen, Christoph Steinbeck and Rajarshi Guha.

*CDK news* is a publication of the Chemistry Development Kit (CDK) project. All articles are copyrighted with GNU's FDL by the respective authors. Submis-

sions can be send to the Editors-in-Chief.

Contact address of the CDK project representative responsible for this serial publication:

Dr. habil. Christoph Steinbeck  
Cologne University Bioinformatics Center (CUBIC)  
Zülpicher Str. 47  
50674 Koeln

CDK Project homepage:

<http://cdk.sourceforge.net/>