

Dynamic Indexing of Chemical Metadata Using Open Tools: Case Study of Open Babel, CDK, and the Blue Obelisk

Geoffrey R. Hutchison, Tobias Helmus, Stefan Kuhn,
Henry S. Rzepa, Christoph Steinbeck,
Christopher J. Swain, Egon L. Willighagen

ACS Fall Meeting
September 14, 2006

“It is easier to use Google to find something from one of a billion web sites than on your own disk.”

— *Steve Jobs, CEO Apple Computer
WWDC, June 2004 San Francisco*



“I can plug my iPod into any computer and it will recognize my music and give me all sorts of metadata: artist, title, type of music...

Why can't I read the chemical metadata off my chemistry files?”

— *Prof. Henry S. Rzepa,*
Spring 2005 ACS Meeting, San Diego, CA





10,000+ chemical files!

Apple Spotlight



- System-wide index and search mechanism
- Full-text searching and restricted metadata (e.g., music files, digital photos, publications...)
- Third-party API for indexing (plugin mechanism for new filetypes)
- Third-party API for searching

ChemSpotlight: Dynamic Chemical Metadata

- Use the system-wide search database (no coding for indexing or retrieval code)
- Includes textual data in chemistry and non-chemistry files (e.g., chemical names, formulas in documents, etc.)
- Multiple retrieval and filtering interfaces (i.e., any third-party search tool works)

ChemSpotlight: Dynamic Chemical Metadata

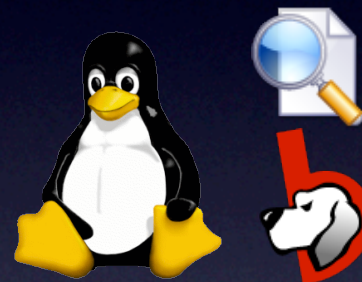
- Concept not limited to Macintosh:



Windows Vista



Windows + Google



Linux: Beagle, Strigi

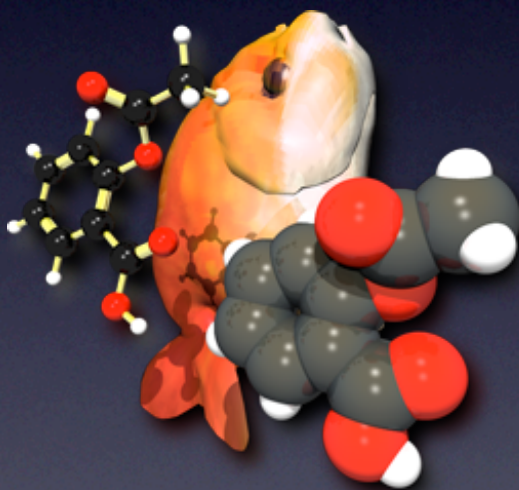
- Index files on-demand as they are modified
- Support for plugins to support new filetypes
- No restriction on file organization

ChemSpotlight: Indexing Architecture



Spotlight

+



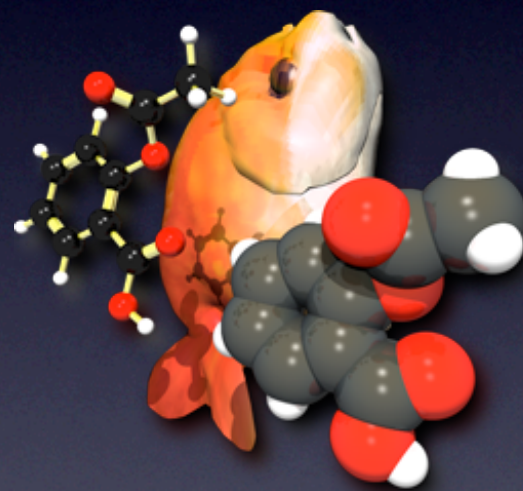
Open Babel

+

~300 lines
of code

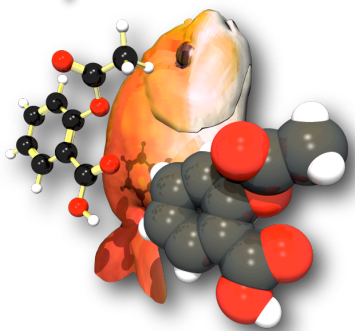
Open Babel: Not Just File Conversion

- A toolkit for chemistry development
- Searching / similarity fingerprints
- Molecular, atomic, bond descriptors
- Atom and bond typing, perception, aromaticity, chirality...



Editor / Viewer

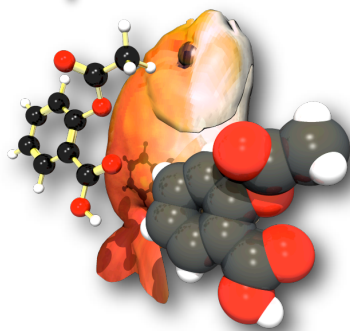
Open Babel



OpenGL
Graphics

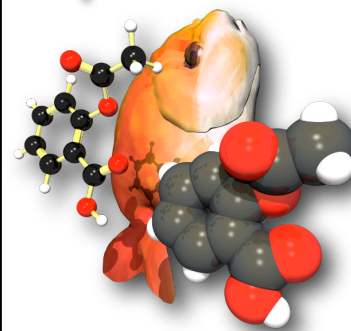
Analysis

Open Babel



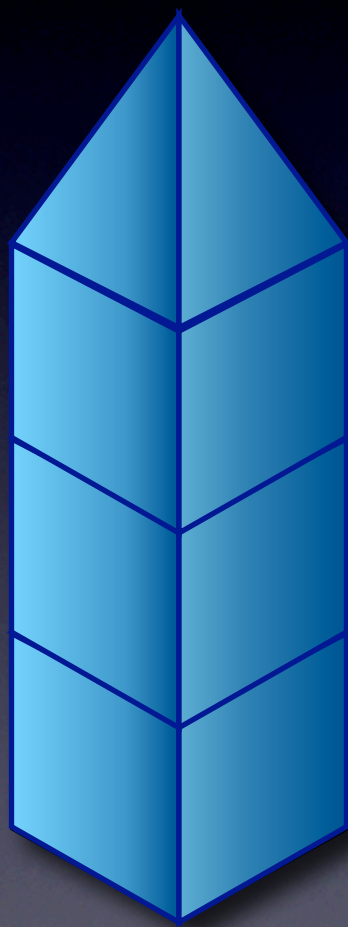
Database

Open Babel



- Code reuse through open source code:
*Focus on problems **beyond** the basics*
New science, not new software development
- Rapid development
- Reduce non-standard file formats & bugs

Blue Obelisk



- “Formed” at Spring 2005 ACS meeting, San Diego
- Open Source
- Open Data
- Open Standards
- Fun

<http://www.blueobelisk.org/>

Stored Metadata

- Residue sequence (for biomolecules)
- Molecular weight & exact isotopic mass
- Molecular formula (e.g., C₄₂H₅₆S₁₅)
- Number of atoms, bonds, residues, molecules (e.g., multi-molecule file)
- Dimensional data: 2D vs. 3D, chirality
- Daylight SMILES
- InChI identifier
- More to come...

Currently Supported Filetypes

- XYZ, PDB, MDL SDfile, Tripos Mol2, CML...
- Limited only by support in Open Babel (currently 57 formats, with more to come)
- In *principle*, ChemSpotlight can also support embedded chemical metadata in XML, Word, PDF, etc. files

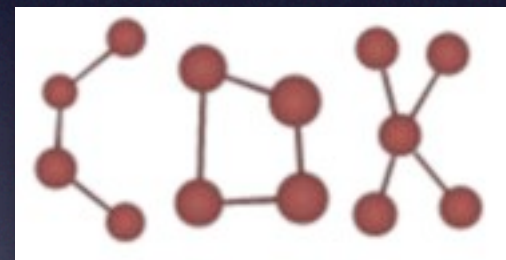
ChemSpotlight: Retrieval Architecture



+



+



Spotlight

iBabel

JChemPaint

ChemSpotlight: Retrieval Architecture

- Generic system-wide text search (via menu)
- Programmable API
- Command-line
- Third-party tools
 - Desktop apps
 - Workgroup searching
 - Web servers



Demos

Conclusions

- Dynamic, on-demand indexing of chemical data via system-wide search features
- System-wide text search and restricted metadata interface
- Open source and extensible
Easy to adapt for your workgroup
(your filetypes, your metadata, your disks)