Data Transfer: Towards a common standard

This document is meant to outline points of perspective and discussion points towards meeting the goal of setting up requirements and guidelines for a common standard for data exchange within the combustion community. This document is meant to be flexible and continually evolving through input of the partners.

The purpose of this document is to provide a framework and focus for discussion during the SMARTCATs mini-symposium on data transfer and for fact collecting afterwards. It is hoped that as you are reading this, perhaps something 'strikes a nerve' enough to warrant a response and discussion.

The next section outlines phase one of the goal of establishing common standards (this is meant to provide guidelines for those presenting their 5 minutes perspectives and for the discussion itself). The sections after give somewhat of a justification the phase one goals.

Phase One: Cataloging

The purpose of this phase is to provide the foundation for entrepreneurs developing (or planning to develop) tools using data or facilitating data exchange. The purpose is to set down basically what exists and what the requirements, expectations and desires of the community are.

The discussion of data exchange is (and has been) continuous and, for the most part, endless. In order to actually convert this endless discussion to concrete results, I suggest that expertise of the SMARTCATS be used to formulate a set of catalogs with respect to data exchange. An important aspect of these catalogs is a prioritisation that would be useful for the next phase of realisation of the aspects and objects expressed in this list.

1. **Perspective/Roles**: What type of users are involved in data exchange. These range from user accessing data to the generator of data to those who implement tools to use the data.
    1. For each role/perspective catalog a prioritised set of requirements, expectations and desires.
    2. Refine this list with respect to the interaction of roles..
2. **Data to disseminate**: What data, regardless (at the moment) of format, can and should be disseminated and at what level (private, shared or public) should be disseminated.
    1. For each apparatus or tool (experiment or calculation) , catalog the data that is generated from raw data to final result and prioritise this data in terms of usefulness and practicality of dissemination (whether for, for example, a software tool, database, supplementary publication information, or published results).
    2. When possible, the role of error analysis/determination should be included.
3. **Data formats**: There are formats that are being used within We are not alone. Considerable effort in related fields has been made that can be directly used by the combustion community:
    1. **General mathematical/computer science**: Catalog of general data types that have already been characterized, for example, simple numbers (including, for example, error bars and units), graphs (from continuous to noisy), surfaces, networks, etc.
    2. **Publication**: Representation and repositories of references
    3. **Chemical**: Chemical data from other disciplines, most notable, for example CML, for structural, thermodynamic, kinetic, calculation, etc.
    4. **Experimental**: Data processing from experiments and devices.

4. **Current Efforts/Projects:** This should be a catalog of projects or ideas that could be the basis for collaboration or build upon
    1. **Current project underway:** Projects in any form dealing with scientific data exchange. This could be specifically within the combustion community or relating to data that would be of interest to the combustion community.
    2. **Political discussions:** In this category I put discussions/projects/proposals that have been, could have been or are being discussed by larger authorities (EU, Combustion Institute, etc.) for a broader purpose. These could give the basis for cooperation or ideas to build upon.
    3. **Current proposals**: Current (and maybe even rejected) proposals that could be interesting for the combustion community

For the actual discussion during the SMARTCATS mini-symposium, discussion should be limited to the definition and elaboration of perspective/roles. In addition, discussion and the utility of current efforts and projects would be useful. For those giving 5 minute discussions, it would be particularly useful (to begin discussion) to emphasise these two points (roles/perspectives and current efforts).

I would ask that the cataloging of data coming from experiment and calculations be done offline.

The following sections are discussions on why and how this list was put together.

Perspectives/Roles

An important aspect of setting up requirements and guidelines is to differentiate the different roles that are involved in promoting data exchange. A particular researcher can in fact have multiple roles depending on the goal. It is particularly important to define these roles and acknowledge how these roles interact.

A preliminary set of roles are:

1. **User**: This is one who is interested in using the tools to acquire and use the data. In this role, the user is interested accessing data in a convenient and efficient way. The user is also interested in what data is available. Of course, this is means there should be a strong coupling to the generator of data. Sub-categories of users could be considered, such as researcher (both academic and industrial users), decision maker/manager, publisher/librarian (data managers and disseminators), etc. and even these could have different levels.
2. **Generator**: This is one who is actually generating data, both experimental and theoretical. The first focus is how much, in which detail and in what form the data should/can be disseminated. An important aspect of this is to make this as painless and efficient as possible so as to not generate more burden. The demands, from, for example, funding agencies, publishers and , not to mention the community as a whole, is slowly ‘forcing’ the generator to have a broader means of dissemination.
3. **Software/Database Developer**: This is the one providing tools for the user to access the information and the one who incorporates the generators data. From the user, the developer takes into account how and in what form the data can be accessed. The developer interacts with the generator by incorporating their data into whatever system they are developing. Both interactions require compromises on both sides in terms functionality, practicality and feasibility.

At the first level, each user, depending on their role, should catalog their requirements, expectations and desires with regard to data exchange purely in terms of their (selfish) goals and should prioritise them with regard to their importance. At the second level, each role’s prioritised list should be analysed by the other roles, in terms of burden, feasibility and practicality, on how it can be implemented within their role.

Which data should be disseminated?

There are two aspects to this question.

First there are many levels to data, from the raw data to process/condensed data that is to be published. All data points involve a process to their development. Ignoring (discussed in the next section) here the practicality of representing the data, the user and the generator should be concerned with degree of usefulness and accuracy of the data.

Second, for a particular piece or type of data the generator is also concerned with which degree of publicity and sharing would be useful and desirable and which degree is allowed due to proprietary reasons.

A question that is becoming increasingly important in the dissemination of data is reproducibility. With the increasing complexity of the generation of data (whether stemming form calculations or experiments), publication of the 'final' result and just a description using words (within the publication) of how this result was obtained is no longer enough. Researchers are asking for access to the data leading up to the final answer. How was the raw data interpreted? How does this interpretation effect the reliability of the raw data (error bars around final result)?

This brings up the question of how much of the scientific process needs (and can be) to be documented. In other fields, a related aspect of data collection is accountability, what data and especially when this data was collected can important when patents (translated to money down the line). A community that is particularly concerned with this is the pharmaceutical community and has lead to the development of electronic scientific notebook.

The bottom line here is that the data generator, in collaboration with the user, should catalog the data that can be shared and at what level it should be shared. Prioritising the usefulness of the data is also important.

For the experimentalist, with each apparatus, the data process chain from raw data to final result should be characterised and cataloged with each level being prioritised in terms of usefulness the community and at what level it would be useful to share to the community.
For the theoretical researchers, with each type of calculation, the same data process chain should be characterised and cataloged.
With this data, the software/database developer can prioritise its incorporation for calculation and dissemination. The users can also prioritise its usefulness in terms of their goals of using the data.

It is important (as discussed below) that this discussion is decoupled from how and in what format the data should be disseminated.

Role of data formats

Both the generator of data and the

The question of data formats has several important aspects:

1. **What data**?: This is a question for the communities point of view, to decide which data (as discussed previously) should be available.
2. **Data type**: This is more of a mathematical/computer science question. Data can range from a simple number to a complex graph or multidimensional surface. Both have Each has its problems. With a simple number, there is the problem of significant digits.

3. **Units**: This is a sub-category of data type. Due to the multidisciplinary nature of the combustion community combined with the historical usages of certain units, a wide range of units have to be accounted for. The generator of data and even the user of data would not be satisfied with a restriction to SI standard.
4. **References**: Very rarely is a particular data point from a single user. How and to what extent should references be incorporated into the data format.

Standardised formats, especially for the software/database developer, ease the incorporation of data into their systems. 'Forcing' the data generator to use these formats eases the incorporation of data.

Historically, the combustion community has adopted several 'standards', for example, CHEMKIN format, NASA polynomials and even that for transport data. But however convenient and useful they are, these formats are slowly placing unwanted restrictions on data. For example, in the CHEMKIN rate formation, expansion to non-Arrhenius behaviour is becoming necessary or expansion of the 7 coefficients of the NASA polynomials are also becoming necessary. As data evolves, the data format must also evolve. In addition, the words 'standard format' for, for example, CHEMKIN, has to be taken with a grain of salt, because looking at how different groups interpret this format can be vastly different. Such 'fixed format' machine readable formats restrict evolution of data and places addition burden on the software/database developer to interpret them.

Taking an extreme stance, the 'only' requirement for the incorporation of data for the software/database developer is that it be machine readable and well-defined. Though possibly inconvenient, software can always be developed to read any well-defined machine readable file regardless of format. Two practical examples of this principle are the following:

1. OpenBabel (www.openbable.org): Within the chemistry community, this software interconverts 'over 110 chemical file formats'. Here the emphasis is on the information that is available within the format rather than the format itself.
2. XML file conversion: Within the computer science community, with XML as the accepted method of internet data exchange, (open source and beyond) software exists to not only covert automatically and XML file to data structures within most (popular) programming languages, but software exists to facilitate the conversion of one XML format to another.

In addition, the data within the combustion community is not unique, data formats already exist, both at the level of 'data type', meaning the mathematical/computer science representation and in the form of what data, including units, references and chemical and physical data from other disciplines.

Arguments for XML formats

1. XML has already been accepted within a wide variety of disciplines, beyond computer scientists. This means that developed formats already exist that can be useful for the combustion community.
2. XML is a flexible format for data exchange shifts the emphasis from ***how*** the data should be represented to ***what*** data should be represented in the first place. This shifts the emphasis from computer science/mathematics to the field of application, i.e. combustion, where the expertise lies.
3. The **how** is reduced to cataloging a set of acceptable tag expressions. Even if the tags and even the order of the tags is not 100% standardized, , general (even public domain) software exists (for example, XSL conversion) to facilitate the conversion of one XML format to another.

4. For software/database developers, since XML is the accepted method of internet data exchange, (open source and beyond) software exists to not only covert automatically an XML file to data structures within most (popular) programming languages.

Current projects/efforts

There are efforts underway to tackle the problem of data exchange not only within combustion and other communities as well. It is obvious that the combustion community is not the only community that is faced with the dissemination problem so it is important for combustion related efforts take note of these outside efforts. And considering the multi-disciplinary aspect of combustion, recognising these efforts will avoid 're-inventing the wheel' for the discipline of combustion. It is also important, to reduce overlap, to make note of efforts within the combustion community.

There are several aspects of current projects and interaction are occurring at several levels:

1. **Data formats**: The data formats needed by combustion are by no means unique. There is considerable overlap with other disciplines. It is important, not only for efficiency, but also of the sake of standardisation that
2. **Repositories/Databases**: There are already considerable efforts both within combustion and in general to set up repositories and databases of scientific information. It is important to take note of these efforts and evaluate how and if they could be useful for the dissemination of data for the combustion community.
3. **Combustion Community**: There are efforts within the combustion community and related communities already that have dealt or are dealing with these issues. These should be made more publicly accessible to promote collaboration.