

# Semantic Storage of Chemical Data

Taylor, K.R.; Gledhill, R.J.; Essex, J.W.; Frey, J.G.  
School of Chemistry, University of Southampton, SO17 1BJ

## Abstract

Chemical data storage is inherently information inefficient, discarding many minor details to ease the burden of handling such data. The application of semantic web technologies and a triplestore allows the practical capture and referencing of a huge amount of metadata on each datum, thereby supplying the means to track authorship and veracity of properties within the database. This augments high throughput simulation, quality control and QSAR/QSPR analysis by minimising the loss of vital provenance when the data is created.

## Data Mining in a Chemistry Context

Chemical data is nearly impossible to describe with just a simple number. There are hundreds of little details that provide the meaning for a number by giving it context; experimental method, ambient conditions, authorship, and start and finish times for the experiment are all potentially informative. Some of these details appear superficial in the context of a database, but their value is readily apparent in quality control and more thorough examination. These things are normally discarded when data is collated owing to variability and the difficulty in describing them formally, thus making such databases easy to create but stripped of valuable information. A common tendency in science is to publish imperfect data, because to be too stringent will result in very few publications and hide potentially useful information away. In consequence, methodology and the fine detail are even more vital, and yet they are neglected in favour of the semantically void supplementary text field.

The chemical information world is also quite hostile to data mining. Intellectual property is closely guarded and the community interested in bulk data is small compared with the community interested in very specific data. Most presently available databases have interfaces that presuppose particular kinds of queries and preclude more interesting questions. This typically manifests in the form of web pages and query forms that a user has to manually complete, working within the constraints of the only interface available. If one database does not provide the answers, another must

be tried with its correspondingly unique interface and unknown content. In this way, many online databases are very like the paper libraries of centuries past. We have access to many of them, but each must be searched independently using whichever indices they have included, and then we must somehow consolidate the results from each.

Consolidation is easy when we are interested in only a few hits, as is usually the case for synthetic chemists, but an utterly hopeless task when mining large data sets. Each set must be filtered, cleaned and altered automatically before it can marry up with any other set, and it all has to be done without common data or file formats. There is enormous scope for improvement in this area if we are to meaningfully analyse data without first spending several months cleaning results one entry at a time.

## RDF for Chemical Data Storage

RDF<sup>1</sup> provides a means to describe data and its metadata together, and when coupled with a triplestore engine it becomes a rapid access database suitable for our purposes. By sacrificing some of the speed of relational databases, we gain massive flexibility in the ways we structure our data, in addition to not being constrained by the strict design limitations of a relational system. RDF makes the relationships between points explicit rather than making holes to drop the data into, and so removes all the assumptions normally found in spreadsheet-like storage.

Here we demonstrate that this approach applied to chemistry creates an intuitive data structure that parallels our own way of think-

ing, hence making investigation of the data and creation of software more simple, while bringing chemistry data storage up to date. The ease with which RDF copes with hierarchical information makes the design process much more straightforward than formal relational database design, while the possibility of cross-linking between data points makes it simple to relate properties to each other and create provenance chains. It is also easy to re-define our data structure to incorporate new changes that we may decide upon, making it a suitable approach to storage in a development environment. Researchers will not always be able to specify their needs well enough for conventional database design processes. The arbitrary level of abstraction possible in RDF could be attained with a relational database but would suffer in terms of working with and understanding the data structures, nor would it provide the portability of serialised RDF.

Figure 1 shows how we choose to encode chemical data relating to one molecule. Objects are marked as ellipses, the arrows show how predicates link objects together, and rectangles are literal values. Literals are the actual data in text form, rather than another RDF object, and are the values that we need to store, while the rest are simply the meta-data that explains what the literals are. Any combination of ellipse or rectangle and the connecting arrow represents an RDF triple. Cardinality is noted on the diagram to help show where the structure can expand, and which parts necessary for a complete description of a single property. A selection of simple state independent properties are held apart at the top of the hierarchy, where they are easily reached for selecting particular molecules. Pivotal amongst these is the InChI<sup>2</sup>, a sort of URI for molecules that allows us to combine records from different databases using a code derived from their structure. Below this are kept the properties of most use to us along with their complete provenance information. Such properties include measured values from the laboratory (melting point, density) and references to files of data which we need not reproduce in RDF such as absorption spectra and molecular geometries. Data files described in our system gain value in the sense of the normal use for RDF, that is describing the content without replicating the data, while supplementing the content with knowledge of how those files were produced. A grossly simplified example of how this schema operates for a single property is shown here in RDF/XML format:

```
<Molecule>
  <has-inchi>1/C6H15N03/...</has-inchi>
  <has-property>
```

```
<Property1>
  <rdf-type><MeltingPoint/></rdf-type>
  <has-quantity>
    <blank-node>
      <has-value>14.4675</has-value>
      <has-unit>
        <Celsius>
          <power-of>1</power-of>
        </Celsius>
      </has-unit>
    </blank-node>
  </has-quantity>
  <has-source><Database1/></has-source>
  <created>2005-05-14T12:41:02</created>
</Property1>
</has-property>
</Molecule>
```

A more complete example can be found at [www.soton.ac.uk/~krt1/escience/example.rdf](http://www.soton.ac.uk/~krt1/escience/example.rdf) and rendered using the W3C RDF validator found at [www.w3.org/RDF/Validator/](http://www.w3.org/RDF/Validator/).

Although it may not be immediately evident how all this additional detail is useful, the benefits are manifold:

- Normal data mining operations are possible eg. Select out all values for this property
- Tightly defined subsets can be selected on any and all criteria eg. author, classes of properties, date, ranges of values
- Data and history are explicit, so the data remains meaningful in the future when the authors have moved on.
- Value and context are distinct, easing computation using the values.
- Detail is sufficient for different kinds of data to coexist eg. predicted versus measured values.
- Provenance information allows value judgements on similar data with differing heritage.
- Provenance information is also suitable for tracking and repeating workflows. Reproducibility is of vital importance in all scientific research.
- Assertions can be superceded but are not deleted and hence recoverable for re-evaluation.

Clearly the use of such a verbose storage method as the basis for a database is wasteful and inefficient. RDF, like all kinds of XML uses enormous amounts of space compared to binary files, while triplestores are generally considered slower than relationally structured databases. These views cannot be denied but the question is whether the benefits outweigh the corresponding performance hit.

Within the domain of Chemistry we are able to show that a meaningful bulk of data can be captured and utilised on commodity computer hardware. Ten million triples is quite reasonably achieved on an entry level desktop PC using AKT's<sup>3</sup> 3store program coupled with MySQL.



One property typically requires around thirty triples to describe, while our data set consists of ten thousand molecules with at least six properties, amounting to two million triples. In a worst-case scenario we can extend this to include up to three million molecules for which we have structures but few or no properties, and would strain existing triplestores beyond proven capacity with several hundred million triples. At these sorts of scales, assertion into triplestores is a significant issue with multiple millions of triples taking days to assert and this is the chief limitation of the approach. It is conceivable that specific hardware could greatly improve its scope for application in the pharmaceutical area where compound libraries and data exist for millions of molecules. Our particular application deals with formulation and development of QSPR (Quantitative Structure-Property Relationship) models on thousands of molecules, and this is where our method is of most use. QSPR models are typically built by regressing on large numbers of variables in an iterative process where initial models are tested and rejected or improved. Our storage method should greatly ease this process and make validation or explanation of anomalous data points significantly quicker.

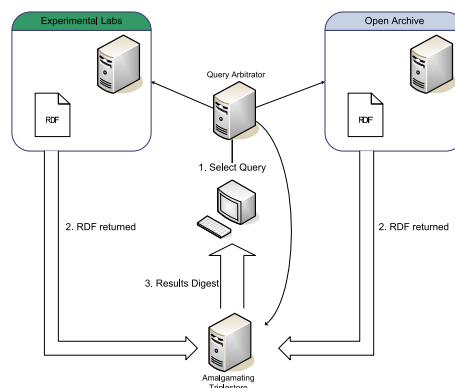
## Further Implications

RDF and the Semantic Web are typically concerned with aggregating knowledge from all across the web without needing to have a centralised store of all that information. A byproduct of using the same technology to describe chemical data is that distributed chemical knowledge becomes feasible and even desirable. As long as RDF conforms to the same schema, it can be transferred from triplestore to triplestore effortlessly and so we can construct arbitrary data sets based on our requirements without needing to handle every piece of knowledge in existence. This helps to counter problems with scalability limits on existing triplestore software. Each triplestore can be asked the same queries using common languages such as SPARQL or RDQL and the answers can be returned in one of two ways. Plain text can be returned should we be in dire need of specific numbers, or instead we can have parcels of RDF sent to us that tell us the things we wish to know, as well as the supporting information that we can explore at our leisure. Just as when creating our semantic database, we do not discard useful information simply because it does not immediately plug into an equation.

Figure 2 shows a possible arrangement whereby several independent data sources can operate together without needing to consider synchronisation, or operating some centralised repository.

This has clear parallels with present efforts towards publication at source and attempts to make data more freely available. A collection of compatible triplestores can store data from all sorts

Figure 2: Distributed Triplestore Arrangement



of areas of study, and using a common schema allow a degree of data discovery. If we are interested in a particular compound, we can ask for all properties, or classes of properties and see what our peers have to offer. This may have significant benefits in cheminformatics where a great deal of time is used preparing structures and drawings of molecules when the chances are they already exist. The same is true for the chemist in need of reference material, whose time can be saved through avoiding tedious library work. The data can differentiate itself so we need not separate it out into distinct databases which we must query each in turn.

## References

- [1] World Wide Web Consortium, Resource description framework <http://www.w3.org/rdf/>, 1997.
- [2] International Union of Pure and Applied Chemistry, International chemical identifier <http://www.iupac.org/inchi/>, 2005.
- [3] The AKT Consortium, Advanced knowledge technologies <http://www.aktors.org/>, 2004.