J Comput Aided Mol Des (2009) 23:391–394
DOI 10.1007/s10822-009-9272-5

WARR'S PIECE

CCDC well groomed: an interview with Colin Groom, Executive Director, Cambridge Crystallographic Data Centre, and Frank Allen, Emeritus Fellow

Wendy A. Warr

Received: 9 April 2009 / Accepted: 15 April 2009 / Published online: 7 May 2009

Dr. Colin Groom (CG, pictured right) was appointed as executive director of the CCDC [1] from 1 October 2008, following the retirement of Dr. Frank Allen (FA, pictured left). He was formerly head of Computer-Assisted Drug Discovery and Investigative Chemistry at UCB in Cambridge, UK. He has a B.Sc. in biotechnology and a Ph.D. in protein crystallography from the University of Leeds, and held postdoctoral appointments at Leeds and at Massey University, New Zealand. Since then, he has worked in the pharmaceutical industry, joining Pfizer Global Research, Sandwich, in 1994, where he established a protein crystallography facility, and then established and led molecular informatics units in both Sandwich and in the USA. He joined UCB (Celltech) in November 2002. He was co-director of the International School of Crystallography: Molecules to Medicines via Crystallography held in Erice, Italy in May 2008. He has been at the forefront of applying structural information to drug design and has a number of publications and patents in this area. He is a Fellow of the Royal Society of Chemistry, and a member of the British Crystallographic Association, and has contributed to Biotechnology and the Biological Sciences Research Council (BBSRC) and the Engineering and Physical Sciences Research Council (EPSRC) activities as a panel member. He is on the Editorial Boards of *Current Computer-Aided Drug Design* and *Current Drug Discovery Technologies*.

Interview

WAW: When I first called you, you seemed genuinely excited and pleased about your new appointment. Where does CCDC fit in with your personal interests and objectives?

CG: During a Ph.D. one generates an empathy with a subject that stays with one forever. Mine was crystallography. So this position is really returning to my roots. My interest is in molecular interactions, be they drugs with proteins in cells, or small molecules with each other in crystals. That's what the CCDC is all about; it's a perfect fit.

WAW: Do you see yourself as a manager or a researcher?

CG: Well the easy response to this is to sit on the fence and say both, but that's not very revealing, so I'll have to

W. A. Warr (✉)
Wendy Warr & Associates, 6 Berwick Court, Holmes Chapel, Cheshire CW4 7HZ, UK
e-mail: wendy@warr.com

say it's a manager. I guess I went through a transition that many scientists go through in their careers, feeling "I never have enough time for my own research and that's what I should be doing". Then I came to appreciate that I quite enjoyed helping others to do their work and they were almost always more able than me! I wish I'd come to this realization earlier. I saw management as a time consuming, necessary evil; it's not, I help people.

WAW: I note the mission statement on your web site: "The CCDC is dedicated to the advancement of chemistry and crystallography for the public benefit through providing high-quality information services and software". Was this written before your time? Would you like to see it changed in any way?

CG: It was written before I joined the CCDC. Whilst it still captures what the CCDC is all about, it perhaps lacks a little punch. I'm keen on the CCDC adopting a "Big Hairy Audacious Goal" (BHAG). This phrase was coined by James Collins and Jerry Porras in a 1996 article entitled "Building Your Company's Vision". It is a form of vision statement, but one that has a finish line. It's a goal that people can shoot for. Ford wanted to "democratize the automobile", Microsoft wants "a computer on every desk and in every home", Symbian would like to "become the most widely used operating system in the world". Expect something challenging from the CCDC soon.

WAW: Some people (especially those in the Open Data movement) seeing the words "public benefit" might think that you ought to be freely distributing the Cambridge Structural Database (CSD) data. What would you say to them?

CG: Well we do already do this. Anyone can request a structure from the CCDC for bona fide research or teaching purposes. We handled around 15,000 such requests last year. The CSD is used in 68 countries. Where institutions genuinely have difficulty paying we will always come to some arrangement. What really has to be funded is the extensive curation that is required for many of the structures we process, and development of the software to derive knowledge from the data.

WAW: How many structures do you have?

CG: The CSD contains bibliographic, chemical and crystallographic information for organic and metal-organic compounds. These are derived from publications in the literature and direct data deposition with the CCDC. The current database has around 475,000 structures; it should reach half a million by the end of 2009.

WAW: Do you get complaints about the cost of the database (or your software)?

CG: Of course most people want to get as much as they can for as little money, no matter what the circumstances. However, there is an appreciation in the community that the CCDC does not exist to make a profit. In fact we're not allowed to. The revenue we generate is used to fund the services we provide. Compared to most commercial software what we provide is a bargain. We also take our commitment to support very seriously and have a team of scientists dedicated to working with CCDC software users.

WAW: If CCDC were a for-profit company you would be accountable to your shareholders; to whom are you accountable?

CG: CCDC has an excellent Board of Governors. They represent a broad cross section of international users and depositors, from industry and academia. None is paid; they do it because they care about the CCDC. But of course I'm really responsible to the users. If we don't provide what they want, CCDC will cease to be. That's quite a strong motivation to get it right for people.

WAW: How do you find out what the users want?

CG: We have just completed a web-based survey; 557 institutions responded from all over the world. The survey concerned the development of CSD: structure deposition, data presentation, queries, and analysis. The database is the center of our universe. This is why we exist.

FA: But we see the CSD as a *system*: the database comes bundled with CONQUEST, Vista and Mercury. Mercury [2] is free of charge: there have been 15,000 free downloads of it for use with Crystallographic Information Files (CIFs) [3] etc. in-house.

CG: We also go to dozens of conferences and make site visits. We have a free user meeting coming up in June. This one will be for small numbers of industrial users.

WAW: The Protein Data Bank (PDB) it is funded differently from CSD. Why is it "free" when CSD is not?

CG: Well of course it's not really free, but the cost is not as obvious to the user. The PDB is paid for by US tax payers and a number of charities, but with the CSD the users pay, either from company funds or by using their research funding. CCDC receives its structures electronically as CIF documents deposited via journals, or as donated CIFs. We have arrangements with over 120 journals to receive CIFs prior to article publication and we work with them to improve data deposition mechanisms. Historically, such data used to be published in full in print for small molecules, but when proteins began to be published this was neither cost-effective for the journals nor useful for the reader, so direct deposition of protein data into the PDB rapidly became the norm in the 1970s and 1980s. Nowadays, most journals won't publish a protein structure unless the author provides a PDB ID-number, indicating that the data are actually deposited and have passed some basic validation checks.

WAW: What are the advantages of being a self financing charity as opposed to be being dependent on government funding?

CG: Well, there are two advantages really. The first is the stability of the organization. The way we get funded at

the moment means that we will be able to provide the CSD and associated software for as long as the user base thinks it is of value. The second concerns accountability. At the moment we are directly accountable to the users. That's the way it should be.

WAW: Is the charitable status of CCDC likely to change while you are at the helm?

CG: The charitable status means that the CCDC has a tax position appropriate to the work that we do; we help scientists do valuable research, and we do research of our own. Whilst our status *could* restrict the type of work we do, we have no desire to do anything other than this. So I really don't want to see this change. The CCDC exists for the benefit of the community; it should stay that way.

WAW: What proportion of your income comes from database subscriptions and how much comes from software licenses?

CG: Well that's a tricky one. We try to adjust our activities so everything becomes self-funding. That way we know we are doing what people want. In the current times it's also hard to delineate a database from software. Although we do not profit from what we do, there are of course profit-making organizations selling software for commercial gain. In a sense we are competing against these, so we need to be a little careful not to reveal too much sensitive financial information.

WAW: So, who are CCDC's competitors?

CG: The CSD is unique but there *are* groups that assemble crystal structures in repositories on the web. These collections may not be sustainable and the structures are error prone. Such a collection is not a database but a set of structures. This "open data" situation is somewhat similar to open access to the literature but most of the crystal structure collections are poorly coordinated: they are not as good as institutional repositories. The CSD is a value-added database. The problem with crystal structure collections on the web is that they have the potential to divert users' attention from the CSD; if users were to fail to appreciate the value of the CSD, it could lead to the database being lost to the community. A collection of files of coordinates is no substitute for the CSD.

WAW: What about the eCrystals project (the archive for crystal structures generated by the Southampton Chemical Crystallography Group and the EPSRC UK National Crystallography Service) [4]? What will be your future relationship with the National Crystallography Service at Southampton?

CG: eCrystals has shown the value of a carefully constructed archive: not just of the results of an experiment, but the data generated and the experiment itself. We will continue to support the UK National Crystallography Service; in fact I have already visited and we are sharing our thoughts as to how we can improve the informatics element of experimental science. We store the results of experiments; Southampton stores the experimental data and methods.

WAW: You have good relationships with some publishers for pre-deposition of structures. Is the American Chemical Society (ACS) a partner?

CG: Indeed. The CCDC and ACS have a long standing relationship. We're keen to facilitate easy access to crystal structures wherever they might be published and are working with the ACS to make sure this continues.

WAW: What about the Royal Society of Chemistry (RSC)?

CG: Again we provide access to crystallographic structures published in all RSC journals. The data is refereed by the RSC, but all structures are also curated by the CCDC.

WAW: Have you made any organizational changes since you took over in October 2008?

CG: I have appointed Ian Bruno to lead the CSD team. Ian has a wealth of experience in scientific software development. He will apply this to ensure that entries for the CSD are identified, curated and distributed as efficiently as possible, with no impact on quality. Our entire software development portfolio will now be led by Jason Cole. He is known to many as a key figure in the development of our docking program GOLD.

WAW: You have about 50 staff. Do you see that number increasing or decreasing?

CG: I don't see the value in growth for growth's sake. If we need more people to provide the tools the community wants, then we may well recruit. Conversely, we strive to be as efficient as possible, in order to reduce the revenue we require.

WAW: In addition to the CSD and related software, what other activities does CCDC have?

CG: Having covered the costs involved in preparing the database we use small revenue surpluses to fund research scientists. Since 1991, the CCDC has funded around 50 Ph.D. students in the area of crystallography. In addition we sponsor many conferences, including initiatives such as the very successful series of Crystal Structure Prediction Blind Tests hosted by the CCDC, and a number of scientific prizes. Through the CCDC travel fund (which was set up using royalties, consultancy fees and the like donated by staff) we also help scientists attend many conferences. Of course CCDC staff continue to be active researchers themselves; the CCDC publishes 20–30 papers a year. [5].

WAW: Do you envisage any diversification in the near future?

CG: Again, we need to take the lead from users. One thing we are keen to do is to maximize the value of the CSD in as many areas as possible. A particular interest of ours at the moment is in materials science. Here the

crystalline form is of fundamental importance, from pharmaceutical materials through to metal organic frameworks for gas trapping. We are also making significant progress in introducing the CSD system to the undergraduate teaching curriculum. A collaboration with Greg Ferrance (Illinois State University) and colleagues in other colleges in the United States is proving very successful. A symposium and workshop at a teaching conference in Indiana in 2008 will be extended at the ACS Fall 2009 Meeting in Washington DC.

WAW: Frank, I have heard you talk about the many structures that never get deposited. Will the situation improve in future?

FA: We have certainly added more donated "unpublished data" in recent years, often from service laboratories with a desire to place structures into the public domain. We also add structures that are starting to appear in repositories like eCrystals. These are, of course, curated in the normal way. In general, though, this remains a thorny issue, since service crystallographers are essentially working for chemists who are presumably the owners of the resulting structures. While there may be good reason not to publish some specific structures, it requires a change of culture for the vast majority of these important studies to be made routinely available to the scientific community.

WAW: You must have been at CCDC for about 40 years. How does it feel to be bowing out?

FA: Thirty-eight actually, so not far off! Well, I'm not bowing out entirely, having become an Emeritus Fellow for which I am very grateful. Principally, this is allowing me to return to CSD-related research which is very welcome. If I can also provide any experience and history that may help in future endeavors, then that'll be great as well. So I still feel very connected to the CCDC, but in a different way.

WAW: Frank, what would you like to see Colin achieve that was not possible while you were in control?

FA: First, greater automation in processing of data to the CSD. Not automation for automation's sake, but automation brought about through intelligent use of the chemical and structural knowledge built up over the past 40 years and more. We have the software building blocks; it remains to form them into a new and upgraded workflow. Colin has mentioned this already as being absolutely fundamental to our future. Data entry compares with a factory production line, but with very highly skilled operatives, our editorial team. They need as much knowledge-based IT help as they can get, to deal rapidly with the more routine structures. They can then concentrate on the really difficult cases, often involving novel structure types. IT helps, for example, by producing chemical names automatically (with tools such as ACD/Name) [6]. There are new tools too, such as wikis and pipelining.

Second, I'd like to see an improved understanding amongst scientists about the various funding models that have provided database coverage in crystallography that is unequaled in any other science. There are issues to be resolved for sure, but they are not helped by communication difficulties. If everything you say immediately appears in some email list, or in some blog or other, then you get very circumspect, to the point of silence, in the face of unremitting proselytizing. This is no way to conduct an important discussion.

WAW: Colin, what are your priorities for 2009?

CG: The top priority is to ensure that the CSD is making best use of changes in technology. We're making many back-end changes to the database, to ease its construction and to allow us to continue to develop search tools. In addition 2009 will see the launch of web access to the CSD.

WAW: Well best wishes for the next 40 years of CCDC!

References

1. Cambridge Crystallographic Data Centre http://www.ccdc.cam.ac.uk. Accessed March 2009
2. Macrae CF, Bruno IJ, Chisholm JA, Edgington PR, McCabe P, Pidcock E, Rodriguez-Monge L, Taylor R, van de Streek J, Wood PA (2008) J Appl Cryst 41:466. doi:10.1107/S0021889807067908
3. Hall SR, Allen FH, Brown ID (1991) Acta Crystallogr A 47:655. doi:10.1107/S010876739101067X
4. eCrystals. http://ecrystals.chem.soton.ac.uk/. Accessed March 2009
5. CCDC publications http://www.ccdc.cam.ac.uk/about_ccdc/company_information/ccdc_publications/. Accessed March 2009
6. ACD/Name http://www.acdlabs.com/products/name_lab/. Accessed March 2009