Technological and Social Aspects of Collaboratively Developed Chemical Information Systems (CIS)

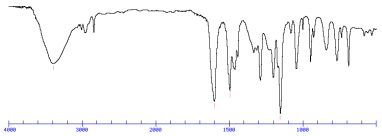
Christoph Steinbeck

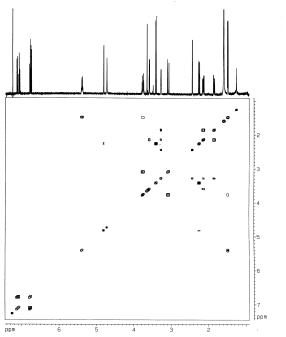
Cologne University Bioinformatics Center (CUBIC)

Chemoinformatics, Open Source and the Internet
A Somewhat Personal History.

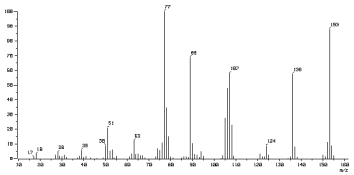
- 1985 Minix, Andrew Tanenbaum
- 1990 World Wide Web by Tim Berners-Lee
- 1992 Release 0.2 of Linux
- 1993 Mosaic Web Browser by NCSA
- 1997 First release of CompChem java library and JChemPaint 2D editor
- 1998 Jmol
- 2000 The Chemistry Development Kit (CDK)
- 2001 JOELib, OpenBabel
- 2002 NMRShiftDB
- 2005 The Blue Obelik Movement, PubChem
- 2006 Bioclipse

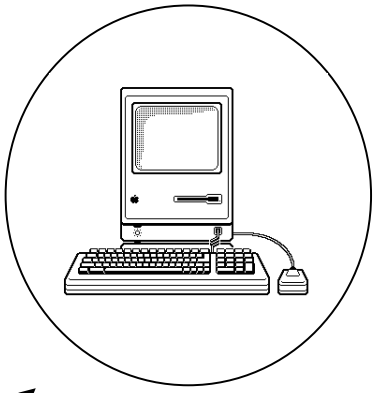
Computer-Assisted Structure Elucidation (CASE)

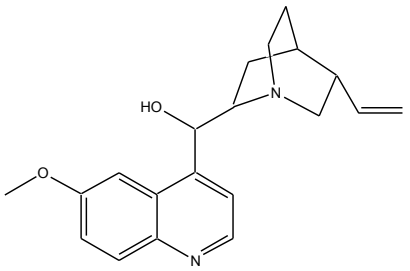




Molecular Formula from HR-MS







- Constitution

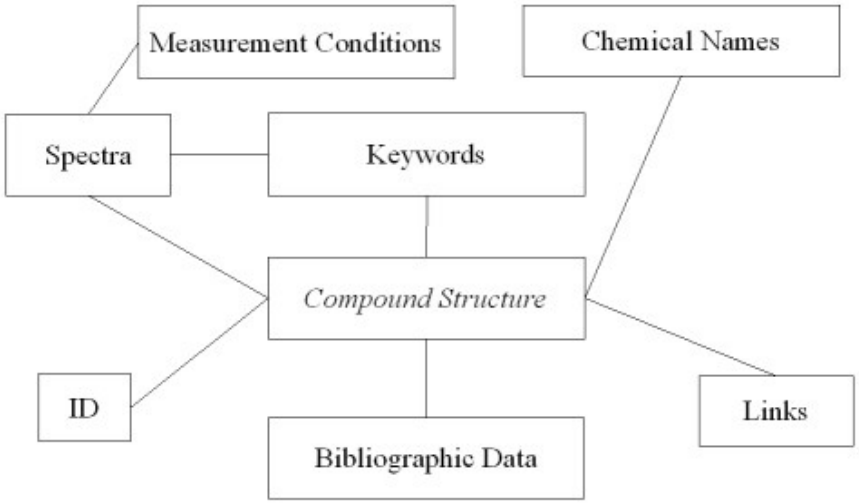
- Conformation
- Chirality
- E/Z Stereochemistry

Further Facts:
- Substructures
- Drug Likeness
- Natural Product Likeness
- etc.

NMRShiftDB (http://www.nmrshiftdb.org)

An open access, open submission, open source database for organic structures and their NMR data

- >19.000 Structures and > 21.000 assigned spectra
- Searching for spectra, structures and facts
- NMR Prediction for ^{13}C-, ^{1}H-NMR and other nuclei
- Corina-generated 3D coordinates
- ACD/Names™ IUPAC-Names
- Open Submission, Peer-Review by human reviewers, published as a daily blog



Steinbeck, C.; Kuhn, S.; Krause, S. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1733 - 1739.
Steinbeck, C.; Kuhn, S. *Phytochemistry* **2004**, *65*, 2711-2717.

Proposition:

Chemical Information Systems can be authored (read: produced and maintained) by their user communities, not only by secondary publishers

Does it work?

- Yes, but.....
  - A lot of contributions from very few people
  - Almost no feedback by a very large, but very silent community.
  - You can write papers, beat the drum: most people will not have heard of your product (related to that: Google ranking)
  - No “Hey, it's not perfect, so let's make it better” attitude in Chemistry

- Are we alone with this experience?

Social Issues

Science & Technology

August 14, 2006
Volume 84, Number 33
pp. 50-53

Online Archives On A Bumpy Road

Digital repositories have garnered mixed reviews—everything from enthusiasm to apathy

Sophie L. Rovner

Libraries have always been the storehouses and caretakers of human knowledge. But with so much scholarly material now being produced in digital form, a new storage option has emerged: online archives that can provide free and (one hopes) enduring access to this material.

OA/OS databases vs. Online archive

- Database requires careful editing and curation of dataset (more work than archive)
- Database utility in everyday work is much more obvious.

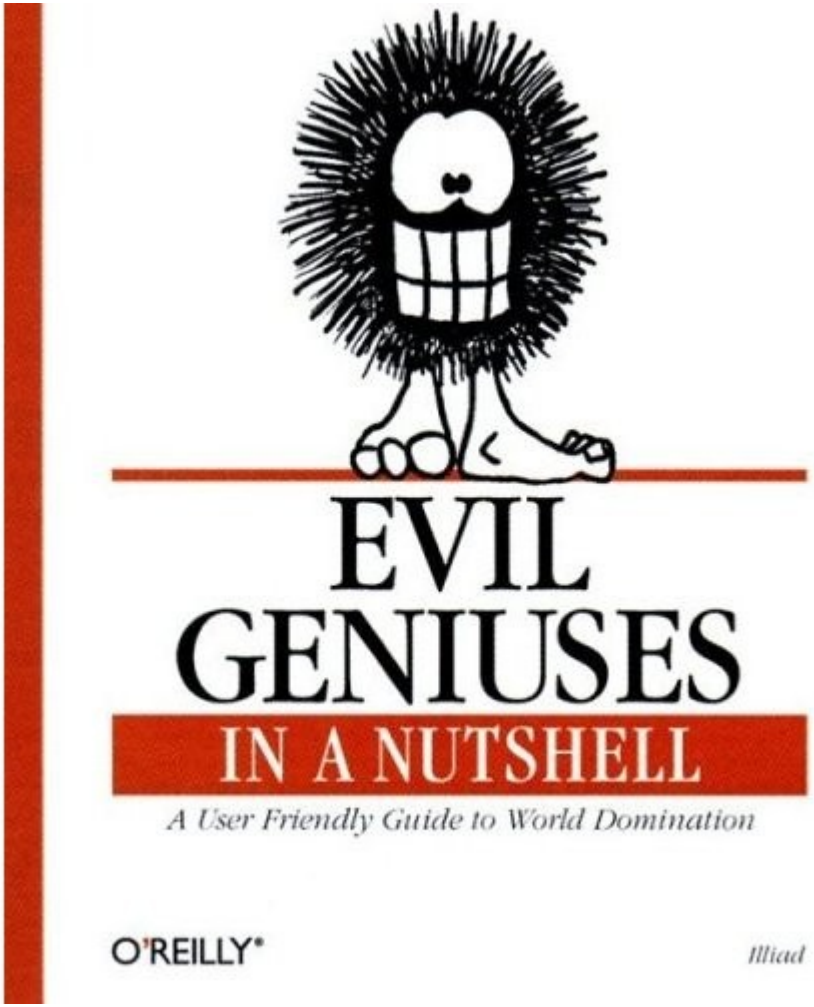
Social Issues: Overcoming the Schweinehund

- It's hard to motivate people to contribute to an open CIS.

- "It's a useful (important, cool, nice to have, ...) system we are building here" just is not enough.

- Ways to get people on board:
  - Force them (nicely): Submission to a peer reviewed journal requires deposition (With advantages for both side).
  - Make their life easier if they do it: the tools which do the quality checking and layout also do the submission to the database.

- Ways to fill the database without the help of humans: Let robots do it. (See next talk in this session and my talk (CINF 101) on Wednesday.)


EVIL
GENIUSES
IN A NUTSHELL
A User Friendly Guide to World Domination
O'REILLY
Illiad

Some random thoughts on technical aspects

- Classical hand-tailored vs. multi-purpose systems.
- Wikipedia is a collaboratively developed information system – partly even a chemical information system
- However:
  - Chemistry has special has special requirements for queries, display and analysis of stored data (full text searches are just not enough).
  - Structure-Property-Databases (most common type of CIS in chemistry) may have complex data relations
- So, can we integrate structure, spectra and other chemistry-specific search features with Wiki technology in a plugin-like manner?

Financial Issues

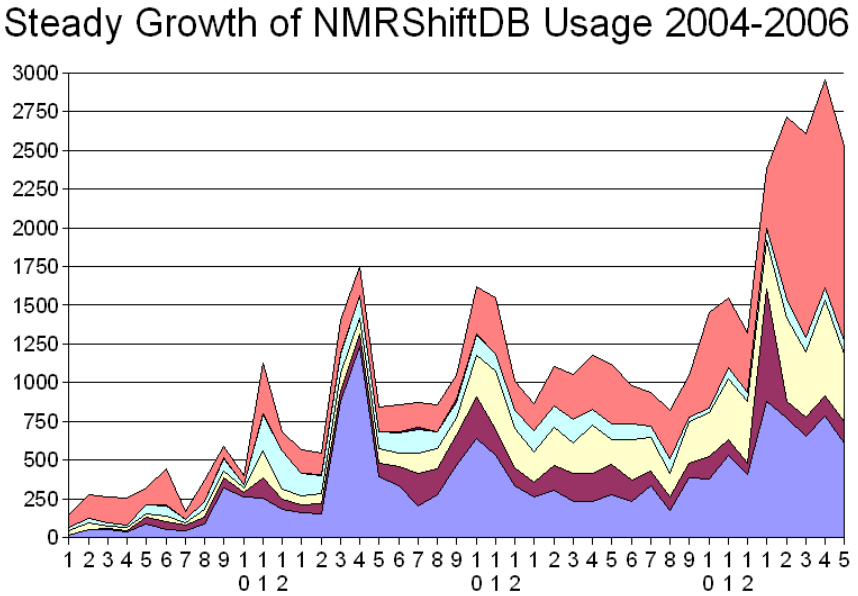
- We can collaboratively build an open CIS, but:
  - If there is no institution making money with the database, who maintains it, i.e.
    - Who maintains the software,
    - Who pays for the server,
    - Provides support, etc.
- Things are relatively easy if you deal with a computer science/hacker community.
- Most chemists are neither of that.

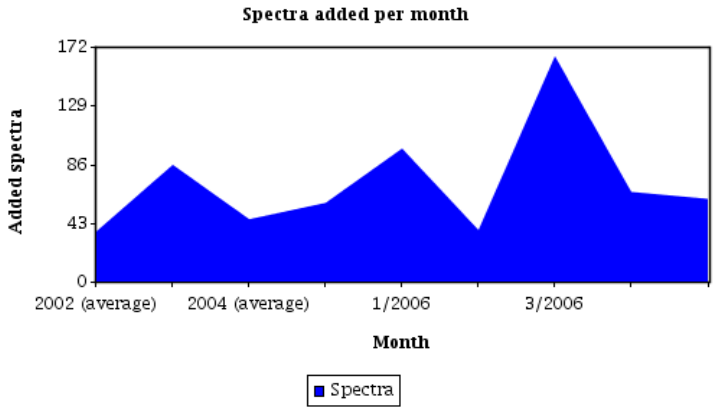
- One solution: National (continental, world-wide) centers for database hosting with guaranteed long-term funding.

NMRShiftDB (http://www.nmrshiftdb.org)

An open access, open submission, open source database for organic structures and their NMR data

- 900 registered users
- Runs on 4 mirrors servers in three geographic locations, two more soon
- Spectra for all compounds now being computed by ab-initio methods by Wavefunction inc.
- Now being fed by Peter's JournalEatingRobot™ (see CINF101)





Steinbeck, C.; Kuhn, S.; Krause, S. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1733 - 1739.
Steinbeck, C.; Kuhn, S. *Phytochemistry* **2004**, *65*, 2711-2717.

Set aside the social aspects: Do we have the technology to do it?

- My answer is “yes” (surprise!)

- Bioclipse
- Chemical MIME Data
- Chemistry Development Kit (CDK)
- JChemPaint
- Jmol
- JOELib
- JUMBO
- Kalzium
- NMRShiftDB
- Octet
- Open Babel
- QSAR
- World Wide Molecular Matrix

The Chemistry Development Kit (CDK)

- An open source Java library for structural chemo- and bioinformatics
- 20-30 developers world-wide
- ~ 66.000 Lines of Code (NCSS), > 6000 Methods, > 660 Java Classes (plus 255 unit test classes)
- Clear object-oriented design
- Easy creation of web services and applications (such as http://www.nmrshiftdb.org)
- Used as library in more than 10 academic and industrial projects
- Quarterly Newsletter “The CDKNews” ISSN!
- http://www.sourceforge.net (development),
- http://www.chemistry-developement-kit.org (showcase site)

Steinbeck, C.; Han, Y. Q.; Kuhn, S.; Horlacher, O.; Luttmann, E., Willighagen, E.
JCIC **2003***, 43*, 493-500.

CDK Functionality

Input/Output

- I/O (CML, MDL, PDB, INChI, ...)
- SMILES Parsing and Canonical Generation

Visualization

- Structure-Diagram-Layout (SDG)
- 2D Rendering
- 3D Rendering

Modeling

- 3D Model-Builder
- Atom-Typing
- Force-Field
- Representation of Biomolecular structures

Chemical Graphs

- Isomorphism detection
- Maximum-Common-Substructure Searches
- SMARTS- and Substructure searches
- Ring searches (Smallest Set of Smallest Rings, All Rings)
- Aromaticity detection

Structure generation

- Deterministic Isomer generator
- Stochastic Structure Generators via
  - Simulated Annealing
  - Genetic Algorithms

Properties

- Fingerprinting
- Gasteiger-Charges
- > 30 QSAR-Descriptors
- NMR-Prediction
  - via HOSE-Codes
  - SVMs

Bioclipse

- Collaboration with Uppsala-University
- JAX audience award winner
- Pluggable functionality for handling:
  - Molecules
  - Spectra
  - Sequences
  - RSS-Feeds
  - And much more

- Version 1.0 now at http://www.bioclipse.net
- Now with NMRShiftDB dataset authoring plug-in (see CINF 101 talk)
- Also includes OSCAR tool for chemical archeology.

Acknowledgment

Stefan Kuhn (Work), German Research Council (Money)

Spontaneous Breakout of Philanthropy

This open community actually makes working in science being fun again.

URLs

http://cdk.sourceforge.net (Chemistry Development Kit, CDK)

http://www.chemistry-development-kit.org (Showcase-Site)

http://www.nmrshiftdb.org (NMR-Datenbank)

http://www.bioclipse.net (Bioclipse Development Site)

http://www.blueoblisk.org (Blue Obelisk Movement)