

Chemists ... and data

Workshop "Data publishing, peer review and repository accreditation: everyone a winner?"

17 January 2013, Amsterdam

RSC and our charter

The object for which the Society is constituted is the general advancement of chemical science and its application and for that purpose:

- to foster and encourage the growth and application of such science by the **dissemination of chemical knowledge**;
- to establish, **uphold and advance the standards of qualification, competence and conduct** of those who practise chemistry as a profession;
- to serve the public interest by acting in an advisory, consultative or representative capacity in matters relating to the science and practice of chemistry; and
- to advance the aims and objectives of members of the Society so far as they relate to the advancement of the science or practice of chemistry.

Where we see the data*

Supplementary information

- “Additional stuff”
- Hugely variable, usually Word/PDF
- Almost all ‘derived’ results vs raw data

Peer review of that data

- Available to reviewers
- Some interchange...
- Comments on spectra, graphs...

but to be honest...

*except crystallographers

- Well defined data format
- Extensive tools for generation, analysis, validation
- Domain repository – CCDC
- Well integrated with publishing
- Community data sharing is the norm

- Request for additional structure data files

Original data at peer review?

- Theoretically great

but issues with reviewers' time, available tools, standard formats

NO requests for more availability of data for review

Data opportunities we've tried

- ChemSpider for community deposition
 - Structures (via standard .mol/.sdf files)
 - Spectra (via .jdx JCAMP files)
 - Reactions with additional 'how to optimise' author input for reproducibility

Not a classic repository, but chemical search engine with data elements

Chemical science researchers?

- Competitive domain
- No community norm of data sharing
- Known difficulties of lab group culture
- Available standards

- “would bury it in the ground if it wasn’t too much effort”
- “our tech transfer office will only let us release data if we’ve proved it’s worthless”

Funder actions and projects

- EPSRC Dial-a-molecule Grand Challenge
- EPSRC Data Management expectations
 - JISC MRD programme
 - PREPARDE

Opportunity

- UK National Chemical Database Service
- EPSRC funded
- Tendered last year
- RSC launched initial service
2 Jan 2013

National Chemical Database Service

- Continuing to offer core database access
- Additionally building a domain repository for chemical sciences
 - Start with what we can do
 - Structured data first
 - Complementary to inst repositories
 - The usual stuff – embargo, standard licences, promoting Open

cds.rsc.org

National Chemical Database Service

The RSC will be operating the EPSRC National Chemical Database Service from 2013-2017

What is the RSC's vision for the Service?

We intend to build the Service for the future - to develop a chemistry data repository for UK academia, and to build tools, models and services on this data store to increase the value and impact of researchers' funded work. We will continue to develop this data store through the lifetime of the contract period and look forward to working with the community to make this a world-leading exemplar of the value of research data availability.

The Service will also offer access to a suite of commercial databases and services. While there will be some overlap with currently provided databases popular with the user community we will deliver new data and services and optimize the offering based on user feedback.

When will the Service be available?

The Service will start on 2nd January 2013, and will be available at cgs.rsc.org

The database services we are working to have available at launch are the Cambridge Structural Database, ACD/ILab and Accelrys' Available Chemicals Directory. The Service will also include integrated access to the RSC's award winning ChemSpider database. As 'live' dates for other services become clear, they will appear here.

How will this affect existing users?

While the CSD and ACD/ILab are continuing, and additional new data sources will be added, access to several of the other existing databases will be ended. Interfaces and access mechanisms will change as we deliver the services differently.

We realise that existing users will be affected by the change in access to many of the databases previously made available through the Service. However, we believe that the components of this new Chemical Database Service will provide greater overall value to the UK academic community.

Available databases

Click boxes for further information and login.

ACD/ILab2 	CSD 
MOS, NPU, CCR 	SPRESIweb 
Available Chemicals Directory 	ICSD 
ChemSpider 	In partnership with the EPSRC 

Data repository

- Use/reuse a priority
- Data availability feeds Dial-a-molecule
- Promotes ELN adoption and use
- Provides data for model building and validation
- Building public APIs and linking to chemical infrastructure

While we're here

- RSC's role in professional development
- Supporting Open Science
- Value returned for data management
- Community norms will change s-l-o-w-l-y

Primary success for CDS?

- When the data is used and reused
 - Use = testing
 - Use = validation
 - Use = visibility
 - Use = reward and recognition
 - Use = acceptance
-
- Building models to create new science

So back to the question

- How do we validate a repository?
- Providing value to build community acceptance of data sharing
- Community validation (from funder and society) + accepted repository good practice

Richard Kidd
kiddr@rsc.org
@rkiddr