Maintaining a Sustainable Scholarly Record

CINF Perspective

Judith N. Currano

Head, Chemistry Library, University of Pennsylvania
Chair, Division of Chemical Information
248th ACS National Meeting, San Francisco, CA
What IS Chemical Information, Anyway?

THE LITERATURE
What IS Chemical Information, Anyway?
Sustainability Issues in Chemical Information

• **The scholarly record must exist**
  – Relevant results must be published in a timely manner and in a location that makes sense
  – Results must be reported accurately according to code of ethics
  – Provenance of data and results must be provided
  – **Key players:** researchers and primary publishers

• **Researchers must be able to access relevant results in a timely and efficient manner**
  – Relevant published results must be easily filtered from background noise
  – Information must be affordable to those who require it
  – Internal data and results must be organized and accessible
  – **Key players:** secondary publishers and vendors, informatics specialists, and librarians

• **The scholarly record must endure**
  – Digital information is much more challenging to preserve for the long-term than paper-based information
  – **Key players:** librarians, primary publishers, and third-party archiving services
**Whence cometh these data?**

- Considerations when taking data
  - Ensure reproducibility
  - Provide documentation/annotation/metadata for the purpose of validity
  - Ideally, have some sort of data validation step

- Considerations when publishing data
  - Annotation for the purpose of discoverability and reuse
  - Context in which data were taken
  - Carefully attribute “once removed” data
Chemical Information Providers to the Rescue!

- Electronic laboratory notebooks can “de-silo” the research process
- Development of internal data warehouses and repositories
- Initiatives in electronic publishing
  - “Pre-pagination”: ASAP, Advance Articles, E-view, etc.
  - “Pre-editing”: As Soon As Accepted
  - Prepublication: Preprints/E-prints (in certain areas of chemistry)
Speed of Research Creates a New Sustainability Problem

Chemical Abstracts issue from 1907. Organized by subject and constructed to be browsed as periodical.

Abstracts Added to Chemical Abstracts by Year and Type, 1907-1945
Speed of Research Creates a New Sustainability Problem

Abstracts Added to *Chemical Abstracts* by Year and Type, 1945-2013
The available literature is increasing exponentially.

Total Abstracts in *Chemical Abstracts* by Year

- Total Number of Abstracts in the Database, 1945: 1,386,786
- Number of Abstracts ADDED to the Database in 2013: 1,497,773

The exponential growth can be described by the equation:

\[ y = 110009e^{0.0569x} \]

with a coefficient of determination (R²) of 0.9578.
So Is the Number of Published Substances

Total Small Molecules Appearing in the CAS REGISTRY by Year

TOTAL Number of Small Molecules in the CAS REGISTRY in 1988: 9,349,316

Small Molecules Registered Per Year

Number of Small Molecules ADDED to the CAS REGISTRY in 2009: 9,817,177
However, Two Things AREN’T Increasing Exponentially

(also €, £, ¥, etc.)

Reproduced from https://staticfree.info/projects/24h_clock/
(accessed July 15, 2014)
How do we read in the “digital age?”

The synthesis of tamandarin M was the synthesis of dehydrotamandarin B. Removal of the coupling with the five-carbon side chain 56 using B0.

Synthetic Studies of Tamandarin B Side Chain Analogues

Kenneth M. Lassen, Jisun Lee, and Madeleine M. Joullie*
Department of Chemistry, University of Pennsylvania, 231 South 34th Street, Philadelphia, Pennsylvania 19104
mjoullie@upenn.edu
Received February 19, 2010

The syntheses of three tamandarin B analogues are described. The goal of these studies was to prepare material to determine their relative therapeutic index and to gain an oversight as to their potential for clinical applications.

Journal of Organic Chemistry screen shots are used with permission from ACS Publications and are copyright © 2014, American Chemical Society (ACS). All rights reserved. Natural Product Reports screen shots are used with permission from the Royal Society of Chemistry.
What Is Out There?

• Introduction of user-friendly, electronically-searchable databases aimed at the end-user chemist

=> FILE REG
   => S C8H9BrO
      Connected to File REGISTRY
      253 Answers stored

=> FILE CA
   => S L1 AND BIOL*
      Connected to File CA
      269 Answers stored

=> D L2 1-13

SciFinder screen shots are used with permission from CAS and are copyright © 2014, American Chemical Society (ACS). All rights reserved.
What is out there?

- First, there was computerized structure searching via line notation. Then came GUI interfaces!

Chemist’s Structural Diagram

Wiswesser Line Notation

[Image of a chemical structure diagram with labels:

1 = saturated 1-carbon chain
V = C=O]

MarvinSketch by ChemAxon
An example of a connection table, used to describe acetone. Each line in the table represents a connection between two nodes.
A substructure search generates a connection table for the substructure of interest and searches for it within the connection table of each substance in the database.
A Markush structure is a generic, patented structure that serves to protect a broader group or class of substances. It is named after Eugene Markush, the first scientist to successfully include such a structure in a patent.

**Substructure of Interest**

**Traditional Substructure Search**

What Is IMPORTANT?

YOU HAVE 4,325,894 HITS

(And that’s just from the FIRST database you tried....)

Social Networking Tools

The Invisible College digitized!

Comprehensives, Compilations, and Expert Opinions

Related References

Scholarship’s “More Like This”
The Chemical Literature Keeps Growing…

... but much of it is still useful to researchers.
Preserving Digital Information: Challenges

- **Persistence**: Data must remain accessible/findable
  - Persistent URLs, locations, etc.
  - Third party archiving services
    - LOCKSS, CLOCKSS, and Portico for journal articles
- **Data Integrity**: Data must remain uncorrupted
  - Need periodic checks for “bit rot”
  - Multiple copies guard against single-site corruption
- **Data Usability**: Data must be able to be used at a later date
  - Appropriate annotation/metadata needed to describe data format
  - Constant upconversion needed to ensure that current software is able to read data

*Standards are of key importance, but are lacking....*
Technological Challenges to Preserving Digital Content

http://www.vintage-computer.com/ibm_pc.shtml
What’s a Poor Chemist To Do?

- Publish when you have something important to say
- Publish in places whose policies are consistent with sustainability
- Ensure that your data are maintained appropriately, both in your lab and where you choose to archive
- Document everything!
- When adopting a new technology, think critically about issues of persistence and sustainability
- Remember: **YOU WANT PEOPLE USING YOUR SCIENCE 100 YEARS (and more!) FROM NOW!**
Acknowledgements

• University of Pennsylvania Libraries
• ACS Division of Chemical Information and Philadelphia Local Section
• Chemical Abstracts Service
  – Roger Schenck
• Members of the ACS Division of Chemical Information and readers of CHMINF-L
  – Erin Bolstad
  – Jeremy Garritano
  – Leah McEwen
  – Andrea Twiss-Brooks
  – William Town
Acknowledgements

• SciFinder screen shots are used with permission from CAS and are copyright © 2014, American Chemical Society (ACS). All rights reserved.

• *Journal of Organic Chemistry* screen shots are used with permission from ACS Publications and are copyright © 2014, American Chemical Society (ACS). All rights reserved.

• *Natural Product Reports* and *Specialist Periodical Reports* screen shots are used with permission from the Royal Society of Chemistry

• MarvinSketch screen shot is used with permission from ChemAxon