Case Study: CSD and the CCDC

Prepared for the ESA Workshop on *Creating and Implementing Sustainability Plans for Data Repositories*[1]

Nancy Maron, BlueSky to BluePrint

Background

The Cambridge Structural Database (CSD) began in 1965 as a manually collated aggregation of data from the literature concerning molecular crystal structures. Today, it is available as an online service that permits researchers to deposit their own data and use that of others. Alongside the online offering are desktop software applications that allow more detailed analysis and application of the data and knowledge in the database. The CSD is supported today by the independent not-for-profit organization Cambridge Crystallographic Data Centre (CCDC), which is the entity that manages the processes needed to curate the data, makes data discoverable, and links it to the scholarly record. The CCDC sustains the CSD through a range of revenue generating activities. Some revenue comes through license agreements with the academic sector, but the majority comes from its work with industry. Each year, the CCDC handles over 70,000 data sets deposited by more than 10,000 researchers and facilitates access to data and knowledge by researchers from over 1,300 organizations worldwide in addition to providing public access to the 900,000+ entries in the CSD.

Sustainability history

When the CSD first started, it was funded by grants in the UK. Over time, as it became clear that this funding was eventually coming to an end, its leaders were obliged to consider other sources of support. Revenue subsequently came through licensing the database and associated software to industry, and to academia. Academic revenues typically came through National Affiliated Centres which took care of collecting fees and copying tapes for distribution to institutions within their regions. In 1987, the CCDC became a registered charity in the UK and today it is fully self-supporting, with no support for core activities coming directly from public funds.

From the late 1990s, several successful research collaborations led to the development of specific scientific software applications, one of which in particular became very popular in drug discovery[2]. According to Director of Strategic Partnerships Ian Bruno, this experience opened their eyes to the benefit of "providing more focused applications specifically targeted at a problem. At the time we saw this as a separate revenue stream." Since then these separate applications have become absorbed into suites of software targeted at specific domains, e.g. drug discovery[3] and materials engineering[4].

[1] Information in this article is based on interviews and email exchanges with Ian Bruno, Director of Strategic Partnerships,CCDC, as well as documentation provided by the director and available online.

[2] GOLD: http://dx.doi.org/10.1006/jmbi.1996.0897

[3] https://www.ccdc.cam.ac.uk/solutions/csd-discovery/

[4] https://www.ccdc.cam.ac.uk/solutions/csd-materials/

Sustainability Model Today

Each year, over 10,000 researchers deposit more than 70,000 data sets into the CSD. CCDC employs over 60 full time staff. This includes a team of about 12 editors, over half of whom have PhDs in structural chemistry, to handle the processing and curation of the data. A team of about 17 software engineers develop and maintain the infrastructure used for creating the database, and also develop the software applications that attract the majority of revenue. In addition to administration and management, while there is no commercial sales force, per se, there is a staff tasked with outreach to industry and academia. Of its £6 million budget, staff time and benefits account for about £4.5 million[5].

CCDC's revenue sources come both from academic institutions (30-40%) and from industry (60-70%), primarily in the pharmaceutical and agrochemical industries. The ratio of academic to industrial sites is approximately 12 to 1. While all individual structures in the CSD are free to all, having a subscription allows access to the full collection and the tools needed to systematically search and analyze the data to generate new insights

With the current balance of revenue sources, the funds received from academic institutions are sufficient to support the ongoing maintenance of the core data, while the funds that come in from industry cover the costs of the ongoing development of the database system and the software tools that are valuable to both those in industry and the academy. The CCDC does not see this as a cross subsidization, where private funds are underwriting academic research; rather they see these two strands as complementary. Both are considered necessary for sustainably making data and knowledge as widely applicable as possible.

The CCDC attracts financial contributions from the academy through a number of different channels including:

- Country-wide licenses – these build on the CCDC's network of National Affiliated Centres and make all data and software available free at the point of use to all academics in a region (there are currently 22 country-wide agreements serving 37% of academic sites)
- Campus-wide licenses – these make all data and software available free at the point of use to all members of an institution (39% of academic sites)
- Research Groups and Individual researchers – priced at a level that could be reasonably included in a grant application (24% of academic sites)

The CCDC seeks different levels of contributions from different regions recognizing differences in economic circumstances and levels of research activity. For some countries, the CCDC subsidizes the full cost of access through its FAIRE Programme[6]. Financial contributions from industry are primarily for access to the database and software across one or more sites. Increasingly these arrangements have been part of more profound research partnerships that provides industry with access to CCDC expertise as well as its technology. Access to individual datasets is free to anyone; the financial contributions are for access to value-added services and software built on the aggregated collection of data.

[5] http://apps.charitycommission.gov.uk/Accounts/Ends79/0000800579_AC_20161231_E_C.PDF
[6] https://www.ccdc.cam.ac.uk/Community/FAIRE/

Building in value, and staying a step ahead

Since the turn of the century, as debate around open access to the outputs of research and the benefits of Open Data has intensified, CCDC leadership noted real pushback on the idea of charging for access to the database. As a result, it focused on lowering the barriers (technical as much as financial) to public services providing access to individual datasets free of charge. It has also developed its free search capability that permits people to look up structures, "if they know what they are looking for –for example, because they want to see if a chemical compound has been studied before, or if they are following a link from a scholarly article."

This move – making the data more freely available and more easily discoverable – was not considered to be free of risk. "At the time, we were aware of what others were doing that could be deemed 'competitive.' Some were pulling data sets from the web and trying to offer them in a relatively simplistic way…. So how far can we go to provide the same thing they do and better, while retaining income?" By coming up with a compromise solution, they hoped to benefit the wider community, who would get access to the data, while retaining sufficient value to provide a basis for financial sustainability. If the balance was right then "those other competitors (would) become a non-issue" and CCDC could continue to thrive.

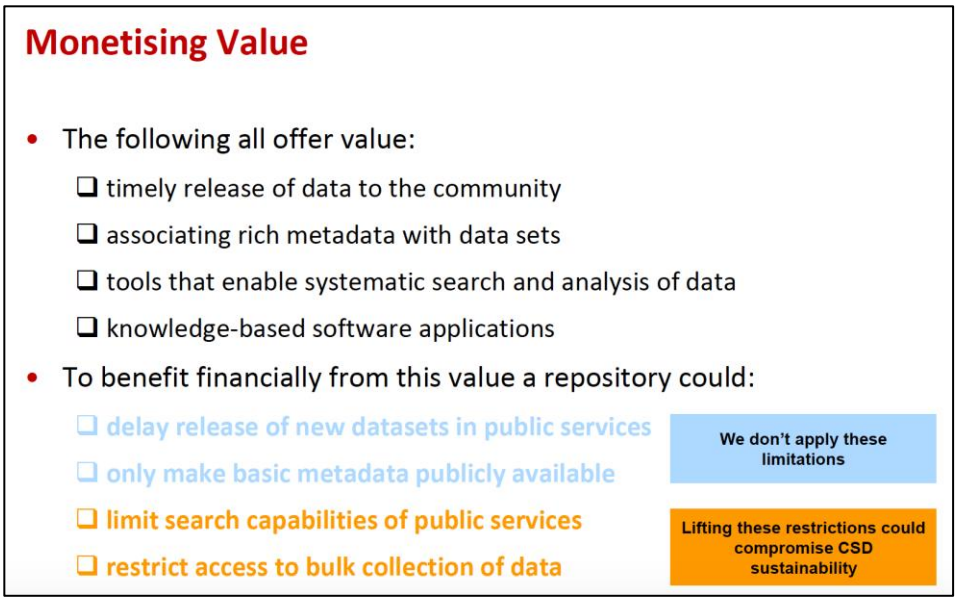
Today, when CCDC leadership think about 'competition' they define this broadly: General purpose repositories for example could be potential threats if people choose only to put their data there – "this would take CCDC out of the workflow and could reduce awareness of the CSD within the community". The solution CCDC seeks is to partner with publishers, offering a means to comply with funder and national mandates to make research data freely available. They highlight their domain expertise as a key benefit and differentiator for scholars who appreciate their more customized approach to managing and sharing their data.

A similarly proactive stance has helped with their industry partners, a critical piece of their sustainability strategy that drives up to 70% of their annual revenue. They have sought to strengthen ties to industry, in particular pharma and agrochemical. Rather than "just licensing software, (we are) making an effort to engage them in the software development process", understanding how we can best make tools available within their workflows and provide support and training tailored to their needs.

As those relationships have grown, CCDC have identified ways to further exploit their expertise, noting a "shift from software-side service to more of consultancy; expert-based service." This is possible, given that the CCDC staff include not only the technical experts who developed the analysis tools, but the PhD-level researchers who understand how to apply them.

Benefits and Tradeoffs of this model

The CCDC leadership talk openly about the tradeoffs inherent in having a subscription model that requires that some functionality be behind a paywall. The slide below, from a recent CCDC conference presentation, outlines their basic argument: some limitations are needed, they feel, in order for them to generate the financial support they need to successfully run CCDC for all those who need it.



Slide from CCDC presentation by Ian Bruno at OECD GSF-CODATA Workshop on Sustainable Business Models for Data Repositories, 3 November 2016 (courtesy of Ian Bruno)

In addition, CCDC view their partnerships with industry not as a necessary evil, but as a real benefit. In addition to the revenue stream they provide, industry partners' focus of applying research is "a good influence – it forces us to make data software in a way that addresses real problems." As Bruno points out, "It is not just about the money, but about their expertise, and the knowledge about the real world problems they are trying to solve. Which feeds back to academia, as well." An example of this was the push by industry partners to develop API-based access to functionality, a move that benefits academic researchers, as well. To make sure they are staying in touch with this segment of their customer base, the CCDC actively engages with this community; at least once a year, the CCDC brings together representatives from companies (consortia) "to us to talk about their experience in using our tools" and to get their input into what future priorities should be.

This focus on commercial partners could potentially put a strain on the relationships the CCDC has with its academic researchers, whose data, after all, is what the firms need to use. To maintain a healthy balance, CCDC "undertakes education and outreach activities focused on academic needs. We interact with academics through crystallography and chemistry meetings worldwide and run workshops. Chemistry librarians (in the US in particular) are key allies, and we make sure we connect with them."

Value of Independence

For the CCDC, revenue generation is meant not just to provide for a minimal operating budget, but to permit the organization to grow as needed, and to plan for the future. "Being independent allows us more flexibility than if we were at an institution." Says Bruno. "We can build up reserves, and investment; This allows us to cover the risk of unforeseen circumstances; student fellowships; and funding for one-off projects."