

The CCPN project: an interim report on a data model for the NMR community

Rasmus Fogh, John Ionides, Eldon Ulrich, Wayne Boucher, Wim Vranken, Jens P. Linge, Michael Habeck, Wolfgang Rieping, T.N. Bhat, John Westbrook, Kim Henrick, Gary Gilliland, Helen Berman, Janet Thornton, Michael Nilges, John Markley and Ernest Laue

A recent workshop discusses the progress toward integrating NMR data into a unifying data model.

The Collaborative Computing Project for NMR (CCPN) was funded approximately two years ago by the Biotechnology and Biological Sciences Research Council (BBSRC) in the UK. The project has three main aims: (i) to facilitate data interchange between NMR-related software and to develop protocols to promote the archiving and exchange of NMR data, (ii) to develop software for the processing and analysis of NMR data, and (iii) to organize meetings to determine the best approach(es) to particular NMR problems and to help train Ph.D. students in NMR laboratories. These aims, and the way in which the project is run, are very similar to those of the CCP4 project, which develops software and organizes meetings for the X-ray community. Since the project was set up, it has gained significant support from the European Union (EU) and from industry, in particular from Astra-Zeneca, Bristol Myers Squibb, Genentech and GlaxoSmithKline.

From its inception the project has been organized for the whole of the biological NMR community, and the early discussions focused on the need for a standard data format for NMR that could be used by the entire NMR community. It has long been recognized that the field of macromolecular NMR has been restricted by the inability to move data between the different processing and analysis packages in a straightforward manner. For example, often one cannot readily use the best bits of two programs alternately, or easily develop and integrate a special-purpose routine. In addition, if data could later be reprocessed in different ways, by the same or different groups around the world, for example for method development, validation, or for quality improvement purposes, the standard data format would be of great value. The first steps toward this goal were taken some years ago by an IUPAC-IUBMB-IUPAB Task Group (chaired by Kurt Wüthrich at the ETH, Zurich,

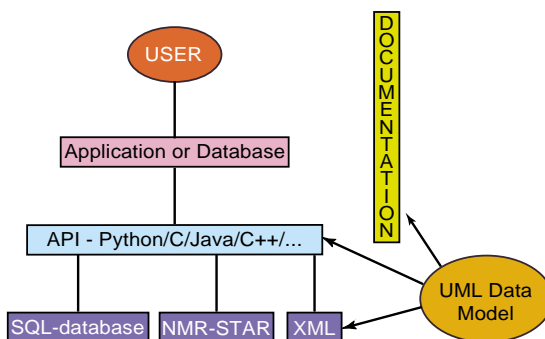


Fig. 1 Implementation of the CCPN project. Users interact with NMR processing and analysis packages as before, but NMR software developers use the APIs to interact with the underlying NMR data. It is envisaged that the way programs interact with the APIs will remain stable over time so that the underlying data formats can change without affecting analysis programs or the databases.

Switzerland), who considered the standardization of data to be reported from protein and nucleic acid structures determined by NMR. This group developed recommendations for nomenclature and data reporting, which were subsequently implemented by the BioMagResBank (BMRB) in the NMR-STAR format (see below).

The development of a complete, interoperable, data exchange and data merging mechanism that facilitates reliable criss-cross would also facilitate so called 'data harvesting'. This refers to carrying forward all the known information about a particular project, for example, in some form of database, from one program to another through all the different stages of analysis up to its ultimate deposition in the public databases. Deposition of both the experimental and structural data to archive databases is an ever more important requirement, but at present one must often input information into the deposition anew. In practice this means that much information is never deposited, because converting the data into the required format(s) simply requires too much effort. The implementation of data harvesting would thus allow the straightforward deposition of a much richer amount of data than at present. Moreover, an automated approach to the collation and deposition of NMR data is clearly

needed with the advent of high-throughput approaches to NMR structure determination. All these problems could be alleviated if there was an agreed way of storing and exchanging macromolecular NMR information between different computer programs.

Macromolecular NMR information is currently stored in a large number of different data formats. For example, we have several formats for raw data acquired on a NMR spectrometer. After data processing, we have many formats for both NMR spectra and the derived data obtained after analysis: lists of signals after peak picking, chemical shift assignments, structural restraints (usually dipolar or J-couplings and Nuclear Overhauser Effects (NOEs)), relaxation data, *etc.* In practice, the development of a common data format for NMR has thus necessarily involved the NMR instrument manufacturers, the software developers in the NMR community, and the public databases, including the primary archives for NMR data at the BMRB and NMR-derived structures at the Protein Data Bank (PDB). Data collection and annotation sites include the University of Wisconsin-Madison (Madison, USA; BMRB), Rutgers University (New Jersey, USA; PDB), the European Bioinformatics Institute (EBI; Hinxton, UK; PDB) and the Institute for Protein Research (Osaka, Japan; PDB). In addition, the project has

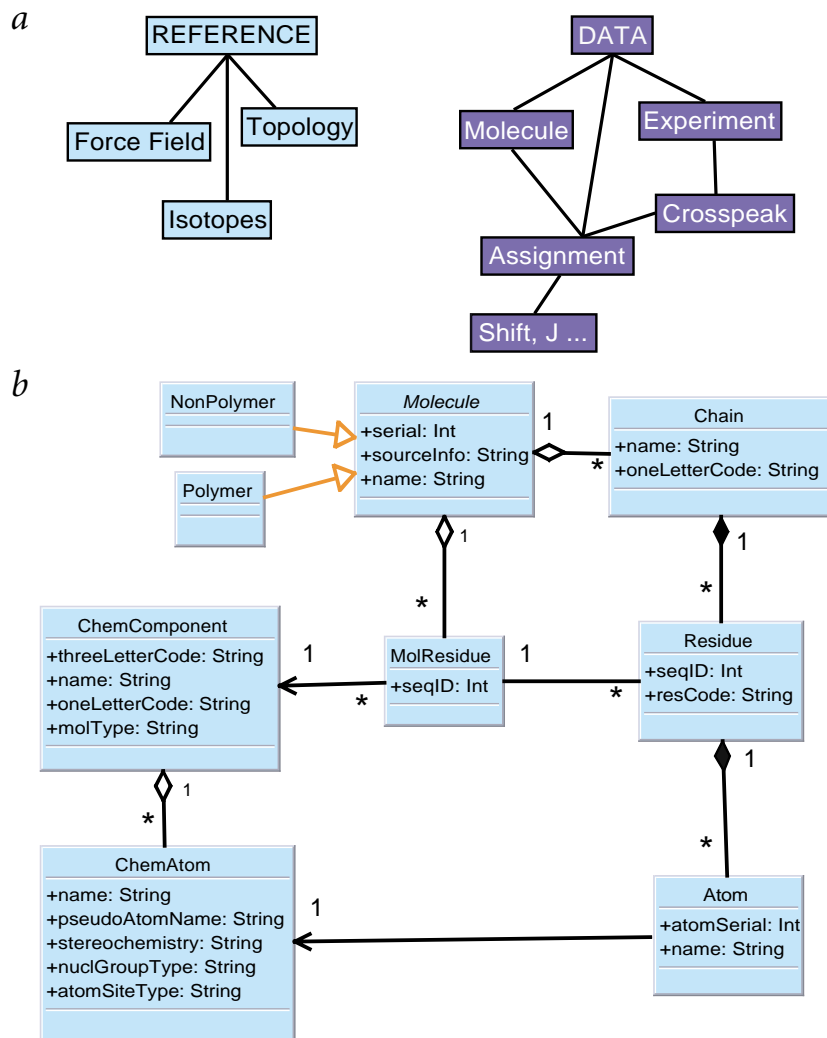


Fig. 2 The CCPN data model. **a**, An overview of the areas covered by the present data model. **b**, A view of the 'molecule' part of the data model (for more details and an explanation of the links, see <http://www.bio.cam.ac.uk/nmr/ccp/datamodel/>).

these data. Such a description remains valid regardless of the file format and organization chosen to represent it. This abstract description then provides the basis for mapping onto existing formats like the PDB coordinate format and the NMR-STAR format developed by the BMRB, as well as for the implementation of new formats such as XML or SQL databases. Structural data has been stored in the PDB coordinate format for many years (see <http://www.rcsb.org/pdb/>) and is used by a very large number of software packages. NMR-STAR provides an existing vehicle for depositing data from a variety of software packages used by the NMR community, is extensible and is in use for depositing chemical shift assignments, couplings, a variety of relaxation parameters, pKa values, atomic coordinates, NMR constraints, and many other types of data at the BMRB (see <http://www.bmrwisc.edu/>).

APIs that allow data to be read into memory, manipulated and written out in several different formats are being produced for a number of computer languages. Our present aim is to provide a close mapping between the CCPN data model and the model that is inherent in NMR-STAR as they are developed in parallel, but in the long term we hope to make these data models congruent. It is envisaged that the manner in which users and databases interact with the APIs will remain stable so that the underlying formats can change, new formats can be developed and even the data model can change, without affecting the NMR processing and analysis packages built on top. By supporting new formats for NMR data (such as XML and SQL databases), it will be much easier in the future to make elaborate queries of the data (for example, to obtain the chemical shift values of the β -methylene protons of those phenylalanines that are next to a glycine in the protein sequence). We have written the data model in UML, an industry standard modeling language that allows a complete and precise description of the data organization. We have also developed computer programs, which generate the required computer code (APIs, input/output routines, data format descriptions such as XML document type definitions (DTDs) and database schemas, documentation, etc.) from the UML description of

also benefited from discussions with members of biological NMR groups, both academic and industrial, in Europe, Japan and the USA.

The issues involved were discussed at two previous CCPN workshops (at EBI, Hinxton, UK and at the National Institute of Standards and Technology (NIST), Rockville, Maryland, USA) where some strategic decisions were made (see <http://www.bio.cam.ac.uk/nmr/ccp/meetings/>). The initial proposal was to develop an exchange format that all NMR analysis packages could read from and write to. However, history suggests that formats and data standards come and go much faster than the data itself, in particular as new methods of analysis are invented. By contrast, data processing techniques and tools are relatively more stable. It was felt that a way to shield the community from such recurring need for modifications in data exchange routines would be to propose and develop an application programming interface (API). In such an approach, the

API would be expected to constantly evolve to accommodate changes in areas related to data exchange, data conversion, data standards and data formats rather than the application program or database itself (Fig. 1). In other words, this approach would permit constant evolution and new inventions in exchange formats without substantially affecting NMR analysis packages and the user community.

To define the structure of the API, and to support both existing and future data formats, the CCPN project elected to concentrate on charting the organization of the data without reference to any particular format by making a so-called 'data model'. For example, a protein is generally considered as a linear chain of residues, which in turn are made up of atoms. A data model describes the relevant objects (in this case molecule, residue and atom), their attributes (atom name, residue type and three-letter code) and the protein's name, synonyms and function, as well as the nature of the relationships between



the data model (Fig. 1). As a result there is only one source document that needs to be maintained and all the remaining code in the project can be automatically synchronized with it. This approach should drastically reduce the amount of manual coding required for maintenance in the future.

The third CCPN workshop on the development of the data model was held at the EBI in November 2001 with the aim of discussing advances in computational aspects of NMR and the draft data model (see <http://www.bio.cam.ac.uk/nmr/ccp/meetings>). The 36 participants represented a very strong congregation of both academic and commercial NMR software and database developers. The draft data model in UML as well as an example API, in the object-oriented scripting language PYTHON, and a storage format with input/output code in XML, together with accompanying documentation (all generated automatically from the UML data model), had been circulated prior to the meeting (see <http://www.bio.cam.ac.uk/nmr/ccp/datamodel>).

The draft data model covers (in varying amounts of detail) the core areas of macromolecular NMR — experiments and spectra, crosspeaks, assignments, chemical shift lists, restraints, molecular topology and molecular structure (see Fig. 2). The data system for crosspeaks and spectra has been designed to accommodate a wide range of experimental aspects, including ambiguous assignments, reduced-dimensionality experiments, *etc.* The system for storing assignments is based on assigning crosspeaks, chemical shifts, *etc.*, to an intermediate object, rather than to a specific atom in the assignment process, where it is not yet known which atoms correspond

to which NMR resonances. For use in the assignment process, the model contains a number of structures to join resonances into groups and to add partial assignment information of various kinds. The data model incorporates the NMR experimental data content already present in the BRMB's NMR-STAR v2.1 and the molecular description is essentially the PDB-to-mmCIF exchange data dictionary developed at the RCSB. The CCPN project is contributing to the BRMB's development of a new version of NMR-STAR (v3.0), which incorporates many of the new areas developed for the CCPN data model. The data model is still developing rapidly and the most up-to-date information can be found at <http://www.bio.cam.ac.uk/nmr/ccp/datamodel>. In addition, the current version of the new NMR-STAR (v3.0) can be found at http://www.bmrb.wisc.edu/dictionary/htmldocs/nmr_star/dictionary.html

At the meeting there was detailed discussion of the draft data model and the data content that should be included. In due course it is planned that the data model will comprise APIs for PYTHON, C, and C++/Java, as well as XML and SQL database storage formats, in addition to the close coupling with the PDB and the BMRB's NMR-STAR.

From the organizers' point of view this was an extremely useful meeting. We are now planning to produce a more complete version of the data model and to present this for comment to the biological NMR community. It is then our intention to write up the work for publication later this year — this will involve all those who have contributed to the project. For it to be a success we now very much need feedback from the wider NMR community,

with comments and suggestions for improvement. It is important to emphasize that this project belongs to the NMR community as a whole and not to a particular laboratory or laboratories. It will only be a success if everyone now provides the invaluable input that will help us make this into a functional data model that is taken up and used by the NMR community.

Rasmus Fogh, Wayne Boucher and Ernest Laue are in the Department of Biochemistry, University of Cambridge, 80 Tennis Court Road, Cambridge, CB2 1GA, UK; John Ionides, Wim Vranken, Kim Henrick and Janet Thornton are at EMBL-EBI, The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK; Eldon Ulrich and John Markley are in the Department of Biochemistry, University of Wisconsin, 433 Babcock Drive, Madison, Wisconsin 53706-1544, USA; Jens P. Linge, Michael Habeck, Wolfgang Rieping and Michael Nilges are in the Unite de Bioinformatique Structurale, Institut Pasteur, 25-28 rue du Docteur Roux, F-75015 Paris, France; T.N. Bhat is in the Biotechnology Division (831), NIST, 100 Bureau Drive, Stop 8314, Gaithersburg, Maryland 20899-8314, USA; Helen Berman and John Westbrook are at Rutgers University, Rutgers, The State University of New Jersey, Department of Chemistry and Chemical Biology, 610 Taylor Road, Piscataway, New Jersey 08854-8087, USA; Gary Gilliland is at the Center for Advanced Research in Biotechnology, NIST, 9600 Gudelsky Drive, Rockville, Maryland 20850, USA. Correspondence should be addressed to E.L. email: e.d.laue@bioc.cam.ac.uk.