Institution: University of Cambridge

Unit of Assessment: UoA5

Title of case study: The Collaborative Computing Project for NMR (CCPN): A novel approach to data exchange between software applications

1. Summary of the impact (indicative maximum 100 words)

Researchers in Cambridge have developed a data standard for storing and exchanging data between different programs in the field of macromolecular NMR spectroscopy. The standard has been used as the foundation for the development of an open source software suite for NMR data analysis, leading to improved research tools which have been widely adopted by both industrial and academic research groups, who benefit from faster drug development times and lower development costs. The CCPN data standard is an integral part of major European collaborative efforts for NMR software integration, and is being used by the major public databases for protein structures and NMR data, namely Protein Data Bank in Europe (PDBe) and BioMagResBank.

2. Underpinning research (indicative maximum 500 words)

Over the past decade, the accumulation, collation and processing of large amounts of data has become increasingly important in biology. This is particularly true for structural biology, which is built around automated data collection and complex consecutive data processing steps. However, the increasing emphasis on software pipelines, on improved productivity and usability, and on deposition of data and results, has until relatively recently proved problematic for macromolecular NMR spectroscopy, one of the prominent techniques in the field. Relevant programs were frequently disjointed and badly maintained, with little provision for combining programs, and exchanging or tracking data.

Recognising these issues, the Collaborative Computational Project for NMR (CPPN; Department of Biochemistry), was founded by Professor Ernest Laue (employed since 1982, Professor since 2000) and computer scientists working in his group; Rasmus Fogh (SRA, 1999-present), Wayne Boucher (Computer Officer, 2000-present), and Tim Stevens (Research Associate, 2002-present), with contributions from John Ionides and Wim Vranken (both then at the European Bioinformatics Institute, Hinxton, Cambridge. From 1999, researchers in the CCPN group worked on a data exchange standard and the integration of software for macromolecular NMR. They developed a precise, detailed data model to serve as a one-stop storage for all data from acquisition through to final deposition, accompanied by extensive data access subroutines in several programming languages (Python, Java, and C[1,2]).

To make it practical to create the several million lines of code in the subroutine libraries and keep them synchronized as the model evolved, CCPN also created a data modelling and code generation framework (MEMOPS[3]) that automatically generates data access libraries from an abstract data model description. Finally, CCPN used the model and subroutine libraries as the basis for the CcpNmr suite of NMR programs, providing quality software for the community, allowing for entry and visualisation of data, and serving as a hub for software integration[4]. The main programs in the suite are CcpNmr Analysis, for interactive display, analysis and assignment of NMR data, and CcpNmr FormatConverter, which converts data between more than thirty different formats using the data model as a stepping-stone. The first test version of the data model and CcpNmr suite was released in September 2003, and CCPN has developed new releases on an ongoing ad hoc basis (latest release was August 2013).

CCPN designed the software framework to support different programs from the outset, reflected in the easy addition of new capabilities to CcpNmr Analysis and new formats to CcpNmr FormatConverter by both CCPN and external groups. Examples include the addition of a metabolomics protocol to Analysis by an external group[5] and the development of solid state NMR modules for CcpNmr Analysis, with subsequent uptake of Analysis by solid state NMR groups[6].

The unique advantage of the CCPN data model lies in providing a comprehensive and unambiguous way to store data and transfer this between programs. Integrating a program with the data model immediately allows it to farm out tasks to other programs and combine with them into pipelines, leading to higher quality, better integrated and more accessible software, and access to the very significant computational resources available through the Grid-based infrastructure results in much faster results generation and large productivity gains.

3. References to the research (indicative maximum of six references)

Publications:

1. The CCPN project: an interim report on a data model for the NMR community. Fogh, R, Ionides, J, Ulrich, E, Boucher, W, Vranken, W, Linge, JP, Habeck, M, Rieping, W, Bhat, TN, Westbrook, J, Henrick, K, Gilliland, G, Berman, H, Thornton, J, Nilges, M, Markley, J, Laue, E. *Nat Struct Biol* **9**, 416-8 (2002)
2. The CCPN data model for NMR spectroscopy: development of a software pipeline. Vranken, WF, Boucher, W, Stevens, TJ, Fogh, RH, Pajon, A, Llinas, M, Ulrich, EL, Markley, JL, Ionides, J, Laue, ED. *Proteins* **59**, 687-96 (2005)
3. MEMOPS: data modelling and automatic code generation. Fogh, RH, Boucher, W, Ionides, JM, Vranken, WF, Stevens, TJ, Laue, ED. *J Integr Bioinform* **7**(2010)
4. A framework for scientific data modeling and automated software development. Fogh, RH, Boucher, W, Vranken, WF, Pajon, A, Stevens, TJ, Bhat, TN, Westbrook, J, Ionides, JM, Laue, ED. *Bioinformatics* **21**, 1678-84 (2005)
5. *The CCPN Metabolomics Project: a fast protocol for metabolite identification by 2D-NMR*. Chignola, F, Mari, S, Stevens, TJ, Fogh, RH, Mannella, V, Boucher, W, Musco, G. *Bioinformatics* **27**(6) *885-886 (2011)*
6. A software framework for analysing solid-state MAS NMR data. Stevens, TJ, Fogh, RH, Boucher, W, Higman, VA, Eisenmenger, F, Bardiaux, B, van Rossum, BJ, Oschkinat, H, Laue, ED. *J Biomol NMR,* **51**, 437-47 (2011)

Key research grants:

- "CCPN – Collaborative Computational Project for macromolecular NMR spectroscopy", BBSRC, 2003 – 2006, £574,468 (Laue as PI)
- "CCPN – Collaborative Computational Project for macromolecular NMR spectroscopy", BBSRC, 2006 – 2009, £1,460,689 (Laue as PI)
- "CCPN – Collaborative Computational Project for macromolecular NMR spectroscopy", BBSRC, 2009 – 2012, £927,786 (Laue as PI)
- "Extend-NMR - Extending NMR for Functional and Structural Genomics", EC FP6, 2006 – 2009, £324,758 (Laue as PI)
- "WeNMR – A worldwide e-infrastructure for NMR and structural biology", EC FP7, 2010 – 2013, £134,584
- "Collaborative Computational Project for macromolecular NMR (CCPN). Supporting biomolecular NMR and community driven NMR software development.", MRC, 2013 – 2016, £762,546 (Laue as Co-I)

4. Details of the impact (indicative maximum 750 words)

Impacts on commerce

By August 2013, there were over 1000 registered installations of CcpNmr Analysis, with 356 registered users on the CcpNmr Users mail feed (the exact number of end users is difficult to quantify given that an important contribution of CCPN lies in the improvement and integration of programs from other sources). A 2012 user survey by CCPN gave a distribution of 23% UK, 34% North American, 34% European, and 10% 'other' installations, with an average number of users per installation of over 4 (corresponding to several thousand users). By mid-2013, the following companies were registered as paying subscribers: Astra Zeneca, Astex, Genentech[7], Novartis[8], Novo-Nordisk[9], Syngenta[10], and Vernalis[11], yielding an annual total fee income of £43,500. All fee income is directly reinvested in the project. Commercial users of CcpNmr Analysis benefit from faster drug development times and lower development costs. For example:

- 'we [Genentech] find it very beneficial to our company. We particularly appreciate the data

model that has been developed and incorporated into the 'Analysis' and 'Format Converter' software packages for the efficient visualization and assignment of NMR spectra. These features greatly facilitate the structural characterization of protein and protein-ligand interactions'[7]

- 'we [Novartis] use the CCPN software for the purpose of visualizing and analysing multidimensional NMR data of proteins, assignment of proteins and chemical shift mapping. These are critical features … the CcpNmr software is absolutely crucial for our work'[8]
- 'The open nature of the software facilitates the integration of the software into our workflow and thus allows us [Novo Nordisk] to spend more of our time tackling the problem at hand'[9]
- 'Syngenta uses the CCPN software for discovering and ensuring the safety of new plant protection products. I find it very beneficial to my work and particularly appreciate the contribution CCPN has made to simplify the transfer of NMR data between software packages'[10]
- 'CCPN provides an invaluable NMR data analysis platform which is crucial to our research [at Vernalis]; at present, no other software package is available which meets our requirements. We feel that this software greatly enhances our productivity'[11]

In addition, CCPN has a special collaboration with Novartis, which involves Novartis registering as a premium member of CCPN (at an additional charge of £14,000), and CCPN developing customized software now being used by the company in structure-based drug design by NMR.[8]

Impacts on practitioners: working protocols have been changed; quality has been improved
Much CCPN code is distributed under the LGPL license, allowing free copying and reuse. As a result, the number of practitioners (NMR users) impacted is substantial, and crosses a wide range of sectors; NMR is a commonly used tool for companies working in, for example, the pharmaceutical, chemical, agro-chemical and petroleum sectors.

The CCPN data model has been chosen to underpin various international NMR software integration efforts, facilitating change in working protocols for NMR software in commercial and academic laboratories worldwide. The Laue group (as CCPN) coordinated the Extend-NMR project (European 6th Framework Programme), a collaboration of European software developers to develop an integrated software pipeline to support functional and structural proteomics. The CCPN data model and CcpNmr FormatConverter were chosen by the eNMR project (7th Framework Programme), as the main tool to exchange data between programs in its development of a Grid-based platform for computationally intensive calculations used during biomolecular NMR data analysis. This led to a larger integration effort, the WeNMR project (7th Framework Programme, currently under way), which "*aims to bring together complementary research teams in the structural biology and life science area into a virtual research community at a worldwide level and provide them with a platform integrating and streamlining the computational approaches necessary for NMR and SAXS data analysis and structural modelling*". WeNMR links the software pipelines developed in Extend-NMR to the web servers for NMR calculations developed in eNMR and now running on the Grid, and opted to use the CCPN data model as its central data conversion utility. As a result of these integration projects, there has been a significant benefit to commercial users. For example: '*We [Novo Nordisk] particularly appreciate the easy integration with other NMR related software packages, which allows us to switch between different structure calculation programs with ease and use the most appropriate tool for each project*.'[9]

The universal appeal and easy expandability of the CCPN data model has led to it being adopted by various NMR databases and programmes used by industry and academia, for example, the NMR part of the EUROCarb database[12]. Programs in the NMR field that have been integrated with the CCPN data model from the Extend-NMR collaboration include: MDD, PRODECOMP, and Bruker's TOPSPIN (NMR processing); ARIA, HADDOCK, and ISD (structure generation); and CING (structure validation). Other programs that have been integrated include: PALES and MODULE (data analysis and assignment); MECCANO (structure generation); RPF (structure validation); and ECI (database deposition). Programs like DANGLE (structure estimation from chemical shifts) and MARS (automatic backbone assignment) have been further integrated to use CcpNmr Analysis as a front end to prepare data. A further recent example is the CASPER program

for determination of polysaccharide structures from NMR data[13].

The CCPN software has also been applied to projects beyond studies of macromolecular structure and dynamics, e.g. metabolomics[14] and fragment-based small molecule screening. For example, Vernalis has stated that the CCPN software '*facilitates our approach of fragment-based lead discovery*'[11]. The company is currently collaborating with CCPN on the development and testing of a module for NMR-based ligand screening, to improve productivity in identifying low-affinity ligands using NMR, and to which the company is inputting expertise and datasets[11].

A significant long-term impact of CCPN on the worldwide NMR user community has been on the quality and quantity of data available[15,16,17,18]. The data model has facilitated the deposition of much more complete NMR datasets in the public databases such as the PDBe (Protein Data Bank in Europe) and BioMagResBank (the publically-accessible international repository of NMR data, University of Wisconsin-Madison). The CCPN software is also making it possible to recalculate results (and thus improve their quality) as new and improved algorithms are developed. Resulting recalculated NMR structures are available in the RECOORD and logRECOORD databases.

5. Sources to corroborate the impact (indicative maximum of 10 references)

7. Letter of support from Scientific Manager, Department of Structural Biology, Genentech
8. Letter of support from Investigator in Biomolecular NMR, Novartis
9. Letter of support from Senior Scientist, Department of Diabetes Biophysics, Novo Nordisk
10. Letter of support from Senior Technical Expert in NMR, Syngenta Ltd
11. Letter of support from Research Fellow, Vernalis
12. EUROCarbDB: An open-access platform for glycoinformatics. Von der Lieth, CW, Freire, AA, Blank, D, Campbell, MP, Ceroni, A, Damerell, DR, Dell, A, Dwek, RA, Ernst, B, Fogh, RH, Frank, M, Geyer, H, Geyer, R, Harrison, MJ, Henrick, K, Herget, S, Hull, WE, Ionides, J, Joshi, HJ, Kamerling, JP, Leeflang, BR, Lütteke, T, Lundborg, M, Maass, K, Merry, A, Ranzinger, R, Rosen, J, Royle, L, Rudd, PM, Schloissnig, S, Stenutz, R, Vranken, WF, Widmalm, G, Haslam, SM, *Glycobiology*, **21**(4), 493–502 (2011)
13. Automatic Structure Determination of Regular Polysaccharides Based Solely on NMR Spectroscopy. Lundborg, M, Fontana, C, Widmalm, G. *Biomacromolecules (2011)* DOI: 10.1021/bm201169y
14. The CCPN Metabolomics Project: a fast protocol for metabolite *identification* by 2D-NMR. Chignola, F, Mari, S, Stevens, TJ, Fogh, RH, Mannella, V, Boucher, W, Musco, G. *Bioinformatics* **27**(6) 885-886 (2011)
15. The Protein Data Bank in Europe (PDBe): bringing structure to biology. Velankar, S, Kleywegt, GJ, *Acta Cryst*, D67:324-330 (2011)
16. RECOORD: a REcalculated COORdinates Database of 500+ proteins from the PDB using restraints from the BioMagResBank. A.J. Nederveen, J.F. Doreleijers, W.F. Vranken, Z. Miller, C.A.E.M. Spronk, S.B. Nabuurs, P. Güntert, M. Livny, J.L. Markley, M. Nilges, E.L. Ulrich, R. Kaptein and A.M.J.J. Bonvin. *Proteins,* **59**, 662-672 (2005)
17. The NMR restraints grid at BMRB for 5,266 protein and nucleic acid PDB entries. Doreleijers, JF, Vranken, WF, Schulte, C, Lin, J, Wedell, JR, Penkett, CJ, Vuister, GW, Vriend, G, Markley, JL, Ulrich, EL. *Journal of Biomolecular NMR*, **45**, 389–396 (2009)
18. Bayesian estimation of NMR restraint potential and weight: A validation on a representative set of protein structures. A. Bernard, W.F. Vranken, B. Bardiaux, M. Nilges and T.E. Malliavin. *Proteins,* **79**, 1525-1537 (2011)