ChemConnect2017: Enhancing Combustion Information and search through semantic relationships

Edward S. Blurock[1]

1. Blurock Consulting AB, Lund Sweden, http:www.chemicalkinetics.info

ChemConnect2017 is an advanced state-of-the-art combustion database and data repository, derived from mechanistic, kinetic and thermodynamic chemical data (both experimental and modelling), organised into a network of interconnecting concepts.
Through the parsing of the data sets, the pieces of information within data sets not only available to state of the art keyword searching algorithms, but through the use of semantic web techniques, also provides interconnections between independent data sets for efficient data exchange and comparison.

1. Intelligent database through added semantic

A key aspect to 'intelligence' is to be able to derive, extrapolate and reason about concepts beyond the rudimentary data at hand. The goal of ChemConnect is to go far beyond being a flat repository of datasets and data points referenced by a few keywords and create a network of semantically enhanced data. The key is to use ontologies, a 'semantic web' tool used to make concepts cognitively accessible to humans and algorithmically accessible to computers. A given ontology not only gives a common (standardized) set of tags with a given domain, but also provides relationships of the concepts within that domain to provide a framework for reasoning. The backbone of an ontology is a taxonomy of concepts (much like the classification of molecular species into isomers, hydrocarbons, functional groups, etc.). In addition, having a 'common' set of semantics and terminologies, role of standardised (accepted) ontological descriptions (similar to the introduction of SI units, everyone talks the same language) increases the potential for the interconnection of previously independent concepts (see the next section for an example).

2. Network of Data Relations

Through the data representation as a relationship, a network of interconnected data concepts is created. A single relationship, within a Resource Description Framework, RDF, has the following form:

subject (keyword) —> Relationship —> object (keyword or data structure)

The power of such a network of connections is found in passively (meaning the creator does not have to have knowledge of other similar items in the database) connecting objects in the database.
For example, in ChemConnect, the subject is an individual keyword, such as a species name or even a reaction (in canonical form). The object in this relationship, which can be another keyword or a more a data structure of information. The relationship is how the object and the subject are related. An example relating a species name to an isomer keyword could be:

ic3h5chcoch3 —> isIsomerOf —> c6h9o

This relationship represents one link in the network and connects to other pieces of data if the object or subject keywords match. Though the isomer relation, the isomers ic3h5chcoch3 and ic4h7coch2 are linked.

ic3h5chcoch3 —> isIsomerOf —> c6h9o <— isIsomerOf <— ic4h7coch2

corresponding: Edward S. Blurock, Blurock Consulting AB, edward.blurock@gmail.com

Information between mechanisms, in this case a mechanism from LLNL (PRF) and a mechanism from Princeton (nC7H16), can be linked through the same species label:

LLNL#PRF —> MechSpecies—> LLNL#PRF#ic3h5chcoch3 —> Species —> ic3h5chcoch3
<— Species <— Princeton#nC7H16#ic3h5chcoch3 <—Mechpecies<— Princeton#nC7H16_red52

As in this example, searching in ChemConnect2017 is traversing the network of related concepts to find the information that is needed. This example also illustrates an extremely useful and important property of representing data in this way.

3. Organisational structure of the database

In ChemConnect the conceptual relationships provided by the ontologies are used in several capacities. This allows for a more flexible use and interfacing of the items within the database. The implementation of the database and the interface with the database is 'driven' by the ontological description. Here are several brief examples:

- **Definitions**: Classification of the individual data items and their relationships with each other.
- **Tagging (meta information)**: Tags for the data items and tags for the relationship between the data items are
- **Searching:** The classifications within the ontology of data items allow the identification of search rules for more effective search. For example, if the keyword is identified as a molecule, then the focus and use of the other keywords will be with molecular search.
- **Interface:** The classifications within the ontology provide information about how the data item should be presented (including allowing for input and modifications) within the graphical user interface.
- **File interface**: The classifications and descriptions within the ontology allow for more flexible interpretation and output of text forms of datasets. This includes the use a different tags for the same data item.

The following sections show further important aspects within the organization of the ChemConnect database.

4. **Use of standard ontologies**

It is important to note that any individual community is not alone in terms of trying to give a semantic context to the diverse items on the net or in a database. In addition, the key to efficient communication and finding connections being concepts is a common language. The World Wide Web Consortium (W3C) has made it a priority to establish established standards and in this case standard ontological descriptions of objects and concepts.
As the following sections elaborate, ChemConnect is using standard ontologies to describe key relationships needed for a database.

4.1 Units

All measurements have units. Though there is a 'standard' set of units, the SI units, not all data conform, mostly for historical or 'common practice' reasons, to these units. For example, though the SI units of every is Joules, it is common practice to use a variety of alternatives, such as calories, or even (for particularly large values) kilocalories or even kilojoules. This puts a requirement on a database to handle the multiple, through unit conversion equivalent, ways to represent the actual numerical value of the measurement.

This becomes particularly important when searching. For example, in searching for a temperature range, the user might use centigrade instead of the SI units of kelvin. It is important to note that this is a common problem within all of science, meaning that others have attempted to solve this dilemma.
ChemConnect approaches this problem with an ontological view. Within the literature, an ontology, the QUDT ontology, exists that describes the relationships between units. The system of units is represented by a hierarchy. In Figure, the Unit class, the class of all units, is first divided into three domain hierarchies. The lowest domain hierarchy, for example, Thermodynamics, this is the specific unit class. This represents the class of the the measurement itself. For example, a measurement could be a temperature 'temperature'. Within this class, there are specific examples temperature units, with their

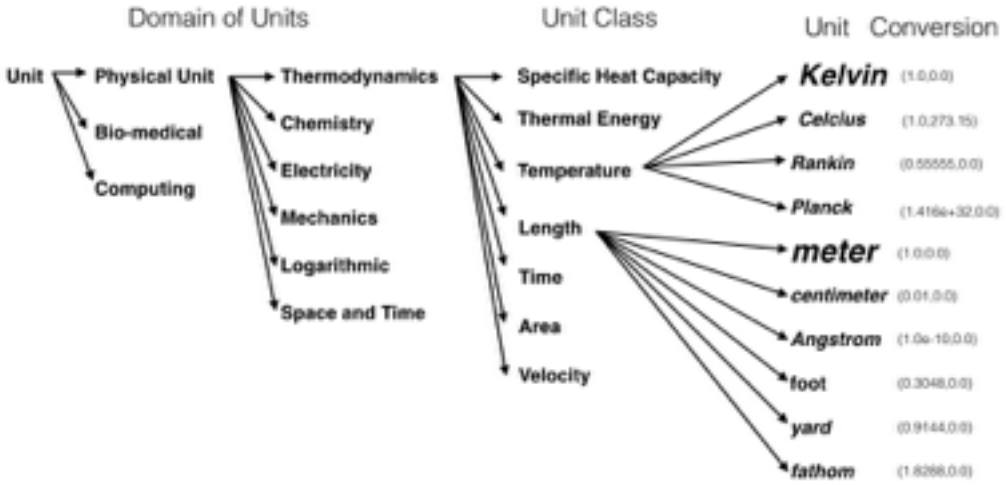

Figure: An portion of the QUDT ontology from 'Unit' class

corresponding units. The reference unit is the SI unit, for example, Kelvin. The conversion units, a multiplicative factor and an additive factor.
Within combustion, there is a special case of conversion, namely rate constants, where this scheme does not just depend on multiplicative and additive factors. The question of units, particularly the Arrhenius constant, follows unique rules and is dependent on the reaction it is describing. The exponential coefficient as a whole is dimensionless, but the Rydberg constant, temperature (always being Kelvin) and activation energy have to have consistent units. The term with temperature to an exponent n is, in end effect, dimensionless, but the Temperature itself must be Kelvin. The dimension of the Arrhenius constant is dependent on the number of reactants of the reaction being described.

4.2 Datasets, Catalogs, concepts and collaborations

The dataset, a set of data points of usually published data, is the central entity of the database as a repository. Typical published datasets are kinetic mechanisms or sets of experimental measurements. Datasets can also be raw experimental results.
ChemConnect goes beyond a simple repository in that the datasets are parsed and the elements semantically organised, using the principles of the semantic web. Each of the data elements are connected into the complete semantic network using RDF expressions.

The organization of datasets and the information within the datasets is modelled after the Data Catalog Vocabulary, DCAT, ontology. The datasets are organised in a hierarchy:

User —> member —> Organisation<— publisher <— Catalog —> Dataset

The User, Organisation and DataConsortium are linked publishers of the dataset and the catalog. Collaborations are represented by DataConsortium membership.
The Simple Knowledge Organisation System, the SKOS ontology, associates a set of concepts and themes with the dataset and catalog (italics are connections and boldface are entity):

Catalog —> *dcat:themeTaxonmy* —> **ConceptScheme** <— *skos:inScheme* —> **Concept** —> *dcat:theme* —> **Dataset**

The ConceptScheme and the Concept (from the skos ontology) are semantic description classes representing the types of datasets and catalogs.

4.4 Traceability and Workflow

No data is independent of other data points. There is a traceable flow of measurements and manipulations that produce the given data point. Experimental data and computations can be traced to their origins of the specific device, such as a rapid compression machine, or computation, such as a computational chemistry method. Both the experiments and the computations have parameters which characterise the particular device or computation. Often there are calibrations associated with the experiment. These calibrations can be for adjusting the device or establishing parameters for interpreting the measurement. For example, a base line for a spectrum should be established, peak assignments for a gas chromatograph, quantifying peak area with know quantities. In addition, there is often a series of manipulations of the data to produce the final published result. For example, in RCM, the experiment measures a pressure trace and this pressure trace has to be converted to an ignition delay time. This trace is particularly necessary in terms of determining the error bounds of the final value.

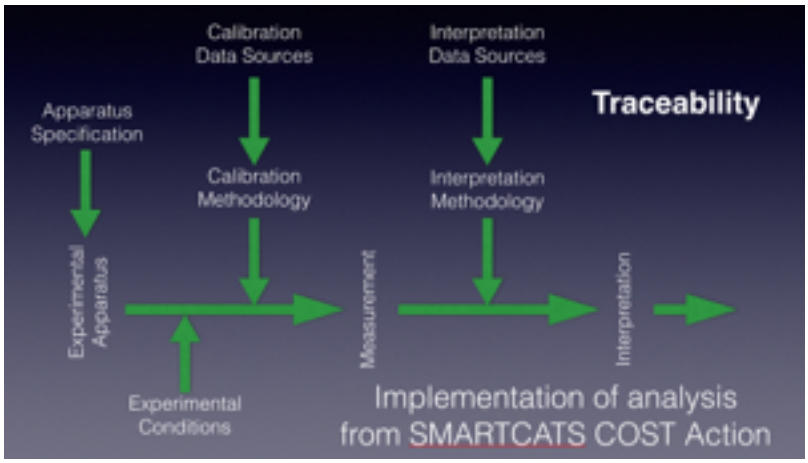

Figure 2: Tracing the path of experimental measurements from the apparatus to the final publication.

ChemConnect tries to capture this traceability by putting these relationships within the workflow diagram of Figure, which was established in the effort of the CM1404 Task Force on data. Once again, this type of workflow and traceability relationship is not limited to combustion experiments, or even the experimental sciences. One accepted (by the W3C) set of terms representing this relationship is the PROV (provenance) ontology:

> *Provenance is information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness*

The PROV ontology defines classes of 'Activities', which in our case can be ideas such as methods or measurements of devices. There are also the 'Entity' class which is the data that is generated (with the connection 'wasGeneratedBy') by the activity or is used (with the connection 'used') by the entity. An RCM measures the pressure trace to produce the actual pressure trace. This can be represented by:

RCM <- used <- measure pressure trace <- wasGeneratedBy <- pressure trace.

In addition, to measure the pressure trace, a set of operating conditions are used:

set of conditions <- *used <- measure pressure trace*

Thus using the provenance ontology, the traceable relationships of the data from measurement to publication can be exactly traced.

4.5 Ownership and access

The concept of ownership of data is not at all trivial and has many perspectives. One complex relationship is all the aspects of legal ownership which involves a combination of patents, copyrights, licences, prepriotory rights, not to mention the relationship between researcher and the respective organization of the researcher. The perspective of a ChemConnect uses the perspective relative to who generated the data, who can manipulate the data (including deleting) and who can see the data (under a search).

Associated with each data point is a tag pointing to a 'DataConsortium' class and who created the data point. The DataConsortium class specifies two lists of who can see the data and who can manipulate the data. These lists consist of keywords pointing to individual users of the database and registered organisations. A special DataConsortium class is the 'Public' class in which the database administrators have entered the data points.

The default DataConsortium is where the generator of the data has sole control and access to the data.

The reason for condensing the accessibility to a single tag is to simplify the search property. When a user does a search, the list of DataConsortium tags is collected. The actual search is performed separately for each tag.

5. Conclusion

The developments within ChemConnect illustrated in this abstract are working toward the goal of providing cutting edge technologies enhancing the availability of the vase amount of combustion kinetic information that is being generated. ChemConnect is not just a repository for datasets. Through parsing of the datasets, the individual data points are isolated. These data points are contextually analysed and put into conceptual resource description framework, RDF. The RDF relationships are enhanced by ontological descriptions of the description terms within the RDF.