Open Data, Open Properties, Open Everything?

A. Ben Wagner, Sciences Librarian

abwagner@buffalo.edu University at Buffalo, NY

University of Toronto Chemistry Librarian Workshop
June 14, 2012

Open Everything

- Open Source
- Open Access
- Open Government
  - http://www.whitehouse.gov/open (Obama's OGI initiative)
- Open Source Religion http://en.wikipedia.org/wiki/Open_source_religion
- Open Cola: http://en.wikipedia.org/wiki/OpenCola_%28drink%29

Open Data

- Open Data
  - The proposition that some data should be freely available to everyone to use, manipulate, and republish.
- Of special interest to chemists
  - Open Property Data
  - Open Spectra
  - Open Crystallographic Data

Useful Reading

Molloy JC (2011) *"The Open Knowledge Foundation: Open Data Means Better Science"*

PLoS Biol 9(12): e1001195.
d**oi:**10.1371/journal.pbio.1001195

- From the *Open Data in Science Working Group* (began 2009); *Open Knowledge Foundation* *http://science.okfn.org/*
- *51% of US researchers – data withholding exerts a negative effect on progress of their research.*

The Resistance Movement

- Nimby effect (Not In My Back Yard)
- Privacy concerns & Getting “scooped”
- Running afoul of ACS’ Ingelfinger Rule
  - http://pubs.acs.org/page/policy/prior/index.html
- Inertia-extra effort-lack of rewards/incentives
- “Good enough” not good enough
- Bad experiences (Burned)

How Open?

- Gratis – free to read
- Libre - free to:
  - Use, repurpose, redistribute, create derivs.
  - Nature of use (commercial/non-commercial)
  - Nature of credit
- If Libre, under what conditions.
  - Creative Commons – Excellent model
  - 6 basic levels
  - http://creativecommons.org/licenses/

Ahead of the game – Envirn. Sci.

- Likely earliest w/Intl. Council of Science forming World Data Centers
  - Archive/distribute 1957-58 International Geophysical Year observational data.
  - Grew to 51 centers in 12 countries
- Reformed into the World Data System in 2009 http://www.icsu-wds.org/
  - Controversial as funding decreased and some data center closed.

Ahead of the game - 2

- Biosciences, especially '…omics'
  - BioMed Central & Public Library of Science
  - Nature Publishing Group (w/caveats)
- Pharmaceutical Science (under public & govt. pressure)
  - Gaulton, A.; Overington, J. P., "Role of open chemical data in aiding drug discovery and design." *Future Med. Chem.* **2010,** *2*, 903-908.
- Big Physics

Open Data - Canada

- Research Data (Science.gc.ca: Research Funding Collaboration: Policies & Guidelines: Open Access:) http://www.science.gc.ca/default.asp?lang=en&n=2BBD98C5-1
- www.toronto.ca/open (who knew?)
- Open Data Pilot Project www.data.gc.ca/
  - Ex: chemicals

NSF – Not really open data - 1

Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples… created …under NSF grants. …where essential, exceptions to this sharing expectation may be to safeguard the rights of individuals and subjects, the validity of results, or the integrity of collections or to accommodate the legitimate interest of investigators.

- http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/aag_6.jsp#VID4

NSF – Not really open data - 2

Proposals must include a supplementary document of no more than two pages…[that] should describe how the proposal will conform to NSF policy [previous slide] and may include:

- the types of data, samples, physical collections, [etc.] to be produced in the course of the project;
- the standards to be used for data and metadata format/content…
- policies for access & sharing incl. provisions for appropriate protection of privacy, confidentiality, security…
- policies and provisions for re-use, re-distribution, & the production of derivatives;
- plans for archiving data, samples… & for preservation of access to them.

Open Data–Chemistry–Blue Obelisk

- Began 2005 - 229th National ACS Meeting
- In response to lack of Open-Data/O-Standards/O-Source in Chemistry.
- Goal: Catch up w/other disciplines
- Focus: molecular, reaction, computational chemistry, spectra, & crystallography.

Open Data–Chemistry–Blue Obelisk

- Success - United researchers w/common interests
- Failed – Reaching wider chemistry community.
- Question: How to best engage the mainstream?
-

Open Data–Chemistry–Blue Obelisk

For more info:

- www.*blueobelisk.sourceforge.net/*
- “Open Data, Open Source and Open Standards in chemistry: The Blue Obelisk five years on.” *Journal of Cheminformatics,* 2011, 3:37. doi:10.1186/1758-2946-3-37

Open Data – Chemistry

- Patents are open data!
- ??

Open Property Data

- NIST Data Gateway
  http://srdata.nist.gov/gateway/
  - NIST Chemistry WebBook
    http://webbook.nist.gov/chemistry/

- ChemSpider http://www.chemspider.com/
- MatWeb* http://www.matweb.com/index.aspx

* Open submission/crowd-sourced

Open Spectra

- Spectra Data Base System for Organic Compounds (SDBS)* http://riodb01.ibase.aist.go.jp/sdbs/cgi-bin/cre_index.cgi?lang=eng
- Sigma-Aldrich Product Catalog Advanced Search http://www.sigmaaldrich.com/catalog/AdvancedSearchPage.do
- Spectra Online* http://www.spectraonline.com/

* Open submission/crowd-sourced

Specialized Open Spectra - Sample

- **Atomic & Microwave Spectra:** *Diatomic Spectral Database* (NIST) – 121 diatomic molecules
- **NMR:** *NMRShiftDB**
- **Geosamples:** *USGS Digital Spectral Library* (U.S. Geological Survey) - UV/Near-IR spectra of minerals, mixtures, artificial, liquids and volatiles and vegetation.

* Open submission/crowd-sourced

Open Crystallography

- *American Mineralogist Crystal Structure Database*
  - *American Mineralogist & The Canadian Mineralogist*
  - *European Journal of Mineralogy*
  - *Physics and Chemistry of Minerals*
- Crystaleye – web spider http://wwmm.ch.cam.ac.uk/crystaleye/
- *Crystallography Open Database** - 120K structures
- *Database of Zeolite Structures*
- *Reciprocal Net** (part of NSDL)

* Open submission/crowd-sourced