Report on the Workshop on **International Scientific Data, Standards and Digital Libraries,** Denver June 10-11, 2005.

I. David Brown

Summary

The workshop reviewed progress in establishing the conditions for the semantic web - what conditions are needed and how well different disciplines meet them. Several weaknesses of CIF became apparent. The structure of CIF does not make it readily accessible to other disciplines, it makes little provision for file management, the arrangement of parent-child links needs to be rationalized, and CIFs should contain information about their own history.

.

+++

I was invited to this workshop, run as part of the conference on **Digital Libraries: Cyberinformatics for Research and Education**, to talk about our experience with the Crystallographic Information File (CIF). Problems of human and computer communication between disciplines was one of the underlying themes of the workshop, and this was well illustrated by the problems of communication even between those attending. Words like *digital libraries, metadata* and *semantic web* were widely used but with different meanings. Fortunately some of the speakers took time to discussed the meanings assigned to these words and to make it clear how they were using them. Before reporting on items relevant to CIF, I will give my own understanding of the concepts behind these words as this will provide some background to what the workshop was about. I will illustrate these ideas with reference to the field of crystallography.

A *digital library* is an archive composed of documents that are stored in digital form, currently this means electronic storage. The documents are assumed to contain *data*, though in this document I will use the more precise term *information*, and for the purposes of this report I assume that each document is composed of a number of individual *items*. Unlike traditional libraries, digital libraries are normally distributed, i.e., the documents are not all held in one place but are stored at many different locations on computers that are linked by the web. The documents in a given library are normally related in some way, e.g., by belonging to the same discipline, and it is assumed that they are managed or curated as they would be in a traditional library. Curation may be as minimal as allowing people to upload files in the appropriate format, or may include extensive quality control over the content of the documents. Libraries differ from web pages in that the latter do not contain a managed archive of documents. An example of a digital library in crystallography would be a crystal structure database (though not all of these are currently accessible from the web). Each crystal structure report would constitute a document and the items would be the values of the individual items defined in the CIF dictionaries. The 'digital data', i.e., the information in the document, may be quantitative (e.g., atomic coordinates) or qualitative (e.g., the colour of a crystal or the name of an author). In scientific digital libraries a document may contain observations derived from a scientific experiment (controlled observations), observations of a natural phenomenon (uncontrolled observations), theoretically derived information, additional information needed to provide the context (metadata - see below) or some combination of these. Some people made such distinctions between the different types of information stored in digital library, but there was a lot of confusion.

The *semantic web* is a term that describes an organization (or the set of organizations) that seek out and assemble information gleaned from different digital libraries. The term is used in a rather wide range of senses. At one end of the spectrum is Google which matches the text of a query with text found in the document and provides the URL of the document.. At the other end of the spectrum are programs (not necessarily yet written) that match the meanings of the query words with the meanings of the items in the documents, taking the search one stage further than just matching text strings. *Rules* are then used to combine the information retrieved from different documents to infer new information. This constitutes added value usually referred to as *knowledge*. Knowledge is information not stored directly in the libraries, but inferred from the information found there. Such rules-based software might, for example, be able to infer the structure (knowledge) of a given crystal based on information (data) about its composition and space group held in the digital libraries. Or, in a more ambitious example, automatically be able to generate rules for determining how the distortions in benzene rings depend on the substituents present, based on information about crystal structures found in a large number of different documents drawn from different libraries.

The term *metadata* is usually defined rather unhelpfully as 'data about data'. The ambiguity of this definition wonderfully reflects the ambiguity in the way the word is used. The meaning assigned to *metadata* appears to depend on the context. It usually refers to information used to locate a document containing the requested information. At the most basic level this includes items such as the citation, keywords, authors etc. In more complex cases it might include items that provide context, such as the temperature at which a structure was determined, or the way in which a crystal was prepared. It could also include such basic information as the formula or the presence of particular types of chemical bond. Clearly one person's data is another person's metadata; the boundary between the two depends on the context.

As mentioned above, there is an attempt to move beyond simple text matching to being able to capture the meanings of words. The favoured way to do this is to use a thesaurus - a list of words having the same or similar meanings. Such thesauri can become quite elaborate, and a program used to construct a query might provide a drop-down menu if it were necessary to clarify the meaning a given term. For example, if one is looking for crystal structure reports that include the data, the search program might ask if the 'data' in question was a powder pattern, a list of structure factors, Bragg intensities, locations of Bragg peaks, lattice parameters or atomic coordinates.

In the context of the workshop, CIF looked like a bit of a dinosaur. True, we were in the game before anyone else, but in the beginning we were not entirely sure where we were going and we compromised in many places in order not to alienate our community by insisting on rules that seemed complex and petty to the uninitiated. However, as the semantic web slowly takes form, these compromises are coming home to haunt us. Further, CIF is tuned very much to the needs of the crystallographic community and makes few concessions to other disciplines. Chemists for example might wish to know the typical length of a given bond but they would find the CIF dictionary confusing and would have difficulty locating the data names needed to retrieve a particular item. For them CIF needs a thesaurus that would identify a 'bond' or 'bond length' with the data names in the geom_bond category. If CIF also included the proposed 'methods' which allow a bond distance to be calculated from the atomic coordinates, looking for a 'bond length' might retrieve the information even if it was not explicitly present in the document..

The point was forcefully made by one speaker that one needs to distinguish between the semantics, i.e., the items that appear in the documents (e.g., as defined in the CIF dictionaries or XML schema), and the syntax, the rules used to assemble the items into documents (e.g., XML). The semantics are relatively stable and have a long life, whereas the syntax can change rapidly depending on fashion and the development of new network protocols. This suggests that there is permanent value in the CIF dictionaries and there is no reason why we should not continue to use CIF in more or less its current form for the transfer and archiving of structure reports within the discipline, but if we wish to be part of the semantic web, so that people in other disciplines can locate and use this information without employing their friendly neighbourhood crystallographer as an intermediary, we need to think carefully about how others may wish to access and use the information we store *(interoperability* is what Humpty Dumpty might have called it).

A few other ideas occurred to me as I listened to the talks. The need for cross-discipline interaction both at the syntactic and semantic levels was obvious if we are not to retreat into our own shell. There are a number of weaknesses in our current structure.

> 1. It is clear that each document should carry its history with it. This means that in addition to giving the normal audit information (the creation date, the program and dictionary (and its URL) used to create the CIF), the CIF should include the same information for all its previous versions. Thus when accessing a CIF, one would be able to see that it was written by, say, the CSD who had obtained the CIF from Acta Cryst. E who in turn obtained it from the laboratory where the structure was determined. The dates when each of these transformations took place, as well as details of any changes made and the dictionary used at each stage, would also be available.
>
> 2. CIF has an ad hoc and barely existent provision for file management because STAR assumes that each data block is independent and needs to know nothing about other data blocks either within the same file or in other files.
>
> 3. Another area of weakness is the ad hoc way in which the parent-child relationships were developed without a clear vision of the structure of the linkages between the concepts. For example, should 'distance', 'atoms' and 'symops' be treated as properties of 'bonds' or as separate items whose relationship is implied by the fact that they are all grouped in the same category? The relationship chosen determines the way in which the parent-child relations are expressed and the way in which the software is written. UML (Unified Modelling language) and RDF (Resource Description Framework) are syntaxes that start with an abstract scheme (network) that makes explicit the relationships that are currently rather muddy in CIF. CIF would benefit from an analysis of these relationships.

It is fair to say that when CIF was originally developed we were pioneering a new way. It is not surprising that there are some loose ends.

The first day of the workshop presented the successes and problems of current initiatives such as the International Virtual Observatory Alliance, Data Standards in Biodiversity, Exchanging Technical Product Data (i.e., in commerce), the Geographical Mark Up Language (GML), Materials Mark Up Language (MatML), an ambitious initiative by the German National Library of Science and Technology, MathML, and of course CIF. The second day was devoted more to the problems of creating the software that can search and analyse the various digital libraries. Here it became clear that there was still a long way to go, though the group at Stanford (which produced Google) seemed to be well advanced in their approach to the semantic web.

In the conference which preceded the workshop Rachel Heery of UKOLN at the University of Bath gave an account of the distributed crystal structure library that Mike Hursthouse is establishing in Southampton. The aim here is to capture all the intermediate files generated during a crystal structure determination and to store these on web-accessible servers in-house. The query software could locate any document (structure determination) in any participating institution and the user would be able to review the whole process of structure determination and refinement (including access to the original diffractometer measurements), not just the summary that is normally available in published reports. There is at least one other similar initiative underway and we can expect further semantic web experiments by other groups.

The proceedings of the workshop will be published in CODATA's Data Science Journal. The texts of the papers presented at the conference (but not the workshop) were included in the delegates' registration packages as the Proceedings of the 5th ACM/IEEE Joint Conference on Digital Libraries: *Digital Libraries: Cyberinfrastructure for Research and Education,* Denver, June 7-11, 2005

David Brown
2005-06-20