June 2015

Crystalline

CSD 50 1965 - 2015

The newsletter of The Cambridge Crystallographic Data Centre

Fifty years of sharing crystal structures

Exactly fifty years on from the establishment of the CSD, CCDC Executive Director, Dr. Colin Groom, reflects on its past and its possible future, in this special edition of Crystalline.

It's 1965. The Voting Rights Act is passed in the US but war rages in Vietnam. Work starts on the concrete form of the Berlin Wall and the Sydney Opera House begins to take shape. The Beatles introduce stadium rock concerts, entertaining crowds dressed in newly introduced mini-skirts. Muhammad Ali flattens Sonny Liston with the Phantom Punch. Alf Ramsey prepares his team for England's first (and, sadly, its last) football World Cup victory to follow and Julie Andrews wins an Oscar for her role in *Mary Poppins*.

Meanwhile, communication improves with the introduction of TAT-4, a transatlantic telephone cable capable of handling 138 simultaneous calls. The first combined television and telephone is unveiled in Stuttgart. Amongst others, Woodward, Feynman, Jacob and Monod are awarded Nobel Prizes. Above the earth, space monkey Baker takes a trip in a Jupiter rocket and the first men walk in space.

Olga Kennard,
© Julia Hedgecoe / National Portrait Gallery, London

In Cambridge, Olga Kennard fulfils a dream of herself and J.D. Bernal by attempting to create one of the world's first numeric scientific databases: a database capturing the crystal structures of molecules. They are driven by their passionate belief that the collective use of data would lead to the discovery of new knowledge which would transcend the results of individual experiments[1].

A team collates, extracts and curates the printed coordinates of all 1,500 published crystal structures. The Cambridge Structural Database is born and shared with the world.

Now grown to contain well over ¾ million structures, the database is in use in almost every chemistry department in the world. Pharmaceutical companies depend on it to drive drug discovery projects. Materials scientists design ever more complex 3D network structures. Information from the database underpins all molecular mechanics forcefields and interaction scoring functions allow putative drug molecules to be docked to their protein receptors. So ubiquitous is information from the CSD that the reliance on this incredible community achievement is often overlooked.

And it is a *community* achievement. Yes, the establishment of the Cambridge Crystallographic Data Centre was key to the sustainability of this resource and yes, the hard work and talent of its founders, staff and trustees over many years have been vitally important, but we should commend most the crystallographic community for such exemplary data sharing practises. From the inception of the technique, the results of every published crystal structure have been shared with the CCDC – allowing us to share them with the world. The scientific publishers also play a tremendous role in this. In an often-critical environment, let's not overlook their involvement in facilitating the on-going development of the CSD. By partnering with publishers, we all benefit from informatics systems that allow referees to see crystal structures which, seconds after publication, are then available to all, on computers and mobile devices anywhere in the world. As you read this now, you or anyone else can pick up a smartphone and see all the data associated with any of the structures in the CSD, while its 3D image rotates smoothly in front of you. The founders of the CCDC might have struggled to imagine this technology, but they would undoubtedly be proud to see the CCDC established as the world's leading structural chemistry research institute.

There is a myriad of lessons to be learned from the history of the CSD and the CCDC, and we should all, as community members, allow ourselves a moment or two to feel proud… but it's much more fun to think about the future.

We at the CCDC know that our informatics systems will continue to improve, as will the technology that allows us to process ever more structures of ever more complex molecules. We'll also continue our own structural chemistry research which drives the development of an ever more capable CSD system.

Continued overleaf

The CSD – How it all started

Dr. Jason Cole, CCDC's Deputy Director, reflects on the long history of the CCDC.

Back in the mists of time, the Cambridge Structural Database was created. But, like the big bang, the more reflective amongst you might ponder, "how did this come about?" or "why is the CSD like this, and not like an Oracle RDBM cartridge"? I will try to explain why the CSD began 50 years ago and how it developed to where it is today - to set curious minds at rest.

Swimming in data

The CSD was established by Olga Kennard in 1965 as a result of ideas and thoughts first expressed by J. D. Bernal, 30 years earlier. Kennard and Bernal recognised a need for collecting structural information. At the time, crystal structures were still rare and special, taking many months of effort to elaborate.

Even so, it was recognised that the quantity of data available would soon go beyond the capacity of the average structural chemist's brain, so some form of indexed access to the data would be essential. The first versions of the system were not databases - they were books containing all of the structural information then available in print form.

Very quickly, however, with the advent of automation via computing, the structural universe expanded. The numbers of structures being determined presented curation challenges for the staff of the CCDC. The first 15 years of the CSD's existence were occupied with developing internal systems to make curation efficient and building relationships with journals to act as the trusted repository for data they published. Such work still continues today.

In these early years the data received was not in an electronic format, and reliable optical character recognition systems were yet to be invented, so the majority of time and effort was in manual data transfer from hard copy into electronic form. If you look closely enough into the FORTRAN code developed at the time, you can find algorithms designed to detect and suggest data corrections to manually typed-up coordinates by cross-referencing to typed-up author bond lengths so that the data was self-consistent.

The development of the "CIF" format in the late 80s to mid-90s was partly a response to this type of issue. By having a solid standard in place, crystallographers could move away from hard copy to electronic submission, removing a significant source of error for our editors. That said, hand-edited CIF files are still a source of issues, even today, and expert curation remains essential to maintain the high quality and reputation of the database.

Getting it to the people

As with all rapidly expanding data resources, the database infrastructure needs continuous development and, every now and then, a

Continued from front cover

Looking further into the future, with the aid of the Centre's official crystal ball (feel free to groan), it's interesting to speculate at what point we'll begin to complement our experimental data with predicted crystal structures. We're not far off the one million crystal structures Sarma and Desiraju suggested might be enough to allow knowledge-based crystal structure prediction. Indeed, the continued development of software, computer hardware, radiation sources and instrumentation means we'll reach this point remarkably soon. Our own internal research in this area already shows great promise.

But I think it's nice to live in a world where we can't predict the future, and any attempts to do so in this piece will inevitably come back to haunt me, so I'll resist the temptation. But I'm glad that Olga and the Sage did. I'm even more delighted that their vision of the future came to be – their dream came true and we should all be grateful for this.

1. The Impact of Electronic Publishing on the Academic Community. From private data to public knowledge. Kennard, O. (1996) Portland Press.

2. Sarma, J. A. R. P., and Gautam R. Desiraju. "The supramolecular synthon approach to crystal structure prediction." Crystal growth & design 2.2 (2002): 93-100.

The CSD – Timeline

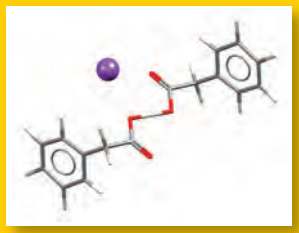
1965 Dr Olga Kennard OBE FRS of the University of Cambridge Crystallography Dept. became the Founder of the CCDC

1967 CDCPPT10 a cis-platin structur

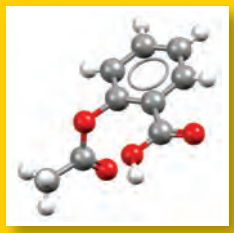
1960

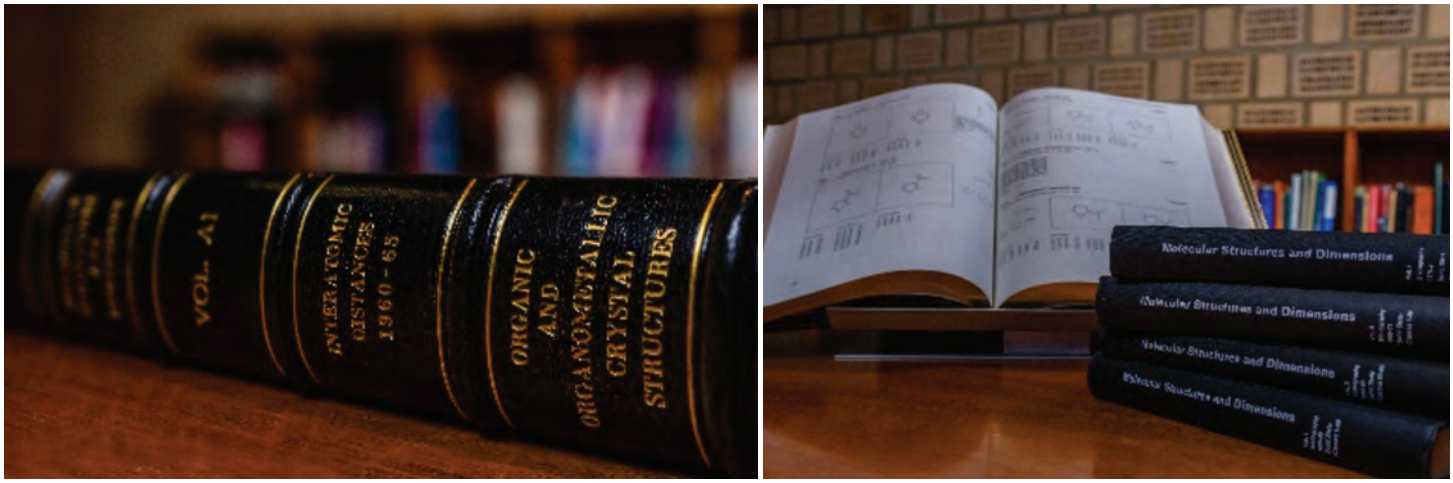
1962 Crick, Watson and Wilkins received the Nobel Prize for DNA Structure

1960 KHDPAC the 1st neutron diffraction structure in the CSD



1963 ACSALA the 1st aspirin structure





Early Version of the CSD

complete rebuild. Before the advent of the internet and networked computers, there was a big problem: how do you get the data to the people? The solution turned out to be a network of affiliated centres (NACs), to act as national sources for the CSD around the world.

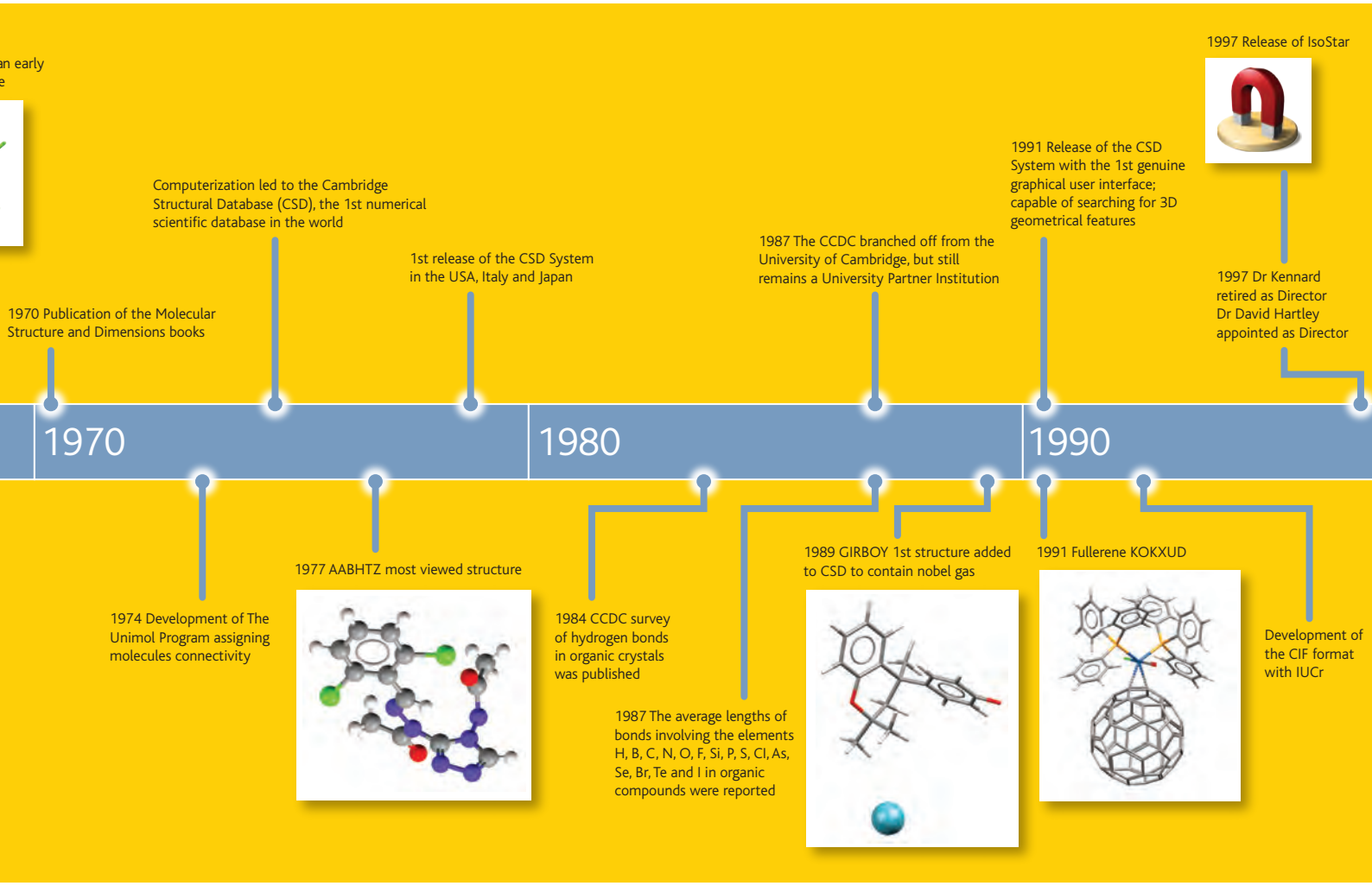
The principle role served by NACs was to reproduce and redistribute the database. The CSD was transferred to magnetic tape in Cambridge. To save on expensive shipping costs, a single copy was sent to each NAC and they reproduced the tapes locally, so that shipping costs to end users were predominantly local rather than international, with the NAC bearing the smaller expense.

Of course, the NACs now serve a rather different purpose. Reproduction of magnetic tapes is thankfully no longer necessary, so they fulfill an enormously valuable role in speaking for their local communities and managing national level agreements to allow unrestricted access to the CSD by all chemists in their country.

What's in a name?

The first "Cambridge Structural Database", wasn't actually called this; in fact it had many names, such as *The Cambridge Crystallographic Data File*, *The Cambridge Crystallographic Database*, *The Cambridge Datafile* or just *The Database*, used within the community it served. The more familiar the user was with the CSD, the less precise the name they tended to use.

This is still the case today; at crystallography conferences in particular, you'll hear speakers



in talks just say *"The Database"* on the safe assumption that everyone knows they mean *The Cambridge Structural Database*, even though there are now many databases that contain structural chemistry and biochemistry content.

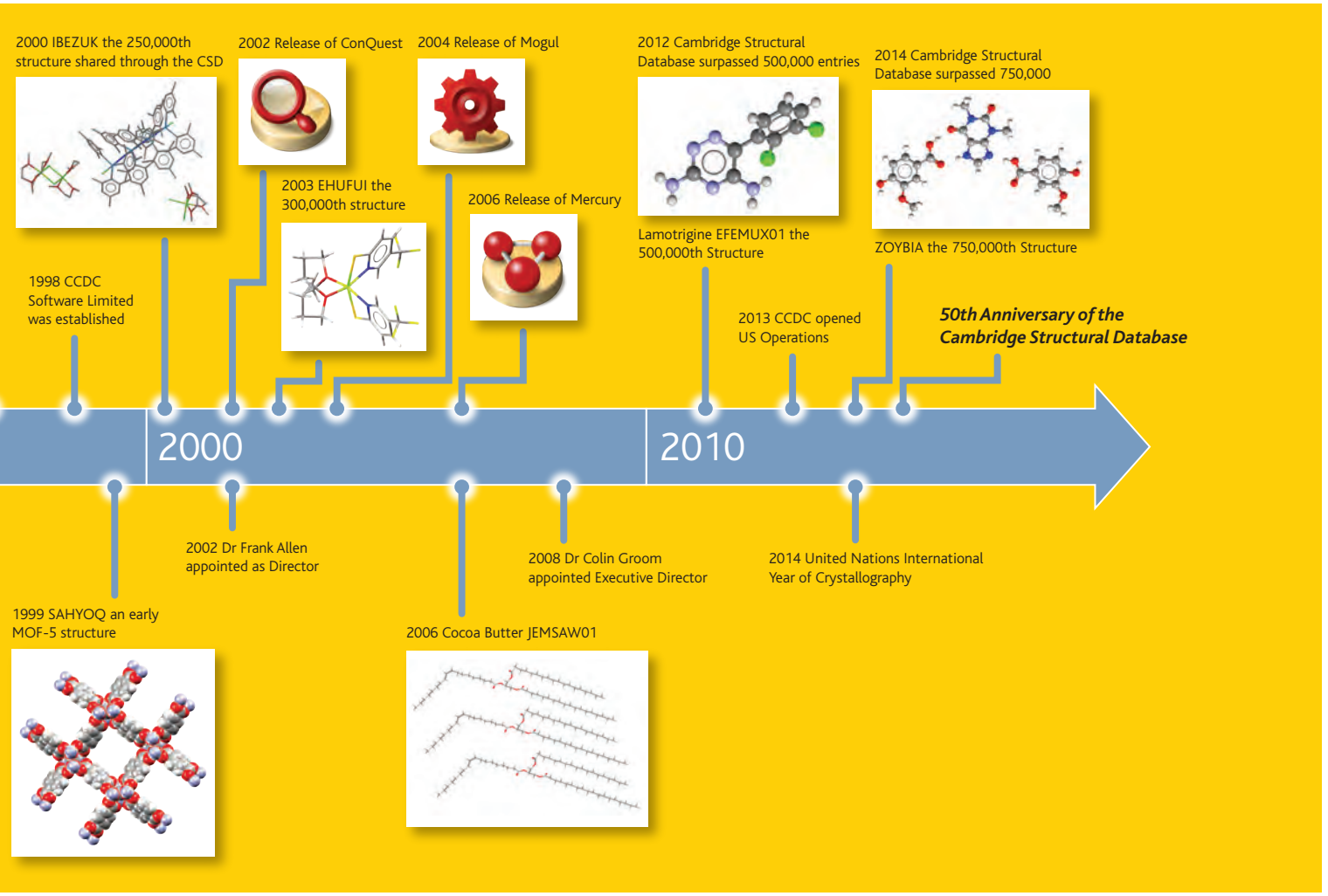
All of the names we used contained *crystallographic;* all of those that originated from within the Centre itself contained *Cambridge*, at Olga's behest. The change to *"Structural"* was an attempt to engage all chemists rather than just specialist crystallographers and to highlight the wealth of 3D data that was available to them. Not all of the CCDC's employees at the time agreed with the change; for example, one felt that it made the database sound like a database of information for architects, but the name *"Cambridge Structural Database"* eventually stuck and became official.

Why isn't the CSD just another relational database?

These days, as software developers, when we think of databases we usually imagine they are built in relational database management systems; so why is the CSD not a relational database? Ultimately the format developed was chosen to minimise size and provide for the most efficient access. Relational databases tended to be good to search and retrieve small packets of information from very large bodies of data by only retrieving the exact data requested and searching based on pre-indexed fields. The CSD had a rather different use case; its users wanted to search and calculate parameters from large volumes of data that didn't naturally fit into a tabular structure with indexing. To have indexed in the relational way would have required very large numbers of indices, each creating an increase in size for the underlying data files. In the 1980s and early 1990s disk space was limited and very expensive, and so instead a system was developed that could index on a relatively small number of 1D and 2D chemical structure features. The indices could be loaded rapidly from disk and so search speeds were very fast, and the size penalty paid was smaller than a classical relational database.

Space saving was a big issue. For example, the CSD's underlying data format involves packing information together where space could be saved. For example; atomic coordinates are "folded-in" with their own standard deviation under the hood, rather than storing two separate numerical values. This worked extremely well but caused a great deal of confusion for young developers (specifically the writer here) as one needed to know the "runes" to reverse the packing process. Nowadays, such considerations might be secondary if we were starting again. We could be far less efficient in our development practices, and just rely on cheap access to fast processors and large memory arrays. (Indeed, recent advances in relational database platform capabilities give us scope to consider this architecture for future CSD generations – so you should watch this space!) The history does, however, help us to understand how the CSD has developed to where it is now – with a fantastically efficient underlying structure, able to cope with the ever-increasing growth in known small molecule crystal data. And we can all be grateful for that!

CSD 50
1965 - 2015



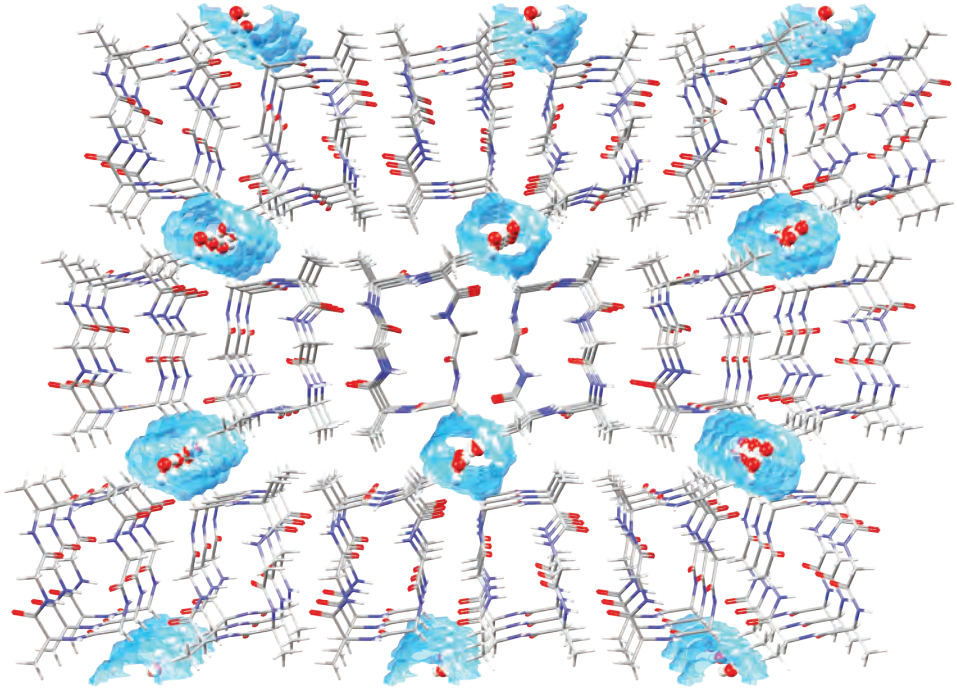
Taking Scientific Communication to the Next Dimension

Scientific communication is frequently seen to be challenging, and this is particularly true for structural sciences, like crystallography. A lot of the concepts that structural scientists try to convey are inherently three-dimensional in nature, such as molecular conformations and intermolecular interactions. The right type of visualisation or physical model can significantly aid the communication of these scientific topics. If a picture is worth a thousand words, then how many is a movie or a 3D physical model worth?

This year, the features available in Mercury for generating high impact structural representations are being enhanced and extended in order to help scientists communicate their work effectively and easily. These options will provide straightforward generation of high quality graphics, movies and even 3D printable model files, allowing users to focus on conveying their science rather than learning a complicated interface or modelling language.

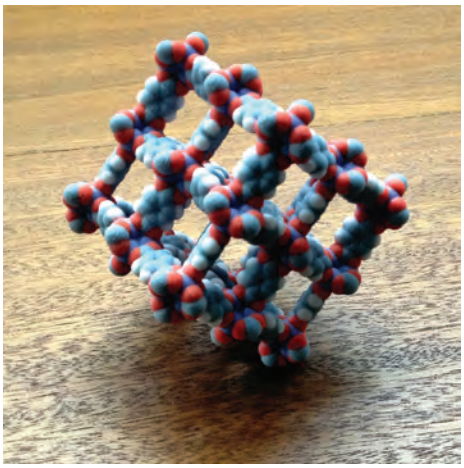
Ray-tracing is a technique for generating images by tracing the path of light and simulating its effects with objects. This method is capable of producing a very high degree of realism and very high impact graphics; in fact most of the images shown in this newsletter are rendered using ray-tracing. For a number of years Mercury has had the ability to output a file representing a structural scene that could be used with POV-Ray (a very effective, and free to use, ray-tracing rendering program) to produce high quality images. An interface has now been developed in Mercury to directly output high resolution ray traced graphics at the click of a button, using POV-Ray behind the scenes. This means that images ideal for important reports, presentations and even journal covers can be generated rapidly and without needing to learn a new interface or scene description language.

High resolution graphics are ideal for posters, papers and reports when structural representations may be inherently restricted to static images, but why be restricted to a static image in a presentation? If you want to show a structural model that is difficult to represent in a 2D image, then a rotating video can be a significant aid. A new intuitive interface in Mercury will allow production of high quality movie files that can be used in presentations or embedded into electronic journal articles without the need to learn any new software or techniques.



High resolution image, rendered using POV-Ray & Mercury, of the water channels in the structure of cyclo(L-alanyl-L-alanyl-glycyl-glycyl-L-alanyl-glycyl) monohydrate (CSD refcode AAGGAG10)

3D printing is still an emerging technology, but it is progressing at a rapid pace – printers are fast becoming financially accessible for companies, universities and even schools. The techniques are ideal for rapid prototyping or for highly bespoke model design. This makes 3D printing really useful for scientific communication where getting across a detailed and specific three-dimensional structural concept can be very challenging.



3D printed model of MOF-5 – a metal organic framework formed from Zn4O nodes and 1,4-benzodicarboxylate linkers (CSD refcode SAHYIK; Photograph: Ilenia Giangreco)

So far, production of model files based on crystal structures has required use of a chain of complex and specialised tools [1-3]. Mercury is now being extended to provide simple, direct output of a 3D printable model file for any atomic structural model including simple molecules as well as extended coordination frameworks or even intermolecular packing arrays. Whether you want a structural model for outreach, education, scientific research or just as a commemorative item for your desk, Mercury will provide easy design of a printable model file.

The new scientific communication features discussed here are available in a software update this summer for all Mercury users (including those using the free version of Mercury).

Dr Pete Wood, Senior Research & Applications Scientist

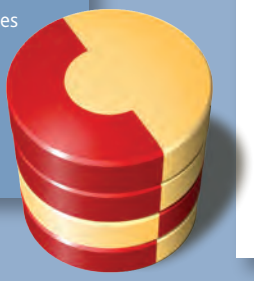
[1] T.-H. Chen, S. Lee, A. H. Flood & O. Š. Miljanić, CrystEngComm, 2014, 16, 5488-5493.

[2] V. F. Scalfani & T. P. Vaid, J. Chem. Ed., 2014, 91, 1174-1180.

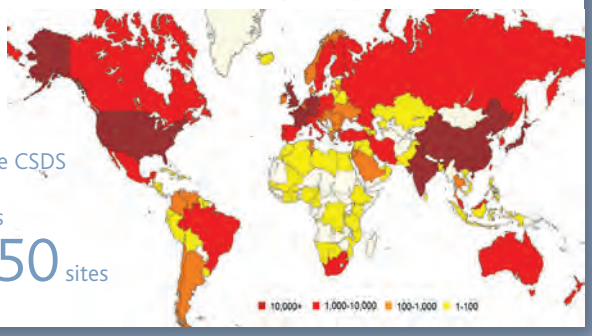
[3] P. J. Kitson, A. Macdonell, S. Tsuda, H. Zang, D.-L. Long & L. Cronin, Cryst. Growth Des., 2014, 14, 2720-2724.

50 years of the CSD in figures

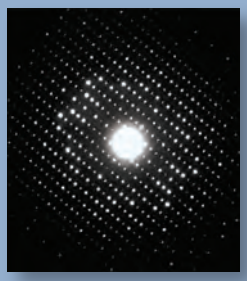
393,529 Publications
1846 Publication sources
18 Languages
303,587 Authors



Number of depositors per country



80 Countries using the CSDS
1,113 Universities
93 Companies at 150 sites



775,654 Entries
1,428,330 Molecules
63,204,673 Atoms
541,706 CSD DOIs

1,952 Publications with the CSD as the topic
76,758 Citations of the of the CSD
8,284 Citing F. Allen et al, (2002)

12,414 Private Communications

Less than 5 seconds to share a structure through the CSD

Events

Date	Conference, Meeting or Event	Venue	Location	Activity
22 Apr 2015	Crystallography for the next generation: the legacy of IYCr	Hassan II Academy of Science and Technology	Rabat, Morocco	Scientific Talk
6 Jun 2015	48th Erice Crystallographic Course; "Engineering Crystallography: From Molecule to Crystal to Functional Form"	Ettore Majorana Centre	Erice, Italy	Scientific Talk, Workshop
14 Jun 2015	The 8th Bologna Convention on Crystal Forms	Zanhotel Europa	Bologna, Italy	Scientific Talk
21 Jun 2015	British Association of Crystal Growth	Queen Mary University	London, UK	Scientific Talk, Exhibition, Workshop
12 Jul 2015	22nd International Conference on the Chemistry of the Organic Solid State (ICCOSS XXII)	Toki Messe	Niigata, Japan	Scientific Talk, Exhibition
19 Jul 2015	Spanish Royal Society of Chemistry (RSEQ) XXXV Biannual Congress	Congresses and Exhibitions Palace of A Coruña (Palexco)	A Coruña, Spain	Scientific Talk
25 Jul 2015	American Crystallographic Association 65th Annual Meeting	Sheraton Hotel	Philadelphia, PA, USA	Scientific Talk, Exhibition, Workshop
16 Aug 2015	250th American Chemical Society (ACS) Meeting & Exposition	Boston Convention & Exhibition Center	Boston, MA, USA	Scientific Talk, Exhibition
23 Aug 2015	European Crystallographic Meeting 29 (ECM29)	Adris Exhibition and Convention Centre	Rovinj, Croatia	Scientific Talk, Exhibition, Workshop
14 Sep 2015	Annual Italian Crystallographic Meeting	Complesso San Giuseppe	Vercelli, Italy	Scientific Talk
20 Sep 2015	XXIII Conference on Applied Crystallography	Czarny Potok Hotel	Krynica Zdrój, Poland	Scientific Talk, Exhibition

Recent CCDC Publications *published since November 2014*

The CCDC team frequently publishes results of their research, which is often the work of collaboration with industrial or academic scientists.

You can find the full list of our publications at www.ccdc.cam.ac.uk/publications. Here are our most recent titles, published since 1st May 2014.

2015

Quantifying the symmetry preferences of intermolecular interactions in organic crystal structures

R. Taylor, F. H. Allen, J. C. Cole, CrystEngComm, 17, 2651-2666, 2015 10.1039/C5CE00035A

Analysis of the conformational profiles of fenamates shows route towards novel, higher accuracy, force-fields for pharmaceuticals

O. G. Uzoh, P. T. A. Galek, S. L. Price , Phys. Chem. Chem. Phys., 17, 7936-7948, 2015 10.1039/C4CP05525J

Small Molecule Crystal Structures in Drug Discovery

C. R. Groom, in Multifaceted Roles of Crystallography in Modern Drug Discovery, NATO Science for Peace and Security Series A: Chemistry and Biology, 107-114, 2015 10.1007/978-94-017-9719-1_9

A review of methods for the calculation of solution free energies and the modelling of systems in solution

R. Skyner, J. McDonagh, C. R Groom, T. van Mourik, J. Mitchell , Phys. Chem. Chem. Phys., 17, 6174-6191, 2015 10.1039/C5CP00288E

Data to knowledge: how to get meaning from your result

H. M. Berman, M. J. Gabanyi, C. R. Groom, J. E. Johnson, G. N. Murshudov, R. A. Nicholls, V. Reddy, T. Schwede, M. D. Zimmerman, J. Westbrook, W.Minor , IUCrJ, 2, 45-58, 2015 10.1107/S2052252514023306

2014

Utilizing Organic & Organometallic Structural Data in Powder Diffraction

J. C. Cole, E. Kabova, K. S. Shankland, Powder Diffr., 29, S19-S30, 2014 10.1017/S0885715614000827

Assessment of a Cambridge Structural Database-driven overlay program

I. Giangreco, T. S. G. Olsson, J. C. Cole, M. J. Packer, J. Chem. Inf. Model., 54, 3091-3098, 2014 10.1021/ci500509y

The Cambridge Structural Database in Retrospect and Prospect

C. R. Groom, F. H. Allen, Angew. Chem. Int. Ed., 53, 662-671, 2014 10.1002/anie.201306438