Acta Cryst. (1998). D**54**, 1065–1070

The Globalization of Crystallographic Knowledge†

Peter Murray-Rust

Virtual School of Molecular Sciences, School of Pharmaceutical Sciences, University of Nottingham, England.
E-mail: peter.murray-rust@nottingham.ac.uk

(*Received 14 April 1998; accepted 9 July 1998*)

Abstract

The rapid growth of the World Wide Web provides major new opportunities for distributed databases, especially in macromolecular science. A new generation of technology, based on structured documents (SD), is being developed which will integrate documents and data in a seamless manner. This offers experimentalists the chance to publish and archive high-quality data from any discipline. Data and documents from different disciplines can be combined and searched using technology such as eXtensible Markup Language (XML) and its associated support for hypermedia (XLL), metadata (RDF) and stylesheets (XSL). Opportunities in crystallography and related disciplines are described.

1. The foresight of J. D. Bernal

Over 30 years ago J. D. Bernal foresaw the coming information explosion and the need to address it:

> However large an array of facts, however rapidly they accumulate, it is possible to keep them in order and to extract from time to time digests containing the most generally significant information, while indicating how to find those items of specialized interest. To do so, however, requires the will and the means. (Bernal, 1965)

Like many other visionaries, Bernal lacked like-minded colleagues and the technology to implement his ideas. It is only now that the world is realising the importance and generating the tools. The explosion is global, epitomized by the WWW, and many people no longer require convincing that the information age is upon us.

It requires new ways of thinking, and the message of this paper is to encourage crystallographers and their collaborators to think in novel ways. To quote from Bernal again,

> [we need to] get the best information in the minimum quantity in the shortest time, from the people who are producing the information to the people who want it, *whether they know they want it or not* (my emphasis).

† Some terms and abbreviations are deliberately not defined in the text but in a glossary at the end. This simulates the approach to the processing of linking and normalizing information.

Information is therefore not the property of the individual or the group but is communal. By being shared and re-used its value may be enhanced well beyond the original motivation of its creation. Bernal recognized this and his foresight led to the capture of crystallographic structures in the community's databases. It is a measure of his success that the databases are widely used as primary sources of information for research and, for example, form a key part of the arsenal for combatting molecular-based disease.

The crystallographic community can take credit for promoting the capture of high-quality curated scientific data. As part of this effort it has pioneered data deposition as an integral part of the scientific publication process. The continued communal development of protocols such as the IUCr's CIF can give reality to Bernal's vision by the next century.

2. Information

There has been much research into how 'Information' should be constructed and made available, which has not proved easy. The heart of the problem is that information requires data to be organized, which can be done in many ways. While many solutions are formally 'correct' they have proved both difficult to understand and expensive to implement. This leads to a 'priesthood' of specialists who design solutions and present them to 'users'. Where there is a clear formal requirement (*e.g.* airline reservations, stock-keeping, *etc.*) and sufficient investment this often succeeds. However, the model does not transfer easily to scientific research where the data are complex, often controversial, and understanding changes rapidly.

A commonly presented view of levels of understanding is

$$\text{Data} \rightarrow \text{Information} \rightarrow \text{Knowledge} \rightarrow \text{Wisdom}$$

with each level bringing complexity, difficulty and expense.

Abstract terms such as these are easy to misuse, so some concrete examples may be useful:

Data: A diffractogram or the 'line printer' output from a crystallographic experiment or program.

Information: A databank such as PDB containing data that has been evaluated, and canonicalized.

Knowledge: The ability to search databanks for a query phrased in human language (*e.g.* 'How does an amide interact with other molecules?')

Wisdom: Is your idea a worthwhile approach to a given problem? How should our community develop a strategy?

There is now considerable investment in the machine-based technology of 'knowledge management'; it is too early to say how valuable the various approaches are. In many cases they are likely to be highly domain-specific and to depend on the way that particular groups of people think.

A key aspect of knowledge is that it will require information from many disciplines. It is unlikely that the originators will have planned for their information to be used in this way, so knowledge tools must be generic. They must cater for many representations, varying in syntax, semantics and ontology. When crystallographic information is used by scientists from other disciplines, the way that we offer our knowledge will be crucial to how easily it can be used and therefore how valuable it will be.

3. Syntax, semantics and ontology

To be machine-readable a document must have an identifiable *syntax*. The PDB records:

```
ATOM     2 CA GLY A 1 -9.899 17.001 13.238 1.00 19.16
HETATM 999 CA         -9.899 17.001 13.238 1.00 19.16
```

are meaningless in isolation. The PDB documents the syntax and provides an online version but for many other formats syntax is often given in a proprietary, out-of-date, paper manual. People often 'guess' formats and there are examples of programs that misread the above records and confuse carbon with calcium. Unfortunately, a huge amount of effort is currently spent globally simply to tackle syntactic confusion.

The IUCr has tackled the limitations of Fortran-based syntax by creating the CIF format (Hall *et al.*, 1991).

This is a simple language for descriptive markup (single and looped name-value pairs) and 'small-molecule' crystallography has used CIF for several years for the archival of its data. CIF had to be extended to cope with the complexities of proteins. The extension for macromolecules, mmCIF, uses not only explicit markup but adds relational structure to the semantics. It is specified as the format to which the PDB will change in 1998. Bidirectional, relatively lossless, conversions between PDB and CIF are being developed, but the community as a whole has had no experience of mmCIF and there are no commonly used tools outside the community of developers. A CIF is *extensible* because it is not limited in the size of loops or the number and variety of named items.

Syntactic analysis of

```
_cell.length_a 12.34
```

provides the name `_cell.length_a` and the value '12.34'. Semantics are added by linking this information to a *dictionary*. This describes the type of the data (in this case float) and the human meaning of the term. A clean separation between syntax and semantics is essential for interoperability of documents. Moreover, if the semantics are to be accessible to machines, a clear mechanism for this must be provided (such as knowing the location and format of the latest mmCIF). Unfortunately, many current formats muddle syntax and semantics and the latter are often assumed rather than explicitly stated (*e.g.* the energy is 'obviously' in Hartrees because that is 'what every theoretical chemist uses').

4. The influence of the WWW

The remarkable growth of the WWW has shown that many people want to pursue Bernal's vision and they expect *knowledge* to be universally available. In many fields the WWW is now seen as the primary source of *information* and crystallographers (along with biologists) have led the way in the scientific disciplines.

There are few examples of effective knowledge management on the WWW but some useful prototypes include:

(*a*) Virtual libraries, home pages, *etc*. These are classifications (usually hierarchical) of a subject, often created by an enthusiast or small group. When well performed, they can be extremely effective ways of finding subject-based information. They work best when the vision of the creator is shared by the community.

(*b*) Search engines. In some cases these are extremely effective ways of searching global information, but are frequently useless due to 'noise'. Metadata should vastly increase the precision of searches and there are initiatives to collect domain-specific data more coherently.

(*c*) Newsgroups, virtual communities. Humans are currently the most effective way of managing knowledge and virtual communication with experts has been enormously effective. For example, the COMCIFs committee conducts most of the business of developing the CIF dictionaries through e-mail and other WWW techniques.

(*d*) Hypermedia (links). The WWW has so far been based on an extremely simple form of hypermedia – inline, single-ended links with no guarantee of robustness. Despite its theoretical flaws, this simplicity has been enormously successful in linking information sources. In particular, following links from a WWW resource is often likely to lead to useful information in a neighbouring domain.

Bernal's plea for the creation of a high-quality, communal, information pool is the simplest, most effective approach to a global resource of crystal-

lographic knowledge. It requires commitment, but if all crystallographers adopt this approach, the individual commitment will be small and relatively painless. The rewards will be immeasurable. The world is clamouring for electronic knowledge, epitomized by the establishment of the W3C. This is a body of several hundred major organizations (many commercial) who want to see the further development of WWW technology. A key requirement is for openness and interoperability; *i.e.* tools and protocols that can be used on any machine anywhere. A browser should be able to 'understand' any set of documents in a consistent way, no matter who created them or it.

The W3C's approach [*structured documents* (*SD*) *and hypermedia*] is generic and can be used in all disciplines. For example, there is strong similarity at the abstract level between the structures of documents describing a Shakespeare play, taxonomy, an mmCIF and an engineering materials catalogue. Mapping *B* factors onto atoms can use the same hypermedia technology as biblical criticism.

5. W3C activities

While writing this I am conscious how quickly technologies become commonplace, but 1998 seems set to be a year where robust, tested IT technologies make a dramatic effect on the global information community. This is a brief summary of the main ones, developed during the last 12 months. They have been enthusiastically received, and will be supported by all major IT-based companies. The driving force is commercial opportunities on the WWW and intranets, but as the technologies are generic they are ideally suited for managing scientific applications. The list below shows the most important protocols that have been, or will be, released this year.

(*a*) Extensible Markup Language (XML). XML is likely to provide the universal syntax for the WWW. It is an extremely simple implementation of SGML, the most commonly used approach for creating markup languages to support structured documents. (HTML is a very simple example of such a language.) Its main points are that user communities can develop their own tagset (*e.g.* <MOLECULE>) and can reliably communicate the structure of their information. ('This molecule is comprised of three other molecules, each accompanied by a textual explanation'.) Since the tags (*e.g.* <MOLECULE>) are meaningless without semantics, authors can customize documents through XLL and XSL (see below). XML 1.0 has recently been accepted by W3C, and both profit and non-profit organizations are already releasing parsers or text editors. We can expect a large number of high-quality generic tools to be announced very shortly. Non-textual applications are also being developed (such as MathML from W3C and the American Mathematical Society). In a complementary manner I am currently developing CML for molecules. Though CIF uses a different syntax, it is sufficiently structured that many operations may benefit from XML-based tools.

(*b*) eXtensible Linking Language (XLL). HTML introduced simple hypermedia but it has limitations of scale (*i.e.* large systems are difficult to maintain) and is not robust (if link targets are moved, the link is not changed). XLL (in final revision) provides arbitrarily complex linking including in- and out-of-line links, typed links and customization of link behaviour. It is also likely to provide protocols for the construction of link databases. This means that highly complex relationships (*e.g.* mappings, annotations, groupings and classifications) can be precisely represented.

(*c*) eXtensible Stylesheet Language (XSL). HTML irretrievably mixes content with form/style; it is normally impossible to locate a given part of an HTML document. XSL provides a clean separation between content (in XML) and style so that different styles (which might include molecular display) are available to authors and readers. Context-dependent rules in stylesheets allow cascading customization in several places: authoring, publisher/web server, client or user.

(*d*) Namespaces and semantics. The W3C expects structured documents to be composed from *information objects,* identifiable under a *namespace* protocol (currently under intense development). Mathematical equations can be created with MathML, text with HTxML, relational data with XML-Data, molecules with CML, and so on. A natural extension would be that learned societies such as IUCr could create their own namespace, identifiable by the uniqueness of the domain name system. A namespace allows identification of the owner of the semantics and ontology and provides mechanisms (probably through URIs) to locate it.

(*e*) Metadata. Metadata describe the role of a hyperdocument, such as the author, the owner, the components, their form, location and content. Authors can describe how their documents are to be used, and search engines can locate them with much greater precision. The Resource Description Framework (RDF) provides a powerful tool (semantic networks) in XML-based syntax and the Dublin Core gives a simple but powerful set of universal guidelines for creating metadata content. A major contribution to the dissemination of crystallographic knowledge will be the creation of simple, systematic terms for metadata.

(*f*) Document object model (DOM). Structured documents (*e.g.* CIFs) are most conveniently represented as trees and the DOM is an abstract description of how to create them. Documents in SD form are then available for searching, navigating, editing, filtering, transformation as well as use in hypermedia and styling. Creating the toolset for all these operations is a large

effort and DOM-based tools should provide valuable support for CIF-based documents.

The attraction of this approach is that it separates the components of information cleanly, and the abstraction allows the *creation of generic tools*. The macromolecular community will need to develop tools for authoring, editing, merging, validating, filtering, transforming and rendering interactive hyperdocuments. The W3C approach provides a head start since *many tools of high quality will be freely available.* Until now 'documents' and 'data' have often been managed by separate technologies; SDs provide a simple way to combine them without loss.

Since the technologies will be very widely available and the XML syntax is 'HTML-like', this approach will be readily accessible to the 'average' crystallographic programmer. A particular attraction of SDs is that 'information objects' from many disciplines can be combined in a single document. Thus maths, chemistry, crystallography, graphics and biology can all be taken from specialist implementations without each discipline trying to represent all components itself.

6. Electronic publication

Scientists publish for many reasons: informing others, establishing ownership and priority, archiving data, peer-review for career progression being the commonest. Electronic technology changes the means of publication and some of the motivation. In particular the static, immutable, non-interactive paper publication is often unable to communicate the real message of the author. However, the motivation and the social aspects of publication (along with a large and potentially vulnerable commercial sector) mean that strategies for e-publication are poorly defined and fragmented. Commercial fears mean that many publications will not interoperate (*i.e.* it may be difficult to link from one publisher to another) and the motivation for change is often for commercial reasons rather than the service of the author–reader community. With library budgets increasingly under pressure through rising prices we can expect to see radical changes in approach.

Of the many new opportunities in e-publication, crystallographers have led the way with the integration of documents and data. A crystallographic 'publication' can also be a data resource for analysis or input into programs or instruments. But this is just the start, since authors can now publish widely without the need for intermediaries. There are many advantages to doing this; new formats can be explored and new technologies developed. It will be the social aspects (*e.g.* technophobia, commercial interests, and fear of losing established means of quality control) that will dictate how quickly innovations are adopted.

In the information age, key aspects will be quality, authenticity, curation and guidance. They are strongly coupled to education and training. There is an oppor-

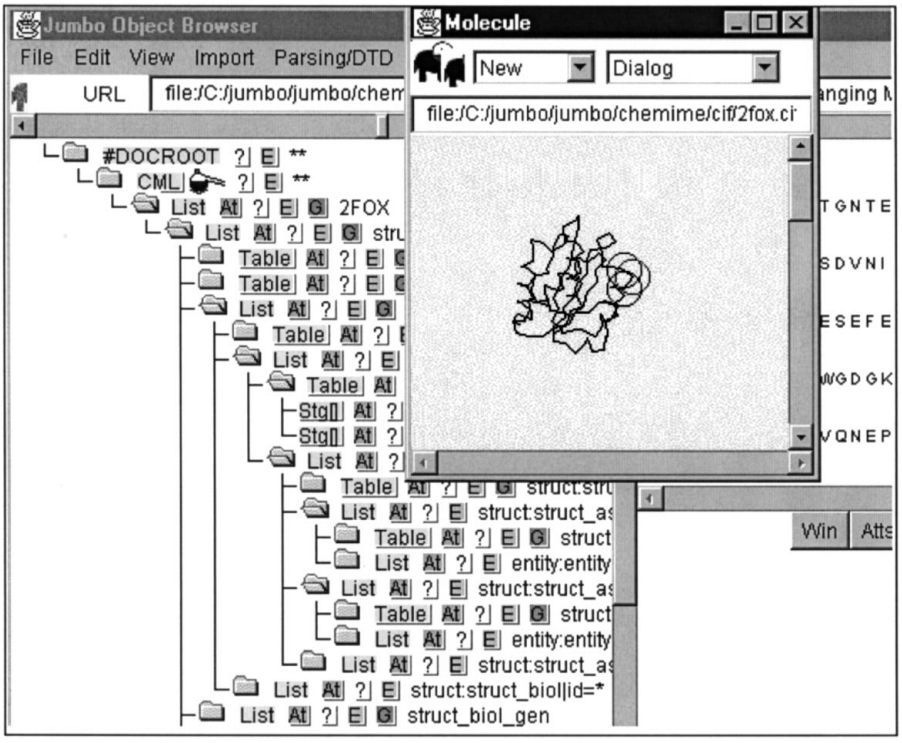

Fig. 1. *JUMBO-CIF* displaying the mmCIF for 2FOX. The tree represents the category structure in mmCIF with automatic joins from CIF tables. Special element-based Java routines display the structure and sequence and relational tables.

tunity for non-profit organizations such as learned societies, research organizations, universities and data centres to expand their role as key points in the information society.

7. Top-down or bottom-up?

Many IT projects involve top-down methodology (*i.e.* strict planning of the final system with increasing detail at each stage). In contrast the success of the WWW has been largely due to robust infrastructure (such as transport protocols and addressing) with an anarchic top level. Here ideas are developed independently as prototypes and the 'fittest' survive. Ideally this can be moderated by the virtual community, with very rapid feedback. For example, recently the XML community built an XML application programming interface (or API) (SAX – Simple API for XML) suitable for use in commercial applications, by a completely virtual process lasting a month. In areas such as bioinformatics where both the science and the IT change on a yearly timescale, long-term planning is often too slow. The W3C approach caters for many independent initiatives by supplying generic solutions.

However, this anarchy brings a need for central trusted bodies. If different groups develop their own ontologies (terminologies), it must be possible to locate these and also preserve them indefinitely. Archival and curation become increasingly important where information is distributed and documents require hyperlinks to be fully understood. The CIF dictionaries are an outstanding example of an open, publicly accessible electronic resource and the scientific community will look to other learned societies for similar tools. Because of their role in publication, such bodies will have a major part in offering a set of tools and ontologies for scientists.

8. A snapshot of the technology

Writing about technology that will be out-of-date when published is challenging! However, the general principles are fundamental and new to crystallography so I give an overview. A document is *parsed* into an abstract structure, most usefully a tree (Fig. 1).

A STAR file can support this model but the structuring in CIF is more limited. (Textual data is better suited by an *event-stream* model.) The tree will usually require rendering either to paper or to a browser, and the presentation will be controlled by *stylesheets.* This will be supported by major browsers to be released within the next few months. As part of the communal XML effort I have written a generic Java-based SD browser (*JUMBO*; Murray-Rust, 1997; Fig. 2), which can be extended to support application domains, especially (but not exclusively) in science. *JUMBO* has over 12 molecular extensions for legacy molecular files (including PDB). A recent prototype (*JUMBO-CIF*) supports both the tree structure of CIFs and the relational tables of mmCIF and related dictionaries. It can navigate dictionaries and link their semantics to data instances. Much current W3C-aligned work is on the interrelationship of trees (DOM), networks (RDF) and relational data (XML-Data). In *JUMBO-CIF* the mmCIF relations can be automatically expanded into trees and links through declarative programming in XML. These XML documents can act as stylesheets for clients who can customize their viewers to show different relations between components which use mmCIF pointers to control display of substructures.

In conjunction with Lesley West, I have also developed the Virtual HyperGlossary (VHG), which uses structured documents and namespaces to create hyperterminologies (Murray-Rust & West, 1997). These are XML-based and use ISO12620 data categories to provide a hierarchical approach to terminology and concepts. We are collaborating with the W3C leveraging action (W3C-LA) in providing a protein structure hyperglossary for the project.

9. The future

The rapid growth of the WWW makes detailed prediction impossible. It is probable that the borderlines between education, research, publication and commerce will shift dramatically. The goal of limitless knowledge beckons, but will be constrained by lack of trained people and the difficulty of navigating the new resources. Groups such as CCP4 and learned societies

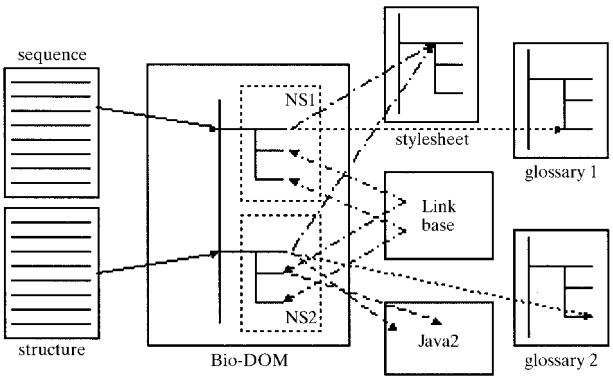

Fig. 2. Integrating documents from different domains. Sequence and structure information is converted to subtrees in an SD. Semantics are added by rendering through stylesheets or linking to domain-specific glossaries. In addition, the components of the two documents can be mapped onto each other through an out-of-line linkbase which might (say) correlate residues in the sequence with active-site information.

will be seen as stable centres for providing direction, support and guidance.

APPENDIX *A*
Glossary

Since I have stressed the role of terminology, this is a small glossary of terms used in this review. The terminology is taken from the approach of the VHG, and shows how semantic information can be structured. In an electronic publication these terms could be extracted into a distributed glossary and re-used in other applications.

Term: CCP4
Full name: Collaborative Computational Project 4 (macromolecular crystallography)
URL: http://www.dl.ac.uk/CCP/CCP4/main.html

Term: CIF
Full name: Crystallographic Information File
URL: http://www.iucr.org

Term: COMCIFs
Full name: Committee on CIFs
URL: http://www.iucr.org

Term: DOM
Full name: Document Object Model
Description: The W3C abstract representation of a document

Term: IUCr
Full name: International Union of Crystallography
URL: http://www.iucr.org

Term: PDB
Full name: Protein DataBank
URL: http://www.pdb.bnl.org

Term: RDF
Full name: Resource Description Framework
Description: The W3C XML-based representation for metadata

Term: SD
Full name: Structured document
Description: A tree-based representation of a document amenable to generic operations such as searching and transformation

Term: URI
Full name: Universal Resource Indicator
Description: A location-independent identifier that should add persistence to the URL concept

Term: VHG
Full name: Virtual HyperGlossary
Description: An XML-based approach to a universal syntax and structure for terminology
URL: http://www.vhg.org.uk

Term: W3C
Full name: World Wide Web Consortium
URL: http://www.w3.org

Term: WWW
Full name: World Wide Web

Term: XML
Full name: eXtensible Markup Language
Description: The W3C's syntax for structured documents

Term: XSL
Full name: eXtensible Stylesheet Language
Description: The W3C's specification for stylesheets, based on XML

References

Bernal, J. D. (1965). *Science in History*, p. 943, 3rd ed., New York: Hawthorn Books Inc.

Hall, S. R., Allen, F. H. & Brown, I. D. (1991). *Acta Cryst.* A**47**, 655–685.

Murray-Rust, P. (1997). *A Gentle Introduction to Structured Documents* and *JUMBO; An Object-Based XML Browser*, in *XML, Tools, Principles and Practice*, edited by D. Connolly. Sebastopol, CA: O'Reilly.

Murray-Rust, P. & West, L. J. (1997). *Steps Towards the Global Linking of Knowledge, Managing Information*, Vol. 5, pp. 34–36. London: Aslib.