Chemical Structure Access & Representation

Authors

Dr. Juan Esteva Intelligent Solutions and Department of Computer Information Systems Eastern Michigan University juan.esteva@intellsolutions.com	Wendy L. Sharp Intelligent Solutions wendy.sharp@intellsolutions.com

The Chemical Exchange Problem

Storing chemical information in a computer is not a trivial task. Many different approaches and formats have been used. Most that worked at all are still with us. For instance, there are between 30-40 important chemical formats—managing them is a formidable and expensive task.

A small part of the problem is that arbitrary methods are used for encoding data, e.g., double bonds can be represented as the integer 2, the real bond order 2.0, the symbol "=", the enumeration DOUBLE, and as repeated connections.

A bigger part of the problem is that most chemical file formats contain information which is meaningless except in terms of a specific program, e.g., "tautomeric", "ring-double", "exo-double" and "fragment double" bonds

The biggest part of the problem is that different programs that process chemical information use different underlying models, e.g., in ab initio or M.O. programs, the idea of "bond" isn't a particularly useful concept. To be useful, we must provide an accurate method for representing the underlying data model.

Approaches that don't work well

Various attempts have been made to solve the problem created by the plethora of chemical file formats. Experience has shown that the two most common approaches don't work well. They haven't solved the practical problem and they keep being rewritten.

Comprehensive file format converters

Software is provided which converts files from one format to another. This is usually implemented by creating "readers" and "writers" which share a common data structure or format.

This approach works well for encoding problems only, i.e. where representational issues don't exist. Such systems are inevitably reactive (must be modified as formats evolve) and usually either inaccurate (there is no central authority) or LCD (lowest common denominator).

"Kitchen sink" formats

An all-encompassing format is proposed which purports to represent every possible kind of chemical information entity.

This approach has all the failings of the previous approach, plus it complicates the problem by introducing yet another format. Furthermore, the new format is so complex that a comprehensive reader/writer can't provide a universal interface because it is prohibitively expensive in size, complexity and support.

Therefore, a solution was sought that would have the following representational characteristics:

- Provide a common representation that provides a base functionality for atomic, molecular and crystallographic information and allows extensibility for other chemical applications.
- Allow containment of chemical information components within documents.
- Reduce the information loss when using different legacy files.
- Carry any desired chemical ontology

In addition this solution would need to provide the functional transactional characteristics:

- Provide classes and methods to facilitate component searches.
- Provide classes and methods to facilitate property searches.
- Offer separate class for Cartesian coordinates.

Figure 1 illustrates the general use case scenario.

CSAR

We have developed a model which makes use of the representational and translational capabilities offered by the Chemical Markup Language (CML)—an application of XML, the eXtensible Markup Language—and complement them with classes that facilitate transactional operations such as searches and creation of collections as described below in Figure 2.

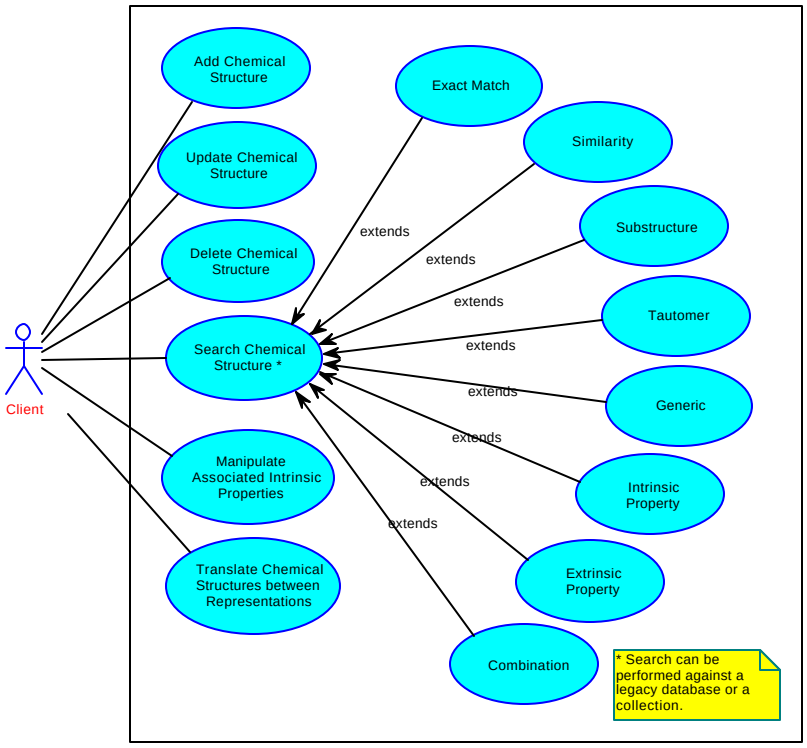


Figure 1 General Use Case

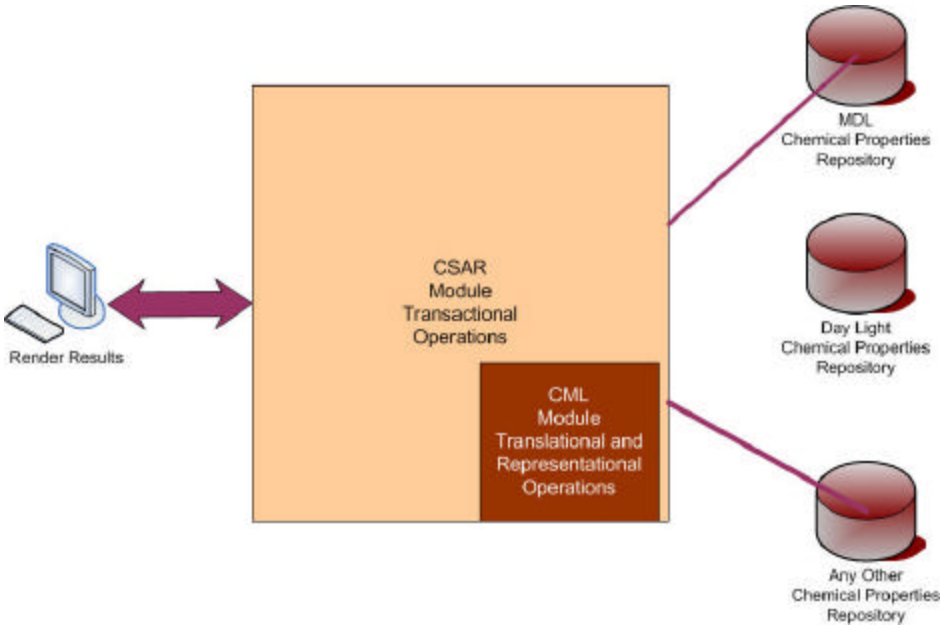


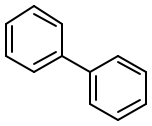
Figure 2 Chemical Structure Access and Representation

Sample Use Cases Solved by CSAR

1 Structure search

A using ISIS/Draw to sketch a molecule for a substructure search of a Daylight database

i simple structure (e.g., biphenyl)

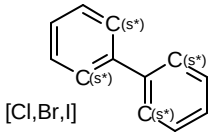


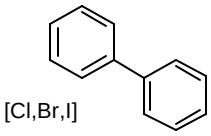
ii More complicated structure (disconnected fragments; atom lists; substitution counts; charges, etc.)

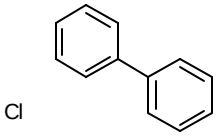
a two fragments (twofrag.mol)

b two fragments + atom list (tfraglist.mol)

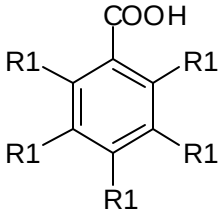
c two fragments + atom list + substitution counts (tfsub.mol)

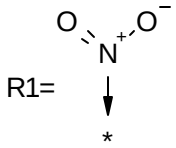




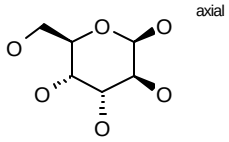


iii R Group Query





iv S-Group data



B using the Ertl Java editor to generate a SMILES to search an MDL database

C browsing the hits

i Browsing regular structures

ii Browsing structures with polymeric constructs

iii Browsing structures to which user does not have rights

2 Use ChemSymphony to search Web-based Available Chemical Directory (ACD) system for suppliers of a given set of structures
   A superstructures of aromatic acid chlorides – very simple substructure to find a large class of compounds (aromacchlor.mol)
   B p-nitrobenzoic acid – search for a specific compound (pnitrobenz.mol)
3 Looking for similar compounds
   A ISIS/Draw front-end to Unity (or RS^3) database back end
   B ChemSymphony front-end to MDL database
4 Registration
   A using ISIS/Draw to generate a molfile for registration into Daylight (convert to SMILES for direct chemical registration, as well as saving the molfile to an Oracle field)
      i Simple structure; everything in molfile translates to SMILES
      ii Complex structure (charges, valence) but properties do translate
      iii Parts of the structure (brackets, S-Group data) do not translate to SMILES
   B using ChemSymphony to generate structures for registration to a Unity database

Conclusion

A revolutionary approach describing the management of molecular information, chemical structure access, and retrieval processes has been described here. The specifications consist of two components, namely the Chemical Markup Language (CML) CMLCore and the Chemical Structure Access and Representation (CSAR). The goal of the proposed specification is to make use of the representational and translational capabilities offered by CML and complement them with classes that facilitate searches and allow the creation of collections. Situations that are not catered for explicitly can be addressed by extending the types and using the conventions described in this document. The standard was developed starting from the practical need to represent complex bodies of data in a natural way. Existing practice and terminology is used wherever this was available and practical.

References

Peter Murray-Rust and Henry S. Rzepa, “Chemical Markup, XML and the World-Wide Web. Part II: Information Objects and the CMLDOM” J. Chem. Inf. Comp. Science, 2001, 41, 1113.

Juan Esteva and Wendy L. Sharp, “Chemical Structure Access and Representation,” OMG Document 2004-08-3.