The Big Data Challenges Associated with Building a National Data Repository for Chemistry

Antony Williams

ICIC Meeting, Vienna

October 14th 2013

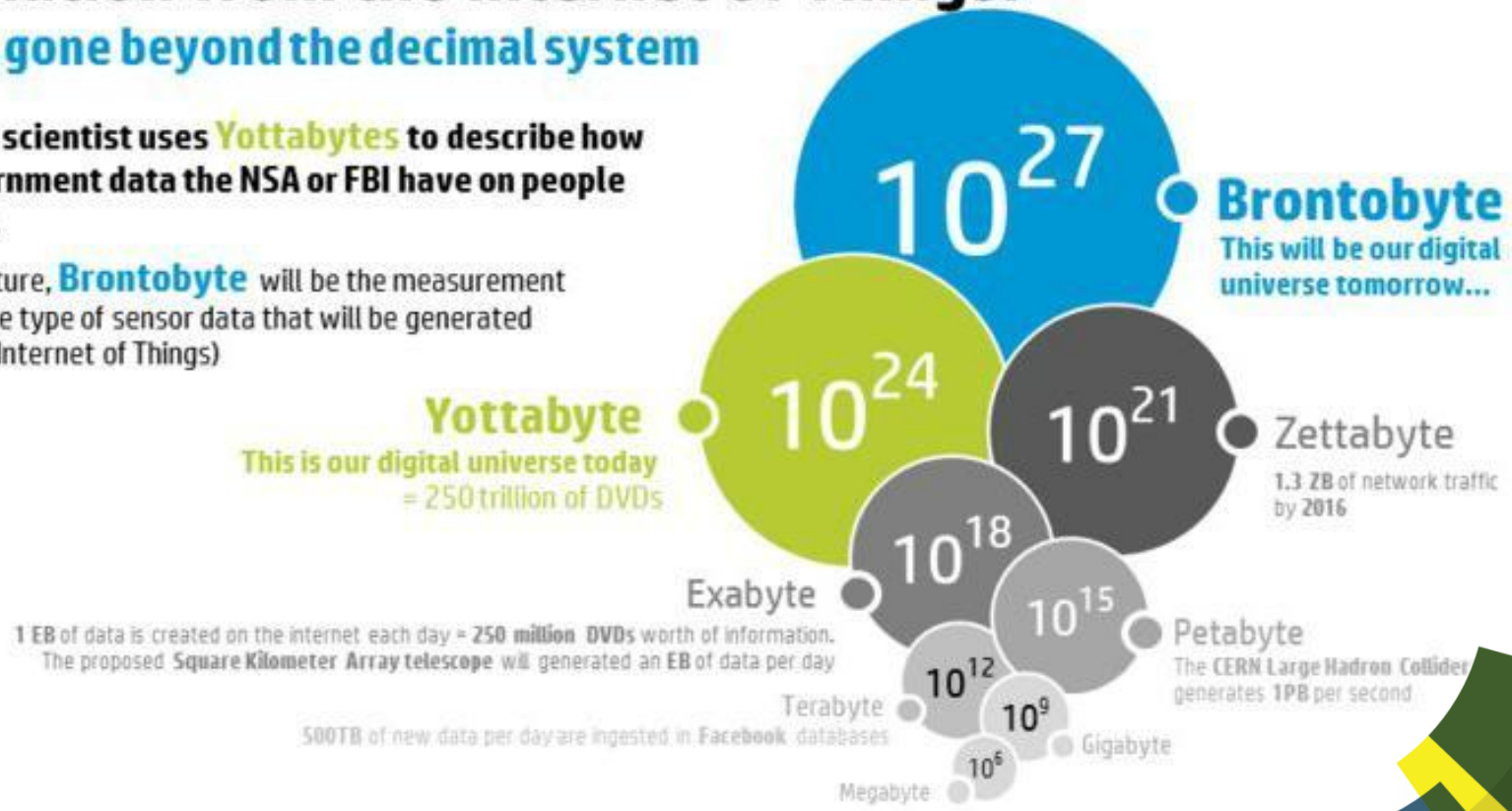
So what is all this Big Data?

Information from the Internet of Things:

We have gone beyond the decimal system

Today data scientist uses Yottabytes to describe how much government data the NSA or FBI have on people altogether.

In the near future, **Brontobyte** will be the measurement to describe the type of sensor data that will be generated from the IoT (Internet of Things)




IN 60 SECONDS...
1 NEW DEFINITION IS ADDED ON urban dictionary
1,600+ READS ON Scribd.
13,000+ HOURS MUSIC STREAMING ON PANDORA
12,000+ NEW ADS POSTED ON craigslist
370,000+ MINUTES VOICE CALLS ON skype
98,000+ TWEETS
20,000+ NEW POSTS ON tumblr.
THE LARGEST SOCIAL READING PUBLISHING COMPANY!!
New Craigslist Ads
320+ NEW twitter ACCOUNTS
100+ NEW LinkedIn ACCOUNTS
13,000+ iPhone APPLICATIONS DOWNLOADED
1 associatedcontent NEW ARTICLE IS PUBLISHED
THE WORLD'S LARGEST COMMUNITY CREATED CONTENT!!
QUESTIONS ASKED ON THE INTERNET...
100+ Answers.com
40+ YAHOO! ANSWERS
600+ NEW VIDEOS
6,600+ NEW PICTURES ARE UPLOADED ON flickr
50+ WORDPRESS DOWNLOADS
25+ HOURS TOTAL DURATION
70+ DOMAINS REGISTERED
60+ NEW BLOGS
168 MILLION EMAILS ARE SENT
694,445 SEARCH QUERIES
1,700+ Firefox DOWNLOADS
695,000+ facebook STATUS UPDATES
125+ PLUGIN DOWNLOADS
1,500+ BLOG POSTS
79,364 WALL POSTS
510,040 COMMENTS
Google
Google Search
GO-Globe.com web technologies

And the World of Chemistry?

Media Releases

CAS Registers 70 Millionth Substance Just 18 Months After Reaching 60 Millionth Milestone

December 6th, 2012

Patents from Asian countries continue to be the leading source of chemistry disclosures

And the World of Chemistry?

Journal of Cheminformatics

InChI in the wild: an assessment of InChIKey searching in Google

Christopher Southan

"The InChIKey indexing has therefore turned Google into a *de-facto* open global chemical information hub by merging links to most significant sources, including over **50 million** PubChem and ChemSpider records."

And the World of Chemistry?

REAXYS®

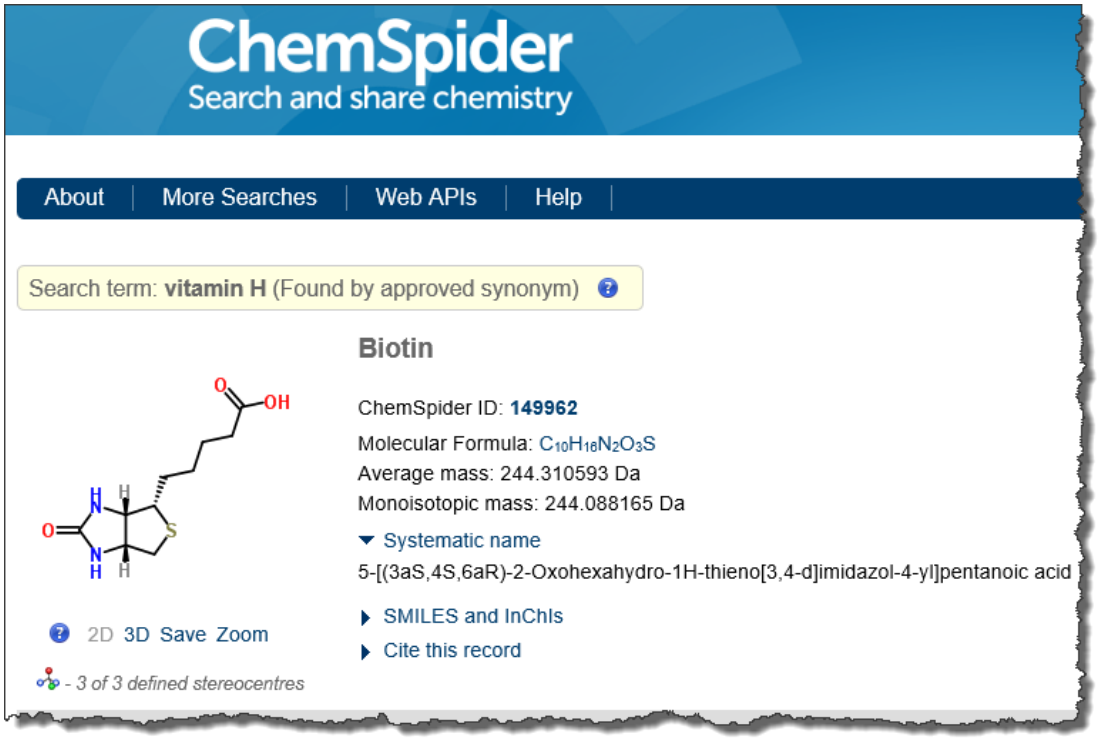
What does the Reaxys Chemistry Discovery Engine offer?

Essential and relevant chemical data

Gain access to over 16,000 periodicals containing 500 million experimentally verified facts.

RSC's ChemSpider

>29 million chemicals from >500 sources



…and the world of Openness

Times have changed…

Open Access funder mandates…

UK Funding Bodies Mandate Open Access

SHERPA/JULIET

Research funders' open access policies

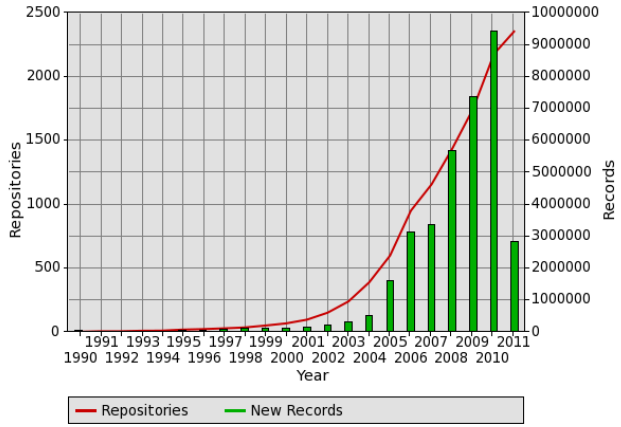
Office of Science and Technology Policy

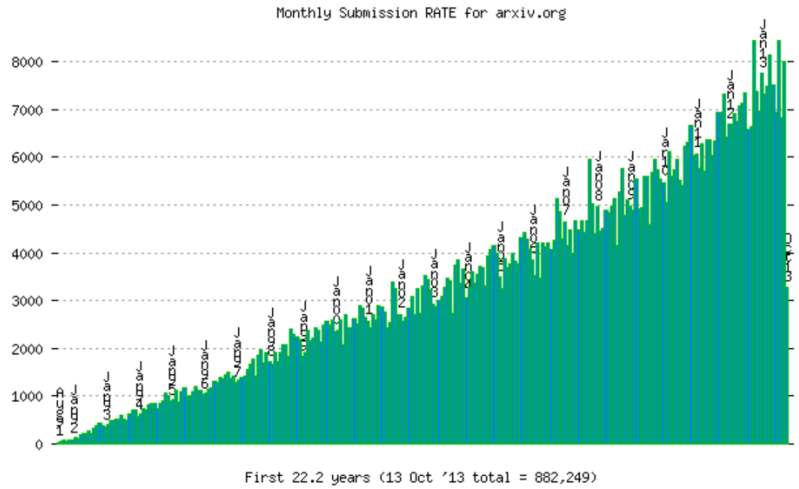
Expanding Public Access to the Results of Federally Funded Research

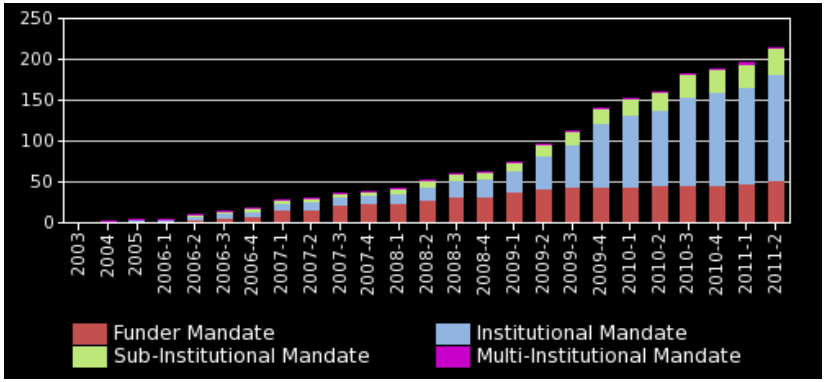
Posted by **Michael Stebbins** on February 22, 2013 at 12:04 PM EDT

Times have changed…

Growth, growth, growth…







Publishers are responding

Open Access at the RSC

Information on Open Access and RSC Policy

RSC launches £1 million Gold for Gold initiative as Open Access transition begins

18 July 2012

ELSEVIER

open access

WILEY Open Access

Open Access at Springer

Learn more about Springer's open access options, including Open Choice and our SpringerOpen portfolio!

Springer

The world of Open Data…

RDA
RESEARCH DATA ALLIANCE

Research Data Sharing without barriers

Open Data are everywhere

- Is Openness and Social Sharing changing the world?
- The cultural experiments in Open Data and exchange are almost daily
- Mobile platforms enhance participation

- **And then what of Chemistry Data???**

Publications-summary of work

- Scientific publications are a summary of work
  - Is all work reported?
  - How much science is lost to pruning?
  - What of value sits in notebooks and is lost?
  - Publications offering access to “real data”?
- How much **data** is lost?
  - How many compounds never reported?
  - How many syntheses fail or succeed?
  - How many characterization measurements?

About Me…as a Chemist

- I've performed a few dozen chemical syntheses
- I've run thousands of analytical spectra
- I've generated thousands of NMR assignments
- I've probably published <5% of all work
- Most of it has been lost
- But things can be different today….
- But it still needs to be associated with me…

What of non-abstracted data?

- How much data generated in a lab, that **COULD** go public, is lost forever?

What of non-abstracted data?

- How much data generated in a lab, that **COULD** go public, is lost forever?
- Public Domain reference databases of value?
  - Syntheses
  - Properties
  - Spectra and CIFs
  - Images
  - Raw data vs. representations of data

ChemSpider

- ChemSpider allowed the community to participate in linking the internet of chemistry & crowdsourcing of data
- Successful experiment in terms of building a central hub for integrated web search
- More people are “users” than “contributors”
- Yet basic feedback and game-play helps

Crowdsourced “Annotations”

- Users can add
  - Descriptions, Syntheses and Commentaries
  - Links to PubMed articles
  - Links to articles via DOIs
  - Add spectral data
  - Add Crystallographic Information Files
  - Add photos
  - Add MP3 files
  - Add Videos

An EPSRC Call

EPSRC NATIONAL CHEMICAL DATABASE SERVICE

Issue date: 06 Jan 2012

"…the identification of the need for a UK national service for the provision of a searchable, electronic chemical database for the UK academic research community."

National Chemical Database Service

- Service for UK Academics
- “Prepaid access” integrating commercial databases and services
- Access to curated data sets
- Provision of prediction algorithms

National Chemical Database Service

ACD/I-Lab2

Physicochemical, ADME and toxicity property prediction (ACD/Labs Inc.).

Further information

CSD

Organic and organometallic crystal structures (CCDC).

Further information

DETHERM

Database of thermophysical data for pure substances and mixtures.

Further information

ICSD

>160,000 inorganic and related crystal structures (FIZ Karlsruhe GmbH).

Further information

Available Chemicals Directory

Provides supplier information for building block molecules.

Further information

ARChem

Retrosynthetic tool for chemical analysis of target organic molecules.

Further information

Chemicalize

Physicochemical property prediction tools with Lipinski-like filters.

Further information

ChemSpider

An online database of molecules from >400 datasources (RSC).

Further information

SPRESIweb

Online chemical structure and reaction database (InfoChem GmbH).

Further information

In partnership with the EPSRC

National Chemical Database Service

- Service for UK Academics
- “Prepaid access” integrating commercial databases and services
- Access to curated data sets
- Provision of prediction algorithms

- Ultimate goal is to federate search
- Development of “data repository”

Development of Data Repository

- Data repository should not just be a data dump – should not be a “big disk”
- Searchable, integrated, segregated repository of data types
- Data access including private, shared embargoed and public
- Delivery of derived models from data
- Integrated to AltMetrics models

What can drive participation?

- What can drive scientists to participate and contribute?
  - Ensuring provenance of their data for reuse
  - Mandates from funding agencies
  - Improved systems to ease contribution
  - Additional contributions to science
  - Improved publishing processes
  - **Recognition** for contributions

AltMetrics

Altmetrics

From Wikipedia, the free encyclopedia

Altmetrics are new metrics proposed as an alternative to the widely used journal impact factor and personal citation indices like the *h*-index.

AltMetrics


Impact
usage
downloads
views
peer-review
expert opinion
citations
alt-metrics
storage
links
bookmarks
conversations

AltMetrics as Scientist Impact

AltMetrics

Altmetric

ImpactStory.

Share the full story of your research impact.

ImpactStory is your impact profile on the web: we reveal the diverse impacts of your articles, datasets, software, and more.

Plum™ Analytics

Measuring Research Impact

Plum Analytics

PlumX Groups Researchers Search

Home / Antony Williams

Show Profile Data Embed Widget

Antony Williams

Connections in Chemistry

Links:

LinkedIn, ScientistDB, ChemConnector Blog, Twitter, about.me, Google Scholar, Microsoft Academic Search, Impact Story, Wikipedia, SlideShare, YouTube, Mendeley, PROskore, ResearchGate, amazon.com, Vizify, visualize.me, Pinterest, ORCID

Researcher from:

Sample Profiles / Royal Society of Chemistry

My passion is connecting people to chemistry. Over the past decade I hel…

tony27587@gmail.com | 919-201-1516

Artifact Summary

Presentation	Article	Paper	Other	Video
268	118	38	37	18

All (505) Presentation (268) Article (118) Paper (38) Other (37) Video (18) Figure (11) Data (4) Book (4) Poster (3) Media (2) Patent (2)

Plum Analytics

Year	Title	Type	All				
			Captures	Citations	Social Media	Mentions	Usage
2013	Dispensing processes impact apparent biological activity as determined by computational and statistical analyses.	Article	20		65		6140
2013	Engaging participation from the chemistry community	Presentation	1		10		821
2013	Approaches for extraction and digital chromatography of chemical data	Presentation	1		5		453
2013	Chemical Database Projects Delivered by RSC eScience	Presentation	1		12		1053
2013	How to build an online profile as a scientist	Presentation	6		20		1255
2013	Magnetism Inside the Human Body: Lessons for Ten Year Olds	Video			8		194
2013	Online social networking for the chemical sciences	Presentation	1		11		1164
2013	Challenging cajoling and rewarding the community for their contributions to online chemistry	Presentation	2		10		1568
2013	Digitally enabling the RSC archive	Presentation	1		4		1438
2013	How ACDLabs Software Tools are used by the Royal Society of Chemistry	Presentation			7		1050
2013	How to Build an Online Profile as a Scientist	Presentation			10		218
2013	Leading Scientists into Openness	Presentation			9		497

Rewards and Recognition

The First Step badge is awarded when a user submits (& has published) their 1st CSSP article.

Congratulations! Your 1st CSSP article has been published. Philosopher Lao Tzu said “A journey of a thousand miles begins with a single step”. In the same way we hope that this will be the first of many submissions that you make to CSSP.

My Profile

Username	Anish Mistry
Security token	sdjkfnsjgneagrgn;ngf;nhrejk
Display Name	Anish Mistry
*Email	made-up@email.com
*First name	Anish
*Last name	Mistry
*Company or Institution	University of Warwick
Street address	The Campus, just outside town
City	Warwick
Province/State	Warwickshire
Postal code/Zip	WA4 1TT
Country	UK
Phone	+44 (0)1111 414141
ORCID Id	0000-0001-5635-0000 Go to ORCID.org
Your roles	Depositor
Research group (Datasource)	The Fox Group

my LinkedIn profile

Awarded Badges

Profile Stats

Leaderboard Position: 6/40
Number of Published Articles: 11
Articles Published:
Dehydration of 3,4-dihydro-5H-Benzo[cd]pyren-5-ol
Reduction of 3,4-dihydro-5H-benzo[cd]pyren-5-one
Chlorination of a carboxylic acid
Hydrolysis of Ethyl 3-(1-pyrenyl)propanoate

Number of compounds mentioned: 61

AltMetrics Feeds

- For our data repository ensure contribution of data will feed out to the AltMetrics platforms
- Every data point, every data download, use and reuse will be associated with the scientist
- Data will be DOI’ed (presently under review)
- Services provided will allow for AltMetrics use

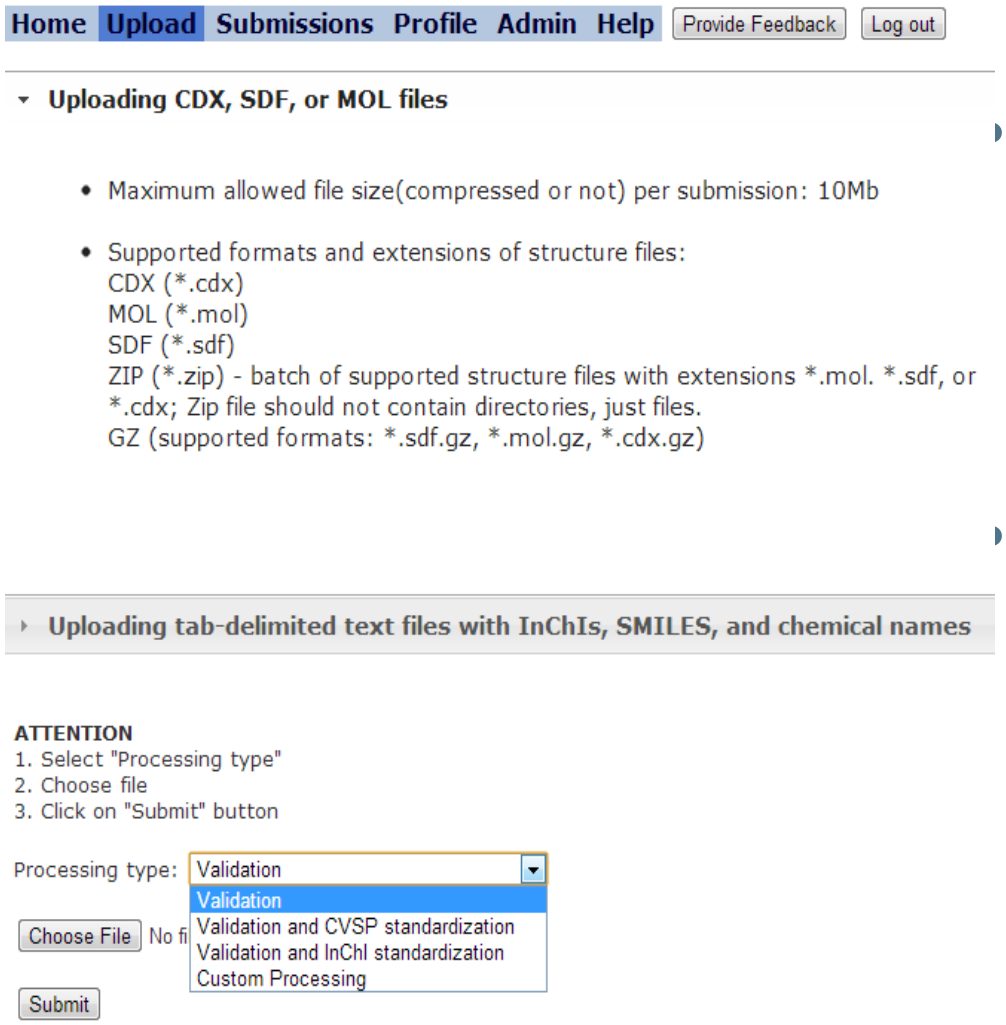
Domain Specific Challenges

- Creating a platform of value not just dumping
  - Searchability, segregation, tagging, use and reuse, collaboration, low barrier to participation
- Quality of chemistry data at source
  - ensuring chemicals are correct
  - reactions map and balance as appropriate
  - file format handling for analytical data types – binary file formats are proprietary
  - valid interpretation of data

Domain Specific Challenges

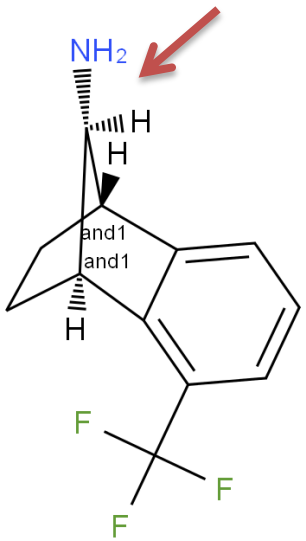
- Quality of data at source
  - ensuring chemicals are correct - VALIDATION
  - reactions map and balance as appropriate – VALIDATION and STANDARDIZATION
  - file format handling for analytical data types – binary file formats are proprietary - STANDARDIZATION
  - valid interpretation of data – VALIDATION and ANNOTATION

Validating Chemicals

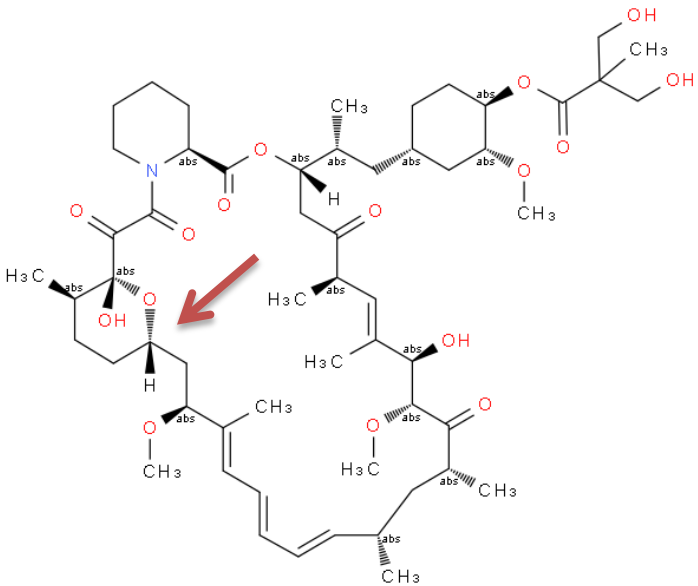


- Community service for validation and standardization of chemicals (CVSP)
- Open rules sets but standard set based on FDA substance registry system

Validating chemicals

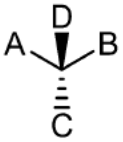


DB08128



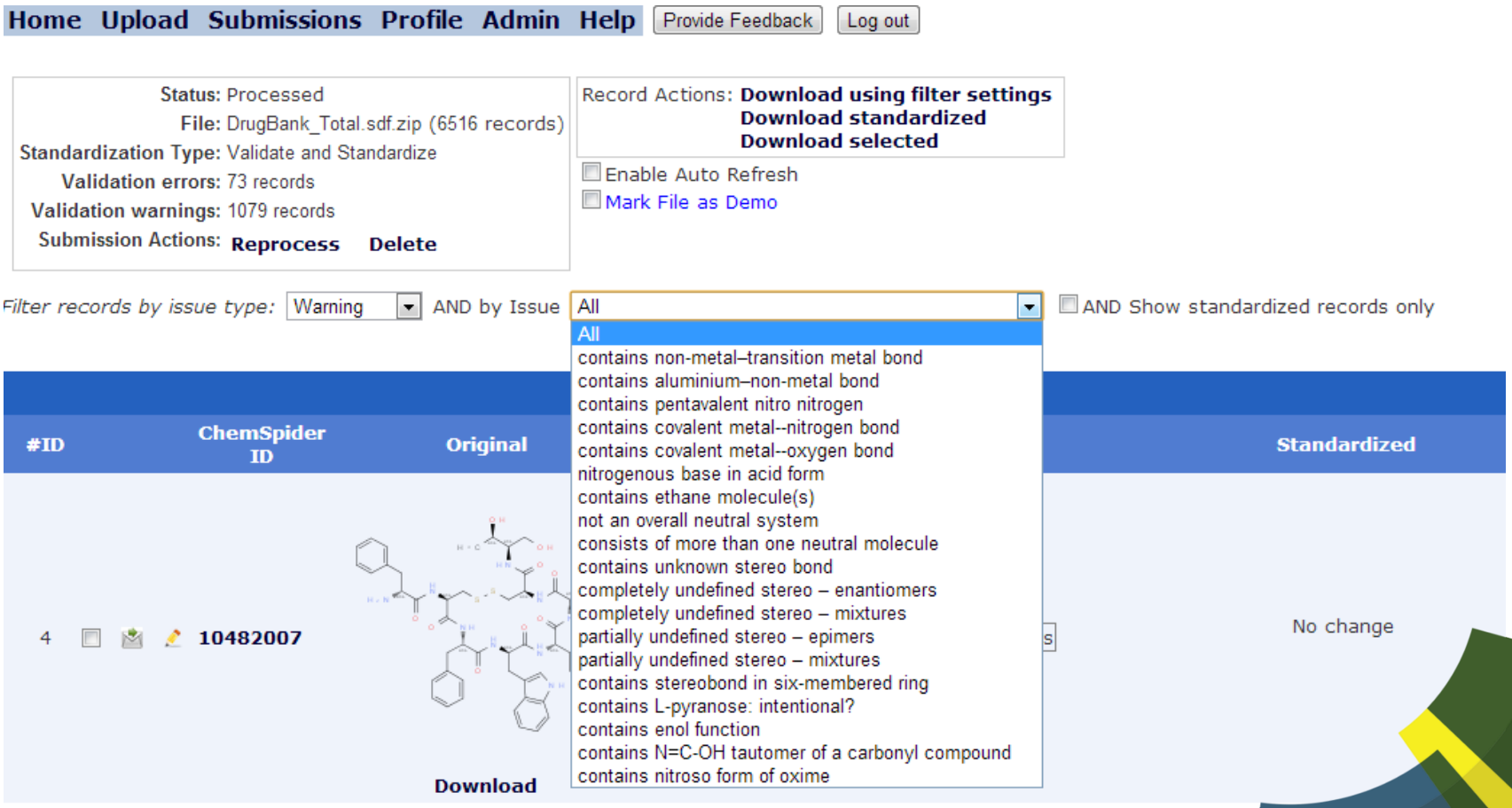
DB06287

J. Brechner, IUPAC Graphical Representation of stereochem. configurations Section: ST-1.1.10



Not acceptable

Standardizing Chemicals



Validated Name-Structure dictionaries for data checking

- Chemical name dictionaries used for:
  - Text-mining (publications, patents)
  - Linking to other databases – think Biology
    - Drug names are incredibly valuable links
  - Searching the web
    - Names link to structures

Difficult to navigate…

Chemical genetics reveals a complex functional ground state of neural stem cells

Phedias Diamandis[1–4], Jan Wildenhain[4], Ian D Clarke[1,2], Adrian G Sacher[1,2], Jeremy Graham[1,2], David S Bellows[3], Erick K M Ling[1,2,5], Ryan J Ward[1,2,5], Leanne G Jamieson[1,2,5], Mike Tyers[3,4] & Peter B Dirks[1,2,5,6]

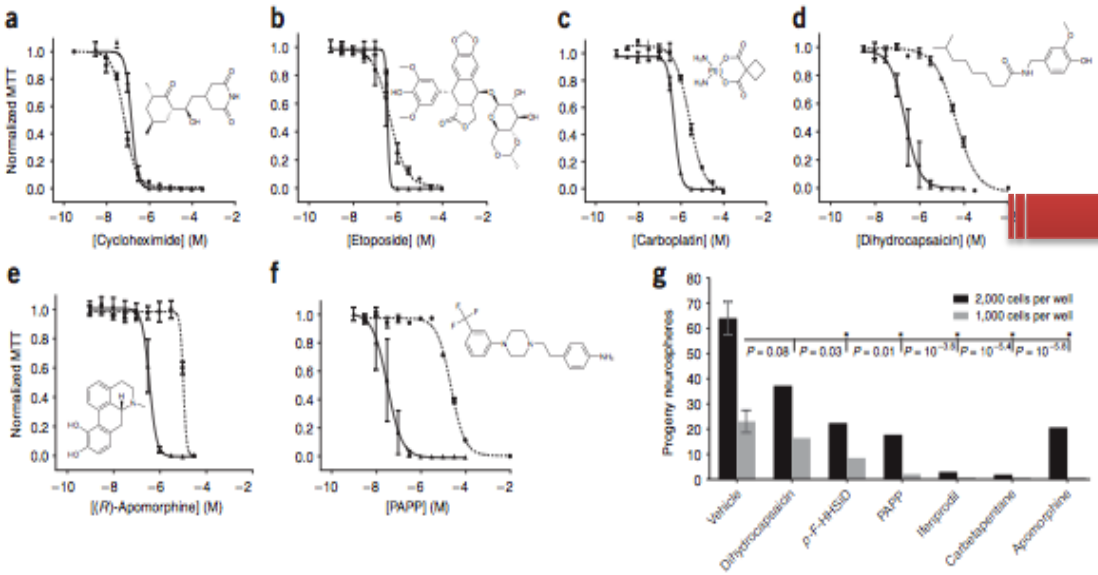


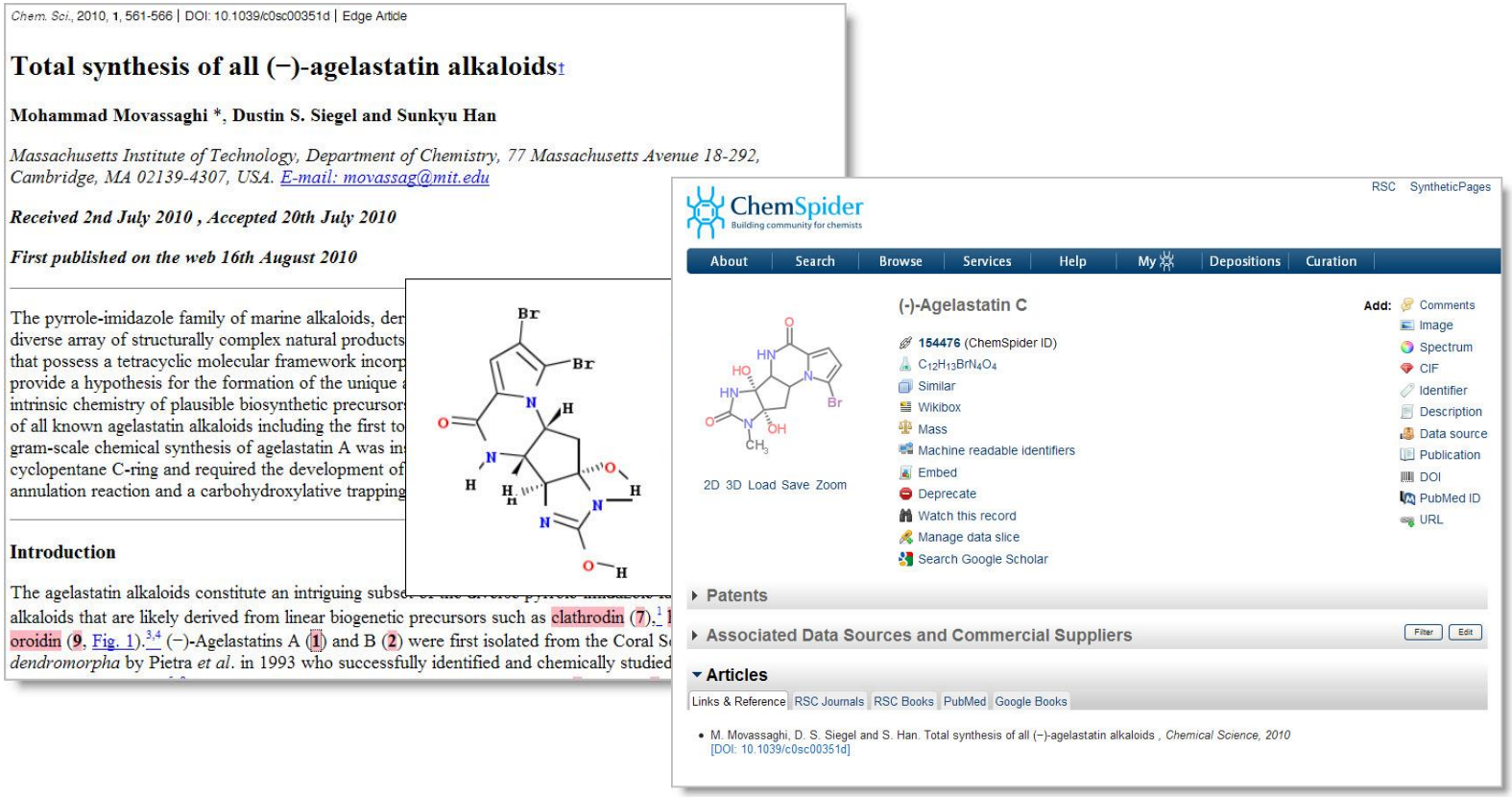
Figure 2 Identification of potent NPC-specific compounds. (**a**–**f**) Dose-response curves and chemical structures of controls: cycloheximide (**a**), etoposide (**b**) and carboplatin (**c**), and of selected newly identified compounds: dihydrocapsaicin (**d**), apomorphine (**e**) and PAPP (**f**). Each plot shows the fitted sigmoidal logistic curve to MTT proliferation assay readings of both astrocytes (- -●- -) and neurosphere cultures (–▲–). Values represent the mean and

Inside our Publication Archive

- How much **data** is in the archive, in the publications and in the supplementary info?
  - How many compounds for ChemSpider?
  - How many syntheses for ChemSpider reactions?
  - How many characterization measurements?
    - Property Data
    - Spectral Data
    - Graphs and charts to be used for modeling?

What if we could capture it all? Digitally Enhancing the RSC Archive

Linking Names to Structures



Semantic Mark-up of Articles

Chem. Sci., 2010, **1**, 561-566 | DOI: 10.1039/c0sc00351d | Edge Article

Total synthesis of all (−)-agelastatin alkaloids†

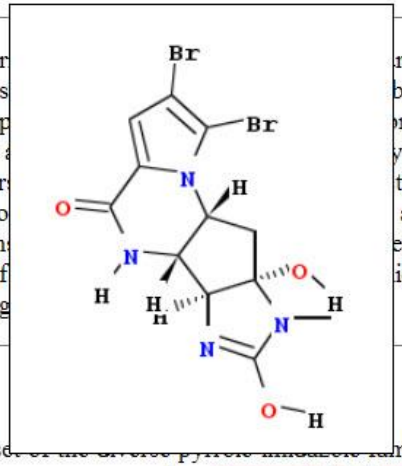
Mohammad Movassaghi *, Dustin S. Siegel and Sunkyu Han

Massachusetts Institute of Technology, Department of Chemistry, 77 Massachusetts Avenue 18-292, Cambridge, MA 02139-4307, USA. E-mail: movassag@mit.edu

Received 2nd July 2010 , Accepted 20th July 2010

First published on the web 16th August 2010

The pyrrole-imidazole family of marine alkaloids, der rsors, constitutes a diverse array of structurally complex natural products bers of this family that possess a tetracyclic molecular framework incorp nnectivities. We provide a hypothesis for the formation of the unique y exploits the intrinsic chemistry of plausible biosynthetic precursor tive total syntheses of all known agelastatin alkaloids including the first to and F. Our gram-scale chemical synthesis of agelastatin A was in enesis of the cyclopentane C-ring and required the development of idazolone-forming annulation reaction and a carbohydroxylative trapping

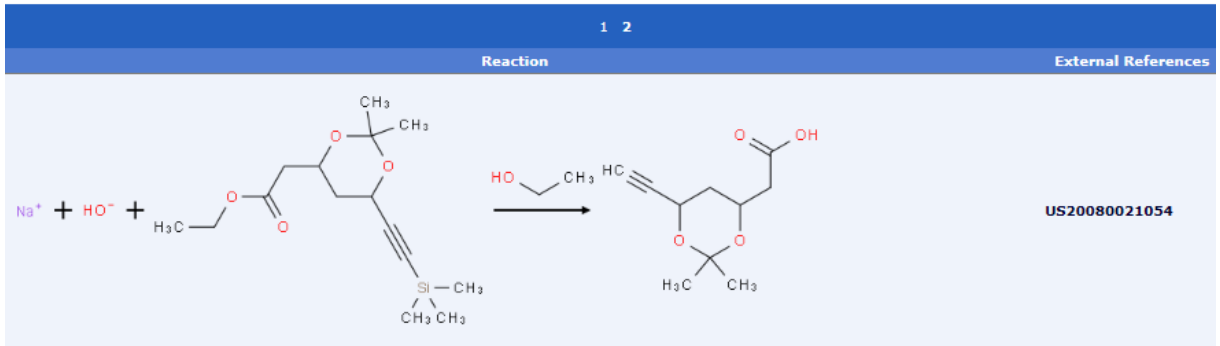


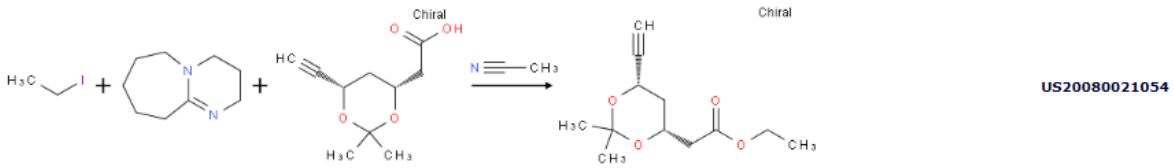
Introduction

The agelastatin alkaloids constitute an intriguing subse mily of marine alkaloids that are likely derived from linear biogenetic precursors such as clathrodin (**7**),[1] hymenidin (**8**),[2] and oroidin (**9**, Fig. 1).[3,4] (−)-Agelastatins A (**1**) and B (**2**) were first isolated from the Coral Sea sponge *Agelas dendromorpha* by Pietra *et al*. in 1993 who successfully identified and chemically studied their unique

Hosting Reactions

- Seed set of over 1 million reactions from patents to develop validation and standardization routines.
- Reactions to be extracted from RSC journal articles, ESI and reaction databases will be examined
- Resulting validation algorithms used at deposition

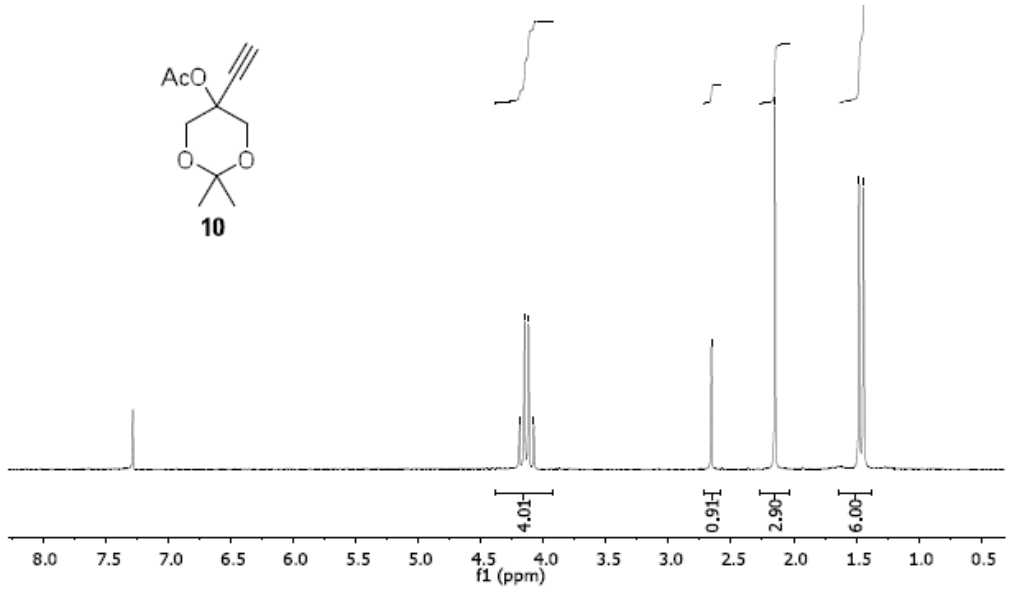


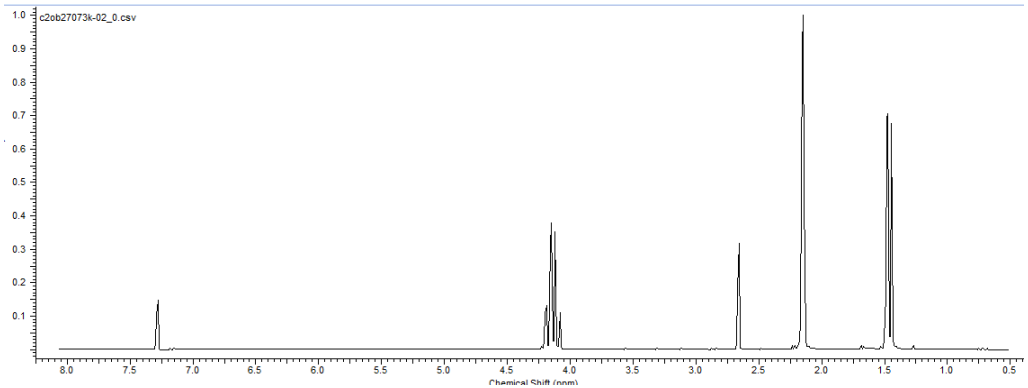


The challenges of analytical data

- Integration of ChemSpider to analytical instrumentation vendors already in place
  - Agilent, Bruker, Thermo, Waters
- Vendors produce complex proprietary data formats and standard formats are required (JCAMP, NetCDF, AniML)
  - ChemSpider already hosts thousands of JCAMP spectra
  - Support of “assigned spectra” in place
  - Data validation approaches understood
  - There are a myriad of analytical data types…

Turning “Figures” Into Data



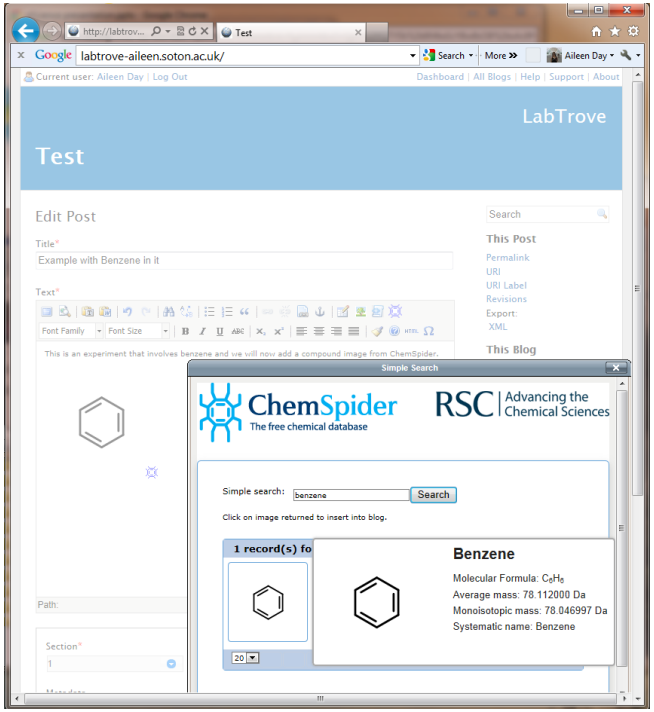


Community Data Repository

- Automated depositions of data – service-based deposition, sweep and deposit
- Integrate to Electronic Lab Notebooks as feeds
- High value would be databases of reference data, but validated by model validation and the community
- National services feeding the repository – crystallography, mass spectrometry

E-Lab Notebooks

- Integration between ELNs and:
  - ChemSpider
  - ChemSpider Reactions
  - Chemistry Data Repository



What do we have in place?

- We are **testing** a data repository on our assets – ChemSpider and our archive of publications
- Working with many collaborators to define needs
- Deposition system for deposition of chemical compounds – hosts >29 million chemicals
- Crowdsourcing curation & annotation platform
- Chemical validation & standardization platform
- Chemical reactions database with >1 million reactions and presently developing RVSP
- Analytical data handling formats (JCAMP preferred)
- And lots in development…

The Challenges Ahead

- Chemistry is NOT just nicely defined structures!
  - Materials, minerals, attached to beads, polymers, ambiguous materials
- Domain-specific measurements
  - File format standards are limited in application
- Encouraging scientists to free up their data
  - AltMetrics, open data mandates, systems
- The data explosion continues
  - 4 years ahead to expand capability

The Future

Individual Scientists

Internet Data

Published Chemistry

RSC | Advancing the Chemical Sciences

RSC Data

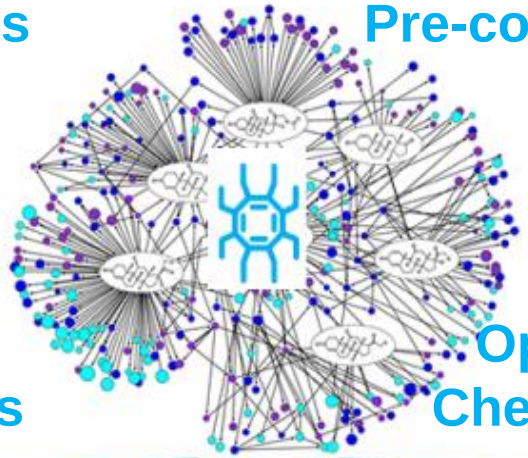
Electronic Lab Notebooks

Enterprise Laboratory Platform

lims eln

Small organic molecules
Undefined materials
Organometallics
Nanomaterials
Polymers
Minerals
Particle bound
Links to Biologicals

Commercial Software
Pre-competitive Data
Open Science
Open Data
Publishers
Educators
Open Databases
Chemical Vendors



Aggregate Data

Search Network

Generate Models

Integrate and Federate

Thank you

Email: williamsa@rsc.org
Twitter: @ChemConnector
Personal Blog: www.chemconnector.com
SLIDES: www.slideshare.net/AntonyWilliams