‘Publishing’ and the Cambridge Structural Database

IUCr Workshop, Osaka, Japan, 23 August 2008

Frank H. Allen

Cambridge Crystallographic Data Centre (CCDC)

12 Union Road, Cambridge CB2 1EZ, UK

allen@ccdc.cam.ac.uk

www.ccdc.cam.ac.uk

Crystal structures in the public domain
So what's new?

- **Original rationale behind scientific databases:**

  "The growing abundance of primary scientific publications and the confusion with which it is set out acts as a brake, as an element of friction, to the progress of science"

  J.D. Bernal (Royal Society Report, London, 1948)

- **Fundamental mission of crystallographic databases:**

  To create comprehensive, value-added and fully validated databases of crystal structure data, with a single-site world repository for each structure type

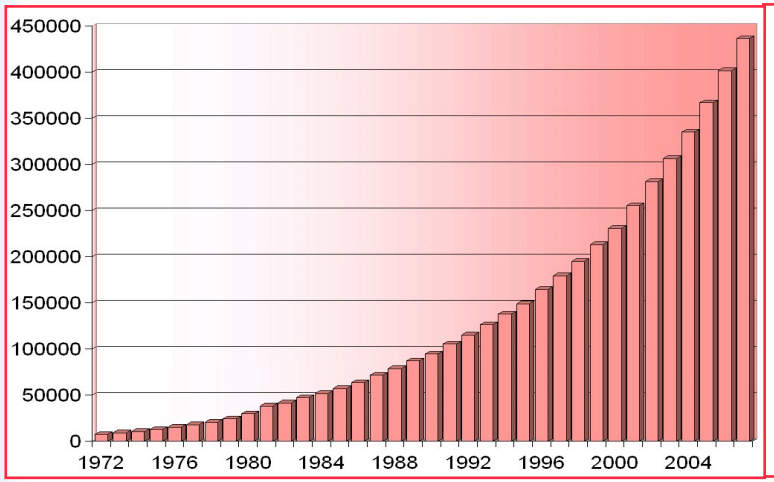
Method of last resort → Method of choice
Crystal structure databases in 2008

January 2008 data			*Total*	*Annual*
CRYSTMET	Canada	Inorganics/Metals	**119,600**	9,000
ICSD	Germany	Inorg./Metals	[>100,000]	[9,000]
CSD	UK	Organic/Metal-Org.	**436,436**	35,000
NDB	USA	Nucleic Acids	**3,730**	500
PDB	USA	Proteins	**48,161**	6,000
All Databases			**607,927**	**50,500**

Growth of the CSD since 1970

"All science is either physics or stamp collecting"

(Lord Rutherford)



Growth 1970 – 2000

453,765 structures

on

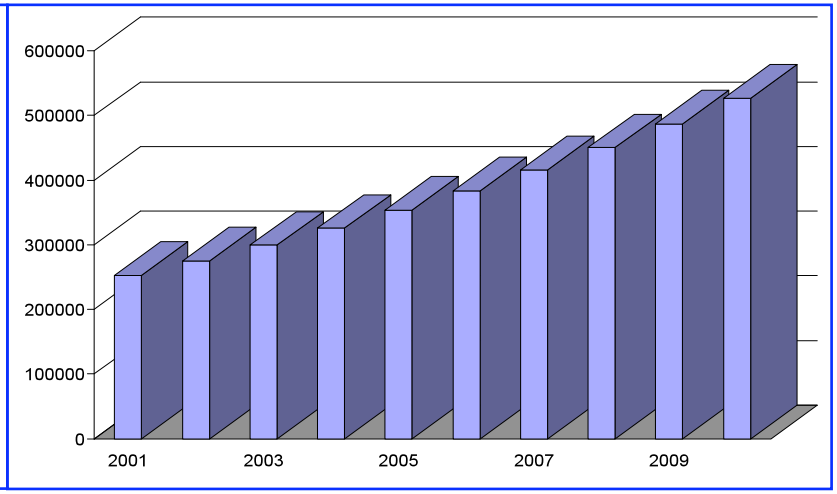
1 July 2008

Projected Growth

2001-2010

>500,000 structures

by end of 2009



CSD: Summary of data acquisition

- **1965 – ca. 1990** Numerical data encoded manually from hard copy journal publications
- **1980 – ca. 2000** Data encoded from journal 'supplementary deposition documents' !
- **1994 onwards** Electronic data deposition via CIF
  Private Communications to CSD (now 5,603)
- **2000 onwards** Appearance of e-only journals
- **2000 onwards** Pre-publication deposition with CCDC for >100 key journals, CCDC archive of deposited CIFs with free access for bona fide researchers
- **In 2008** 1,254 literature sources cited in the CSD
  Top 10 journals yield 49% of structures
  Top 30 journals yield 75% of structures
  99% of data arrive electronically in CIF format

CSD: Vital issues for the future

Internal

- Improving the CCDC's current systems
  - Extend interactions with journals and repositories
  - Improve software as N(struct) increases

External

- **Data repositories and OD archives**
  - Purpose, information content, organisation, oversight, intended user base, data quality, citation in papers?
- **Massive reservoir of unpublished data**
  - How to attract more structures into public domain?
- **Funding data storage and preservation**
  - Sustainability of repositories (and databases)?

Data repositories

- CCDC has always supported the creation of repositories to improve availability of novel structure data: we include data from these e-sources, properly attributed

E-repositories cannot develop in vacuo:

- Who determines content, standards, protocols, formats, quality control (refereeing)?
- Do they extend databases by storing diffraction info.?
- What is relationship to conventional publications?
- Will repositories be updated (new data, references, etc?)
- Will the CCDC have to build up its own list of active repositories, mirroring our journals list? .OR.
- Will there be an overarching ‘federation’ of repositories?

Global Open Data Archives

- Crystallography Open Database (http://cod.ibt.lt/)
  - Entirely self-deposition? Crystal structure results only?
- CrystalEye (http://wwmm.ch.cam.ac.uk/crystaleye/)
  - Aggregates data from journals, expects to aggregate from repositories and will also encourage self-deposition. Crystal structure results only?
- CCDC will pick up relevant novel structures from these sources, but issues noted for local repositories also apply, including oversight, organisation and funding.
- Interactions hampered by the nature of OD polemics!
- Note also:

  IUCr Crystallographic Archive (a current proposal)

Crystal structures in the public domain
Every crystal structure is valuable!

F.H. Allen, *Cryst. Rev.*, 10, 3-15, 2004

An increasing percentage of novel structures are never published in Journals: about 75% are unpublished in many labs.

- As throughput increases, this situation can only worsen
- The log-jam has shifted inexorably to 'placing the data into the public domain', i.e. the 'publication' process
- The scientific community is losing valuable data resources

This is the major challenge facing databases and repositories: how to maximise the number of structures in the public domain

Crystal structures in the public domain
Brakes to the publication process

- Sheer pressure of time – process labour intensive
- 'Ownership' – chemist or crystallographer?
- Responsibility for publication – chemist or crystallographer?
- Structure is not as expected: loss of interest – who 'owns' the data then?
- Refereeing: chemistry is rejected and with it some good crystal structure(s) – what then?

Need for academic recognition or 'kudos', or enforced 'publication' by funding agencies!

Sustainability: Funding and 'Business Models'

- Aggregating, validating, maintaining and deploying databases or repositories requires
  - Funding –management, scientific quality control, hardware etc.
  - Clear expectation of longevity
- Existing databases are funded by:
  - Subscriptions – academia, industry (CSD, ICSD, CRYSTMET)
  - Government agencies – no user charges (PDB, NDB)
- CCDC
  - 1965-1989: public funding, but encouraged to recover costs
  - 1989- : non-profit charitable trust, break-even budget
  - Now: International deployment (70 countries) – subscription, but charitable discounts (up to 100%) for developing countries

Sustainability: Funding and 'Business Models'

- Repositories and OA/OD archives
  - Agency start-up and development funding
  - Long term: international funding, national funding, local funding?
- Issues
  - Government Agencies good at pump priming but not longevity (except the PDB)
  - Long term commitment? - reduces ongoing research resources?
  - Policies may vary dramatically from country to country
  - No clear universal message on establishment and funding of institutional archives
  - Viability: valuable only to specialist crystallographers? or to a broader spectrum of scientists (cf. existing databases)?

We must avoid:

A situation that gives rise to a paraphrase of Bernal:

"The growing abundance of data repositories and the confusion with which they are organised and managed acts as a brake, as an element of friction, to the progress of science"

Acknowledgements

CCDC Staff, May 2008