A semantic GRID for molecular science

Peter Murray-Rust [a], Robert C Glen [a,] Henry S Rzepa [b,] James J P Stewart [c,] Joe A Townsend [a,] Egon L Willighagen [d,] Yong Zhang [a]

[a] Unilever Centre for Molecular Informatics, Chemistry Department, University of Cambridge, Lensfield Road, Cambridge, CB2 1EW, UK, [b] Chemistry Department, Imperial College, London, SW7 2AY, UK , [c] Stewart Computational Chemistry, 15210 Paddington Circle, Colorado Springs CO 80921-2512 US, [d] Laboratory of Analytical Chemistry, Toernooiveld 1, 6525 ED Nijmegen, NL

Abstract

The properties of molecules have very well defined semantics and allow the creation of a semantic GRID. Markup languages (CML - Chemical Markup Language) and dictionary-based ontologies have been designed to support a wide range of applications, including chemical supply, publication and the safety of compounds. Many properties can be computed by Quantum Mechanical (QM) programs and we have developed a "black-box" system based on XML wrappers for all components. This is installed on a Condor system on which we have computed properties for 250, 000 compounds. The results of this will be available in an OpenData/OpenSource peer-to-peer (P2P) system (WorldWide Molecular Matrix - WWMM).

Introduction

Over 30 million chemical compounds are known, many of importance in healthcare, biosciences and new products. It is of fundamental importance to know their properties, including implications for safety. The UK's Royal Commission on Environmental Pollution (http://www.rcep.org.uk) has recently emphasised the importance of having this information and stresses the very low percentage of compounds for which adequate data are available.

It is common to attempt to model the biological and other safety properties of molecules from known physical properties. Sometimes these have been measured but in most cases they must be predicted. Many properties can, in principle, be calculated by solving Schroedinger's equation, although this was often prohibitively expensive. In particular calculations scale badly, often $O(N^3)$ to $O(N^6)$ where N is the number of atoms or electrons. However recent advances include:

- O(N) scaling (usually through localisation of parts of the molecule in the algorithm).
- Farm-like availability of unused compute cycles on non-specialist machines in heterogeneous environments.
- Semi-empirical parameterised methods (QM Hamiltonians such as MOPAC's PM5 are calibrated against experimental properties).

Raw computer power is often not the major challenge. We report below the automatic computation of properties of 250, 000 molecules but stress the analysis and dissemination of results is much more problematic. The data and codes are extremely heterogeneous with no common infrastructure and often based on column-based FORTRAN-like input. Novitiates must study at the feet of experts till they master it. Anecdotally we estimate that input errors render many millions of jobs wasted globally per year. The programs have virtually no interoperability, and numeric output is often poorly defined without explicit scientific units. Metadata (when, where, why, what, how) are universally absent.

Paradoxically this is one of the best test beds for developing a semantic GRID! The underlying (implicit) semantics are extremely stable (the molecule-property relationship (Fig. 1) dates from the 19th century).

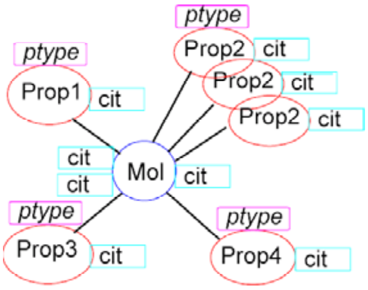


Fig.1. A Molecule has many properties (perhaps with repeat measurements) defined by their types (`ptype`) and citations (`cit`).

The codes themselves are very reliable with excellent implicit semantics. In most cases algorithms (if not always source code) are fully described and agreed. Moreover there is now great demand for computational chemistry for many domains outside chemistry, such as materials, safety, biosciences, earth sciences and nanotechnology. These "customers" increasingly want a "black-box" approach that provides "useable" results on demand.

This challenges much current practice where users are expected to understand the physics of the calculations, and the many pitfalls. Often there has to be an “expert on tap". Whilst the quality of input and the interpretation of output are suspect, this is still essential but we believe that the process can be increasingly "semantically wrapped". A set of rules can decide which molecules are unsuitable for calculations, and what level of accuracy can be expected or afforded for the others. For example a small rigid molecule containing light elements will often give excellent results on a routine basis while large floppy molecules, those with metals or unpaired electrons are immediately filtered out.

Most importantly we provide metadata for each job. This allows the customer to make their own decision as to whether the results are "fit for purpose". Computer-based tools can also analyse the results both for known problems and to discover types of molecules that show pathological behaviour. The traditional code manuals can evolve to a rule-set taking decisions or advising users on options.

The Chemical Semantic GRID

In our earlier vision of the Chemical Semantic Web [3] we foresaw scientists asking for chemical information on demand, often without knowing the details of the science involved. This slightly heretical approach is driven by the pace and heterogeneity of multidisciplinary science, exemplified well in bioinformatics. A scientist must retrieve data from many domains and integrate them without the help of human experts. A key factor in the Semantic Web is transferring expertise into computer representation ("ontologies").

Here we extend this to a Semantic GRID with high-throughput computing on demand. In molecular science this is challenging as it does not map easily onto current informatics practice. There is no equivalent to the publicly funded international bioinformatics institutes that provide Open Data (see below). Most published molecular data is published piecewise in individual journal articles and subsequently partially aggregated by fee-charging organisations. Most is never captured in re-usable e-form; we believe that >90% of computational chemistry published in the peer-reviewed literature is therefore effectively unavailable for a GRID.

Thus culture and business practices are the greatest problems in developing a semantic GRID. There is virtually no formalisation of computer semantics, with each supplier of resources (especially codes and databases) taking a self-centric approach and expecting the user to tool up for their (implicit) semantic model. We are therefore providing a radically different approach based on distributed ontologies, P2P installations, Open Source and, equally important, Open Data.

Open Data

In molecular science the largest barrier to a global semantic web is the difficulty of re-using data. Authors expect their published data to be re-used, but data aggregators require payment for their collections. These are frequently only available on per-entry search and almost always prohibited from redistribution. This stifles exploratory data-mining (a great success of 19th century chemistry) and hinders the creation of transformed, filtered, and aggregated e-handbooks. There is an increasing amount of "grey web data" where sites provide molecular information (in sizeable amounts) but without provenance or intellectual property rights (IPR) for re-use. A typical problem is that some sites make their data freely searchable but discourage spiders and do not provide a complete datafile for download. We believe that, given the tools and zero effort, most authors would welcome the publication of their data in an Open re-usable manner.

There are about 30 million published molecules (and Chemical Abstracts indexes each with a unique semantically void numerical identifier ("CAS number")). We estimate that only 1% are available in Open collections. Of these the largest is the US National Cancer Institute (NCI) which provides 250,000 molecules (although currently without any experimental properties). Other public "grey" sources include the webbook from NIST, ChemIDPlus from NIH and the ligands from the Protein Data Bank.

We summarise "the rights of molecules":

- Most scientists and their funders intend peer-reviewed publication of molecules and data to be re-usable by the community.
- All such molecules and their published properties should be freely and openly available to the global community.
- All molecules should carry the author's metadata on provenance and IP.
- Molecules and their metadata must be freely distributable and incorruptible without hindrance.

We refer to this as Open Data; it can be redistributed and re-used automatically without further permission. If transcluded or aggregated it must be preserved intact with the author's provenance and metadata. Authors can digitally XMLsign their data/metadata to ensure this and our system can, if needed, be configured to reject unsigned data.

We stress that current copyright and IPR must be respected. Data must only be entered if the contributor has the right to do so. Unfortunately the precise position on re-use of electronic scientific data is unclear and we shall omit discussion here.

Model for a semantic molecular GRID

Our model involves the **capture of data at source with minimal effort on the part of authors**. All components are Open Source and Open Data allowing anyone to consume or contribute data or source. Since this model is not widely used in molecular science we have to provide concrete incentives! We therefore offer free computation of molecular properties through a "black-box" of wrapped QM programs. Authors prepare their paper as normal (MS Word and a range of XML-unfriendly legacy molecular editors) which is automatically converted to XML (either on our site or in an Open toolkit). This XML is then fed into the high-throughput "black-box" which calculates properties. The results are transformed into XML and returned to the author.

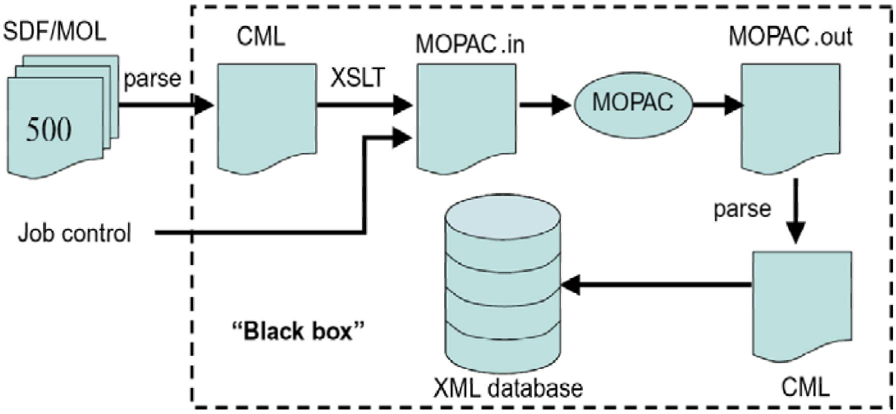
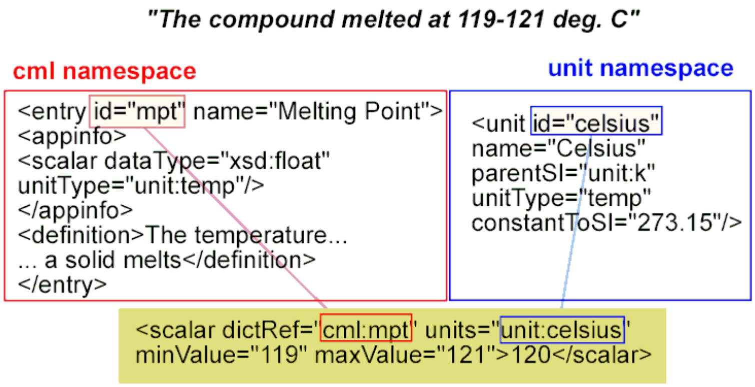


Fig. 2. Our black-box for high-throughput computing of molecular properties. Legacy data is transformed to CML, combined with XML job control and transformed to legacy program input. The output is parsed to XML and stored in an XML repository.

The contributor need know nothing of the details of the QM calculations but the final results (molecular geometry, energy, dipole, charges, frequencies, etc.) are of immediate value. The tradeoff of this barter is that the molecules and their data are made Open to the whole community. Initially we shall publish them on our site where *anyone* can re-use them and we are actively discussing with digital libraries how the data can be Openly archived.

The model relies on shared extensible metadata and ontology. All datuments [2] (data+documents) are in XML and can use any commonly agreed languages. We expect that MathML, XHTML, DocBook, SVG, and CML (Chemical Markup Language) [1] will be common. CML is an extensible family of components supporting molecules, reactions, spectra, computational chemistry, etc. For general scientific data we have created STMML [2] to support data shapes (scalar, array, matrix), datatypes (integer, float, date) and geometry (`vector3`, `plane3`, etc.). STMML also supports dictionaries (our model for extensible ontologies). A data component must be linked to a dictionary entry describing the human-readable meaning and the machine-understandable semantics (data types, constraints, relations, etc.). Dictionaries are loosely based on concepts from XMLSchema and are extensible in that anyone can create their own. Dictionary entries are identified through namespacePrefixed entry references so that they can be globally uniquified and located.



```
<entry id="mpt" name="Melting Point">
<appinfo>
<scalar dataType="xsd:float"
unitType="unit:temp"/>
</appinfo>
<definition>The temperature...
... a solid melts</definition>
</entry>
```

```
<unit id="celsius"
name="Celsius"
parentSI="unit:k"
unitType="temp"
constantToSI="273.15"/>
```

```
<scalar dictRef="cml:mpt" units="unit:celsius"
minValue="119" maxValue="121">120</scalar>
```

Fig. 3. Semantics are added to scientific objects by linking to communal dictionaries (`cml` for chemical concepts; `units` for scientific units).

The strategic implementation is designed to evolve communally; there is no central control. The only constraint on contributors is that they agree to honour the infrastructure and ethos of the system. All contributions will contain metadata identifying the owner so that they are responsible for the content. They agree to use the same dictionary *structure*, basic metadata (Dublin Core), to avoid namespace collisions and to provide unique identifier(s) based on URIs.

The technology will be Openly distributed. This allows a P2P system to evolve where each site holds only the subset they are interested in. Some may hold rich data on a subset of molecules, others may have sparse data on as many molecules as they can find. Many may simply publish their own work, knowing that it can be permanently archived in re-usable form.

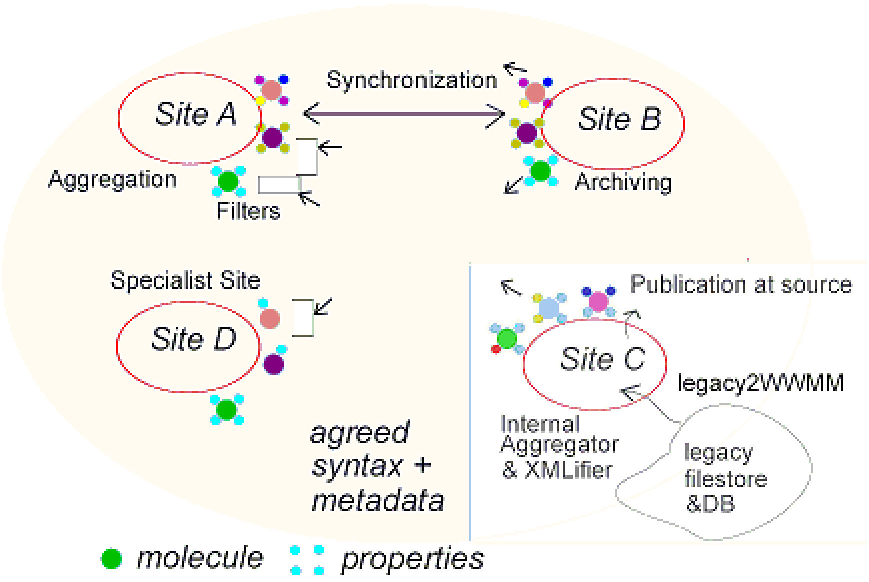
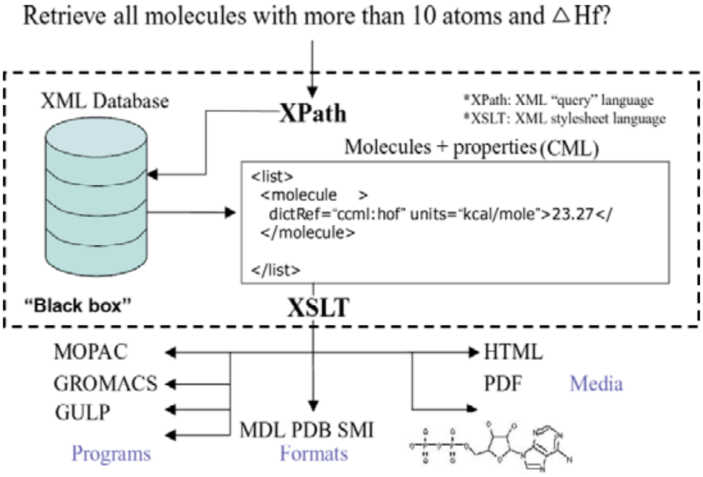


Fig. 4. A P2P system (the WorldWide Molecular Matrix (WWMM [4]) based on this infrastructure. Ellipses represents sites; large circles represent molecules; small ones their properties. Sites can have different selections of each and robots can synchronize subsets of data.

Implementation

The fundamental design criteria are:

- **Data contributors need not know the basis of the technology.** Authors can continue to use their current legacy toolset and we convert the results to XML. (We are also developing XML-based tools and providing libraries for current developers.)
- **Run-anywhere**. Almost all components are 100%-Java. In some cases portable contributed tools in C/++ (with source) are used.
- **Open Source**. This includes XML infrastructure (`ant, tomcat, Xindice, saxon, FOP, batik, etc.) and molecular science (JUMBO, Jmol, JChemPaint, CDK, JOELib, OpenBabel, etc.).`


Retrieve all molecules with more than 10 atoms and △Hf?
XML Database
XPath
*XPath: XML "query" language
*XSLT: XML stylesheet language
Molecules + properties (CML)
"Black box"
XSLT
MOPAC
GROMACS
GULP
Programs
MDL PDB SMI
Formats
HTML
PDF
Media

```
<list>
 <molecule    >
  dictRef="ccml:hof" units="kcal/mole">23.27</
 </molecule>

</list>
```

Fig. 5. A 100%-XML query system. XPath-queries are submitted to the XML databases Xindice which returns CML. The CML can be input to programs, transformed to legacy molecular formats, converted to multiple media and rendered in CML-aware tools.

The primary molecular key is the new IChI unique identifier from the International Union of Pure and Applied Chemistry. This is generated automatically from the chemical structure and we use it to index molecules in the XML database Xindice.

For high-throughput of computation we have created a generic `ant` library in which legacy codes are wrapped as a black-box. Condor has been installed on 20 Windows teaching machines (777MHz). Results are parsed and packaged as a series of HTML files

Methodology, Results and Discussion

The 250,000 molecules from NCI were converted into XML (JUMBO) and split into 500 batches of 500. Each batch is run as a single Condor job, with choice of three protocols (varying speed and accuracy). Results are parsed to CML and transformed to SVG/CML/HTML pages (Fig. 7.) with an interactive display (`Jmol`). The CPU time for each molecule varied by 10^6 (0.3 secs to 4 days). There were significant cost benefits in applying rule-based triage in the later stages.

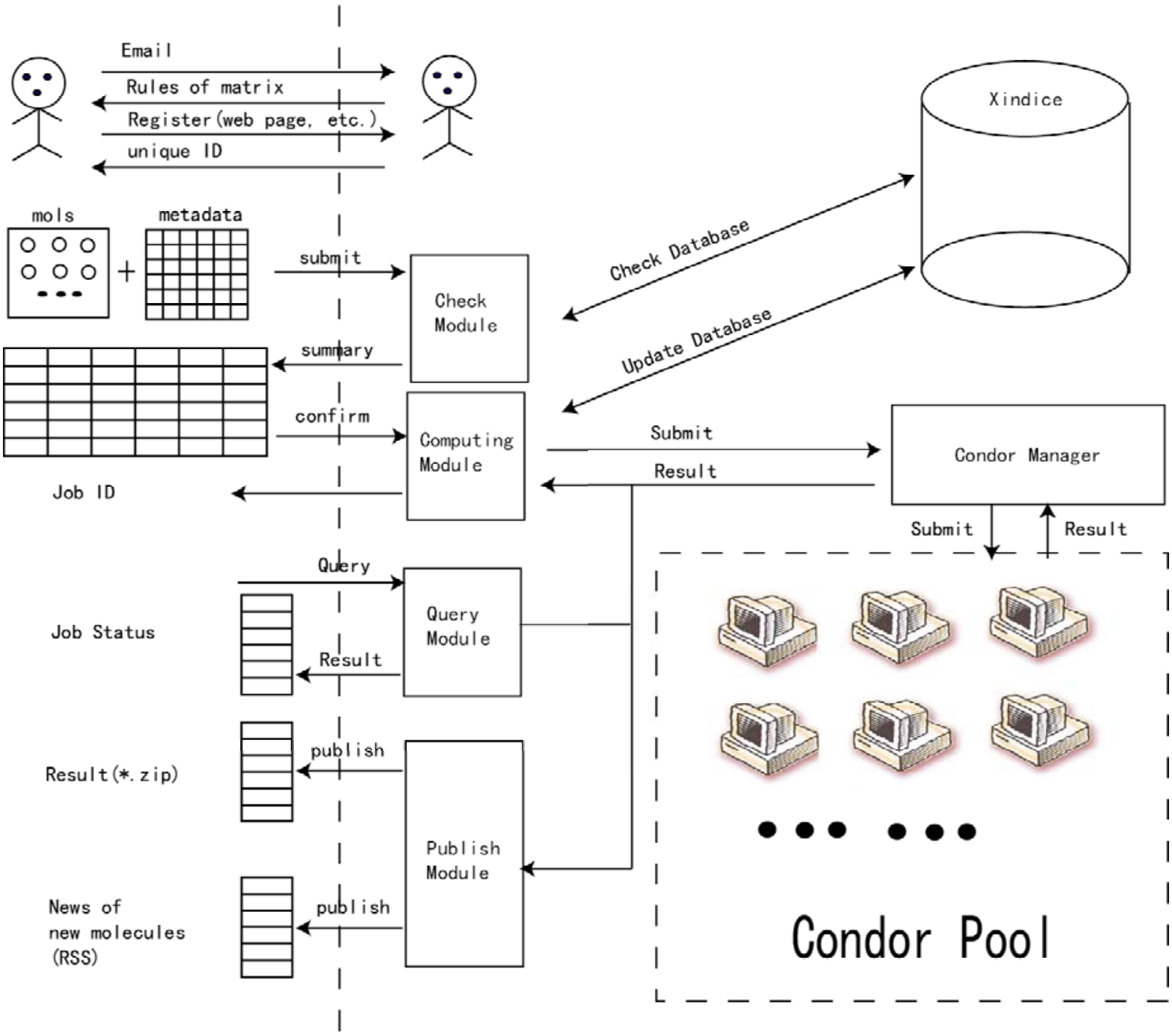
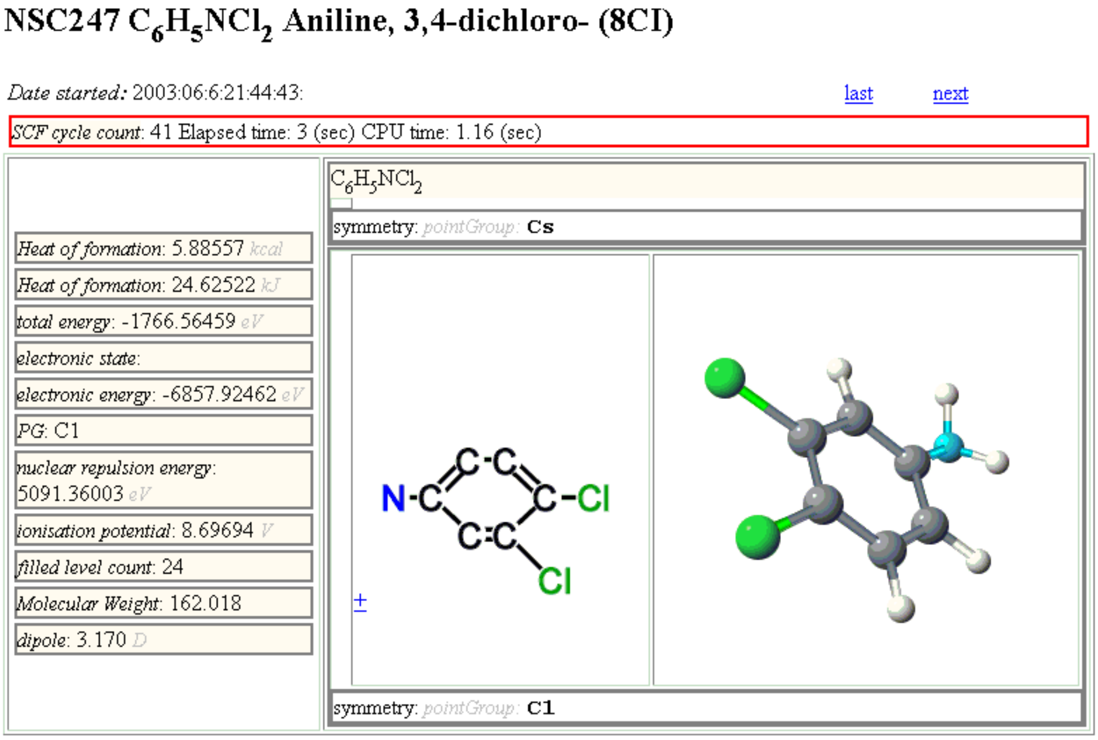


Fig. 6. Workflow through the Condor system. Users register with our site, then submit a bundle of legacy data and molecules. These are checked for validity and novelty and then submitted to Condor. Job status and final results are published Openly on our site.

The Condor pool has run without problems for 3 months and will shortly be GRID-enabled through Condor-G. The granularity allows easy job tracking, but variation in CPU times suggests smaller batches (less than 50 molecules).

Xindice has been loaded with a test set of 10,000 molecules. Xindice `document` granularity must be at individual molecule level. Retrieval of a single entry by indexed element or attribute shows good performance (ca 10 ms on a 1.5GHz Linux machine). We noted that indexes were often large and apparently wasteful of space. Even when indexed, retrieval time seems to depend on the number of *hits* and can be slower for large amounts of retrieved data.

Elsewhere we have been working with molecular science publishers on converting authors' manuscripts to ultra-finegrained XML/CML with good retrieval and precision. This opens the prospect of authors depositing their manuscripts at source before/at/after publication.



NSC247 $C_6H_5NCl_2$ Aniline, 3,4-dichloro- (8CI)

Date started: 2003:06:6:21:44:43:

last next

SCF cycle count: 41 Elapsed time: 3 (sec) CPU time: 1.16 (sec)

$C_6H_5NCl_2$

symmetry: pointGroup: **Cs**

Heat of formation: 5.88557 kcal

Heat of formation: 24.62522 kJ

total energy: -1766.56459 eV

electronic state:

electronic energy: -6857.92462 eV

PG: C1

nuclear repulsion energy: 5091.36003 eV

ionisation potential: 8.69694 V

filled level count: 24

Molecular Weight: 162.018

dipole: 3.170 D

symmetry: pointGroup: **C1**

Fig. 7 One of 250, 000 results

We thank many members of the OpenSource community, including OpenBabel (http://openbabel.sf.net), Jmol (http://jmol.sf.net), CDK (http://cdk.sf.net), JChemPaint (http://jchempaint.sf.net) and JOELib (http://joelib.sf.net) and the generic projects listed. We thank the DTI/eScience project (YZ), Unilever Research (JAT) and the Marie Curie Training Sites scheme (ELW). Thanks to Steve Stein and colleagues (NIST) for data and IChI, and Dan Zaharevitz (NCI) for support and molecules. We thank Mark Calleja and colleagues in Earth Sciences/NieES for advice on Condor.

An interactive XML version of the paper will appear on the proceedings CDROM. More information and demonstrations of the WWMM are at `http://wwmm.ch.cam.ac.uk`.

References

[1] "CML Schema", P. Murray-Rust and H. S. Rzepa, J. Chem. Inf. Comp. Sci., 2003, 43.

[2] "STMML. A markup language for scientific, technical and medical publishing", P. Murray-Rust and H. S. Rzepa, Data Science, 2002, 1, 1-65.

[3] Chemical markup, XML, and the World-Wide Web. 3. Toward a signed semantic chemical web of trust. Gkoutos, G. V., P. Murray-Rust, H. S. Rzepa, and M. Wright, *J. Chem. Inf. Comput. Sci.* 2001, **41**, 1124-1130.

[4] The World Wide Molecular Matrix - a peer-to-peer XML repository for molecules and properties. P. Murray-Rust, R. C. Glen, Y. Zhang and J. Harter. 163-164 “EuroWeb2002, The Web and the GRID: from e-science to e-business", Editors: B. Matthews, B. Hopgood, M. Wilson, 2002 The British Computer Society.