Position Paper for W3C workshop

Cambridge MA, 27th-28th October 2004

Babel-fish in Cheminformatics

"If only there was a Babel-fish for each scientist in every domain"
Not attributed to Arthur Dent, Hitchhiker's Guide to the Galaxy

ABSTRACT

There are several initiatives for standardisation in Life Sciences, notably LSID, OWL and RDF. Perhaps less well know are initiatives from the Cheminformatics working group from the Life Sciences Research group of the OMG. Whilst bioinformaticians have been blessed with virgin territory for the nurturing of interoperable tools, applications and services, the cheminformaticians have wrestled with legacy formats, vertical applications and bespoke systems for almost two decades. There has been no perceived business need to take advantage of technologies such as XML (or so vendors would have you believe) as mission critical data has resided in proprietary silos since the 1980's, unable to be transferred to open formats due to fear of loss of information.

The information bottleneck in Cheminformatics is the standard representation and description of small molecules. The CSAR initiative will be discussed in detail elsewhere, but this paper indicates the clear business benefits of having a standard set of semantics to describe compound collections, the lifeblood of any life sciences company involved in therapeutic research.

USE CASE BACKGROUND

There are over 8 million commercially available compounds from a disparate and varied source. It is estimated that only 2 million of these chemical structures are unique, with a large proportion being structural duplicates whilst having greatly varying attributes.

Any life science company involved in the development of novel chemical entities (NCEs) will routinely purchase compound collections from a range of suppliers for their own use.

The compound collection data is used routinely for "virtual screening" where chemists will use computational methods to analyze compound collections and select suitable sets of compounds to purchase.

THE BUSINESS CASE FOR SEMANTICS IN COMPOUND COLLECTIONS

A standard representation of the attributes of commercially available chemical substances would enable compound suppliers to provide ***consistency*** on values such as price per unit weight, purity, name, synonym, delivery time and physical state.

The compound collection data is typically sent from commercial suppliers on CD-ROM in ASCII format, usually as a SDF (Structure data format) file. There is ***no consistency*** maintained for the attributes of the data between compound collections. For example, attributes such as price, compound name, purity, weight and state ***differ greatly*** in their usage and meaning. These attributes are distinct and separate from the chemical structure representation.

The step of normalizing compound collection data is ***time consuming***. A typical set of CD-ROMs from 5 commercial suppliers takes an information scientist ***2 weeks*** to normalize and then convert into other formats suitable for import into databases.

In an attempt to provide compound collection data “pre-normalized”, several vendors provided aggregations of compound collections for an ***annual fee***.

There are in excess of 400 commercial compound suppliers who provide their data to a vendor such as MDL. This is then converted to SDF or RDF (Reaction data format) and sold as the ACD (Available Chemical Directory). These file formats limit the scope of representation of the compound collection data. There are other vendors such as Accelrys and CambridgeSoft who provide this data in their own application-specific format that is not compatible with the SDF or RDF format.

The major disadvantage with compound collection data provided by aggregators is that the ACD information is at least ***one year out of date*** by the time it is available for use.

So the business benefits of a standard set of semantics for compound collections is clear – save time, save money, share up-to-date accurate information.

REQUESTING PROPOSALS FROM INDUSTRY

In an effort to address this problem, requests were made for proposals defining data structures, which would allow commercial compound suppliers and purchasers to more readily exchange data. These standards were intended to form a common basis (framework) upon which services related to compound collection data could be built.

The scope of the RFP was limited to definition of data structures in support of collection, storage, retrieval, management, curation, communication, and analysis of compound collection data.

Any proposal submitted should seek to present a standard representation for compound collection data from commercial compound suppliers. This would include well-known suppliers such as MayBridge, Asinex and Specs as well as smaller specialist suppliers such as Peakdale or Bachem. The scope included compound collections known as “natural products” (compounds found occurring in nature) as well as small molecule collections known as “drug-like” (compounds that are man-made).

Internal proprietary compound collections owned by individual Life Science research companies were outside the scope of this RFP. The proposal scope did not extend to user interfaces or visualization services.

ISSUES

Proposals should have discussed how a suitable XML schema representation would facilitate the sharing of compound information between compound supplier and customer.

Proposals should have discussed how this representation may be used in other relationships, such as online ordering, database searching, compound brokerage or inventory systems.

Proposals should have shown how this representation may have been used by ***purchase order systems*** such as BAAN, SAP and Sage.

Particular care should have been given to discussion of how this representation may be used to facilitate ***real-time updates*** on compound collection data held by suppliers, such as ***compound stock availability*** or ***discontinued stock***.

It was envisaged that one potentially useful application of a suitable XML schema representation is in the standardization of the compound collection data held by each commercial supplier. This would allow ***real-time updates*** of data held in a relational database that could then be made accessible to the compound purchaser, facilitating the ***automatic ordering*** of compounds from an ***online web service***.

OUTCOME OF RFP

The result? Not one single letter of intent was submitted, despite a lot of initial interest and promise of involvement from both compound vendors and software suppliers. Initiatives such as this need a critical mass of support from big customers, not just small biotechs with limited budgets! This paper challenges the W3C to champion chemionformatics initiatives (the forgotten cousin of Bioinformatics) and help to bring vendors to bear on one of the biggest problems in Life Sciences research!