The Very Open Data Project: Characterizing Combustion Kinetic Data with ontologies and meta-data

Edward S. Blurock*
REACTION, Lund, Sweden

Keywords: Combustion Data, ontology, XML, database, XML Schema, meta-data.

The movement of *Open Data* stems from a priority at the national, European and international levels to make scientific data, especially that that stems from public funding, freely accessible. Beyond political and financial considerations is that science thrives on interaction. With modern science, especially with the explosive use and availability of electronic media, this translates to sharing electronically data between groups. The goal of the *Very Open Data Project* is to provide a software-technical foundation for this exchange of data, more specifically to provide an open database platform for data from the raw data coming from experimental measurements or models through intermediate manipulations to finally published results. The sheer expanse of the amount data involved creates some unique software-technical challenges. One of these challenges is addressed in the part of the study presented here, namely to characterize scientific data (with the initial focus being detailed chemistry data from the combustion kinetic community), so that efficient searches can be made. A formalization of this characterization comes in the form of schemas of descriptions of tags and keywords describing data and ontologies describing the relationship between data types and the relationship between the characterizations themselves. These will be translated to meta-data tags connected to the data points within a non-relational data of data for the community.

The focus of the initial work will be on data and its accessibility. As the project progresses, the emphasis will shift on not only having available data accessible for the community, but that the community itself will be able to, with emphasis on minimal effort, will be able contribute their own data. This will involve, for example, the concepts of the 'electronic lab notebook' and the existence and availability of extensive concept extraction tools, primarily from the chemical informatics field.

The combustion community of researchers is fairly unique within the scientific sector in that is an extremely diverse set of researchers composed of, among others, engineers, organic chemists, physical chemists, theoretical chemists, computational modelers, physicists, mathematicians, computer scientists, modelers, experimentalists and various combinations of them. This diversity has consequences in not only terminology used within each community, but, relevant for this study, also has consequences in the data produced and managed.

The goal of the project is to make the entire range of combustion research data available. This means not just accepted *standard* values, but values that have changed, values under different contexts (for example, reaction rates and reactions as they are used in different mechanisms) and historical values.

In addition, more complex models such as detailed combustion mechanisms are not treated as a coarse grained single entity, but decomposed into fine grained data, such as individual species with their corresponding thermodynamic data, reactions with their corresponding reaction constants. The goal of the database should also not just be storage of data constants, but these constants are also stored in a form that can help in the analysis of, for example detailed mechanisms. Data is stored with several access points. For example, a species name listed in a mechanism can have a link to its isomer form (the only 'structural' data available in CHEMKIN form). This isomer link can then have links to other species labels in other mechanisms for, for example, comparisons. Furthermore, the species will have links to their use in reactions. Reactions in CHEMKIN form have links to the individual product and species labels and correspondingly their isomers. This interlinking can aid the combustion modeler in not only comparing the use of isomers in different

mechanisms. How do specific isomers react in different mechanisms? When the 2D structural data found in automatically generated mechanisms is added, then the possibility of linking specific isomers to 2D-structural data (or even 3D structural data if the 2D structure has computational chemistry origins) is available. In addition, pathway information within the mechanism linking the evolution of one species to another can not only help in analyzing and comparing mechanisms, but also in determining similarities between mechanisms.

A sampling of characteristics and relationships that should be characterized is that the descriptions:

- Should have the **Type** of data: the property being described (classifications/ontologies already exist): kinetic rate, temperature, entropy, pressure, …
- Should reflect the **specific property** the data point represents: property relative to a specific molecule, reaction, model, …
- Should reflect the **history** (time) and **origin** (research group, experiment, theoretical) of the data point
- Should simulate a **"social network" of sharing** information between groups: private data, group data, public data, proprietory data, ….
- Should reflect **interdependency** of data: data points are derived from others, they are used in other complex models
- Should reflect **quality of data**: temporary data (trying it out), optimized, modified, standard accepted data, use in a single/multiple/general models, …

From a software technical point of view, this interlinking of fine grained information at many levels poses new challenges in which traditional methods are not viable. 'Scalability' is a term often used now to describe methods which can handle the large amount of information that this project is proposing in the long term.

The aim of this project is to set up a ***service for the management of a social network of decentralized data*** for the researchers within the combustion community. The purpose of this study is to take into account the needs of data exchange within this diverse community, established through an extensive survey, and encapsulate it into a service for the community. The development and implementation of the service will be in intimate interaction within the combustion community, enhanced by activities of the European Community SMARTCATS COST Action CM1404.