

The Shift in the Currency of Scholarly Information: Implications for Library Collections and Acquisitions

Hilary M. Davis and John N. Vickery
North Carolina State University Libraries

15th Annual North Carolina Serials Conference
Chapel Hill, NC - March 30-31, 2006

Introduction

The unit of information currency is shifting from the journal article to the dataset.

- Primary method/medium by which knowledge is exchanged
- GIS, Genomic, Chemical, Census, Market Reports, Astrophysical

Outline

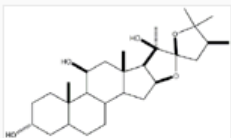
- Indicators
- Stakeholders
- Implications for libraries
- Reflection & Scenarios
- What next?

Indicators

■ Data as commodity

Compound Data Index

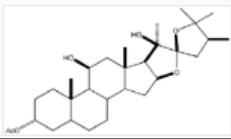
From the following article
[Functional characterization of IREs by an inhibitor of the RNA helicase eIF4A](#)
Marie-Eve Bordeleau, Ayaka Mori, Monika Oberer, Lisa Lindqvist, Louisa S Chard, Tatsuo Higa, Graham
Nature Chemical Biology 2, 213-220 (2006)
doi:10.1038/nchembio776



Compound 1
Hippuristanol

[View in PubChem](#)

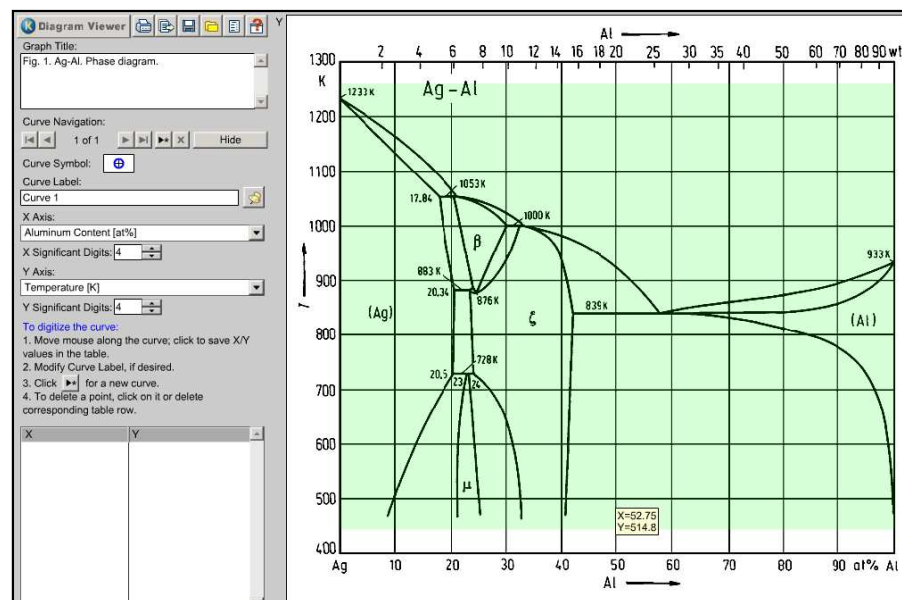
[Full size image \(4 KB\)](#) | [Download ChemDraw file of structure \(3 KB\)](#)



Compound 2
Hippuristanol 3-acetate

[View in PubChem](#)

[Full size image \(4 KB\)](#) | [Download ChemDraw file of structure \(4 KB\)](#)



Indicators

- Data as commodity
- Legislation to protect datasets/databases

H.R. 3261: Database and Collections of Information
Misappropriation Act

Indicators

- Data as commodity
- Legislation to protect datasets/databases

H.R. 3261: Database and Collections of Information
Misappropriation Act

- Growth and manipulability of data

In 2002, 5 exabytes of new information was created
(that's 37,000 times the size of the Library of Congress)

Indicators

- Data as commodity

- Legislation to protect datasets/databases

H.R. 3261: Database and Collections of Information Misappropriation Act

- Growth and manipulability of data

In 2002, 5 exabytes of new information was created (that's 37,000 times the size of the Library of Congress)

- Publication requirements



Indicators

- Data as commodity
- Legislation to protect datasets/databases

H.R. 3261: Database and Collections of Information Misappropriation Act

- Growth and manipulability of data

In 2002, 5 exabytes of new information was created
(that's 37,000 times the size of the Library of Congress)

- Publication requirements
- Public/Private partnerships



What the stakeholders are saying...

Most of the discussion is among researchers and scientists

"Celera's treatment of their human genome sequence data shows the sequence's nature as a economic good as well as the possibility of it being a private good or commodity. The company can control and sell access to the data, and outside researchers have no recourse if they are denied or cannot afford access."
--William Bell, 2003

What the stakeholders are saying...

Most of the discussion is among researchers and scientists

“DNA sequences thus obtained by the database user are still not copyrightable. However, only by copying a copyright-protected music file from the database could the external user obtain the DNA sequences. Such duplication may be copyright infringement, and this encoding approach therefore may provide a form of IP protection.”

--Stemmer, 2002

What the stakeholders are saying...

Most of the discussion is among researchers and scientists

"For most scientists, having the right to download data does not mean much if the extracted information cannot later be reutilized and republished."

--Maurer, Hugenholtz, Onsrud, 2001

What the stakeholders are saying...

Most of the discussion is among researchers and scientists

- Revolutionizing science and engineering through cyberinfrastructure (Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure 2003)
- Scientific data and information (Report of the Committee on Scientific Planning and Review Assessment Panel 2004)
- Draft Report of the American Council of Learned Societies' Commission on Cyberinfrastructure for Humanities and Social Sciences (American Council of Learned Societies 2005)
- Long-lived data collections: enabling research and education in the 21st century (Report of the National Science Board 2005)

Library-land discussions...

Discussion in the LIS literature is dominated by the big thinkers and open data advocates



"the practice of scholarship in many, many fields now is shifting and is now integrally involved in computation, in simulation, in large scale data collection and observational data sets, the control of very sophisticated data acquisitions systems. That data software, visualization and observation now are every bit as important and as significant as traditional monographs and journal articles."

--Clifford Lynch, 2005

Library-land discussions...

Discussion in the LIS literature is dominated by the big thinkers and open data advocates



"... the boundaries of what would conventionally be thought of as a 'paper' or 'article' can be scaled both up and down...an article could be disassembled down to an individual marked-up component such as one atom in a molecule, or instead aggregated into a journal, collection of journals, or ultimately into the semantic web!"

--Rzepa, Murray-Rust, 2001

Library-land discussions...

Discussion in the LIS literature is dominated by the big thinkers and open data advocates

"...a future unit of communication should not discriminate between media types and should recognize the compound nature of what is being communicated."

-- Herbert Van de Sompel, et al., 2004

Implications

- Lots of data, highly scattered
- Value-added integrating services
- Laws to restrict use of data

- Current business models
- Budgeting
- Selection & Evaluation
- Negotiation
- Licensing

Current business models

- Institutional membership (e.g., ICPSR)
 - Library pays for access for entire institution
- Sponsor creation of data (e.g., State of NC GIS)
- Serials-like continuations (e.g., NIST)
- One-time payment per dataset (e.g., Market Reports)
- Ad-hoc arrangements (e.g., GIS)
 - Handshake, case-by-case basis
- Lack of model / Transitional (e.g., NCCS)
 - No business models for dealing with academic needs

Hindsight

- Print to electronic journals
- What did we expect?
 - Lower costs
 - More flexibility
- What did we get?
 - Higher costs
 - Complicated licensing arrangements

The big question:

Who is going to control the datasets market?

Budgeting

- Allocating money for datasets
 - Not a traditional part of library budgets
 - Weigh costs and benefits
 - How much do we allocate?
- Prepare for hidden costs
 - Interfaces for datasets
 - Data maintenance
 - Additional hardware and software requirements

Selection

- What do we collect?
 - Request driven
 - Data producer driven
 - Proactive selection
- Approval plans?
 - “Yankee Datasets Peddler”
- Capture data at the source
 - Institutional repositories

Evaluation

- Quality of data
 - Peer-reviewed
 - "... peer-review would only serve to complicate and slow down the scientific process. Plus peer-review is not thought to be necessary by the scientists for this type of data because it is generated by highly specific techniques that are not subjected to further interpretation."
-- Brown, 2003
- Quantity of data
- Interface
- Format

Negotiation

- Better positioning
 - Be aggressive negotiators
 - Know value of datasets
 - Leverage local expertise
- Data in a user-*unfriendly* format?
 - Bargaining-power
 - Hosting data

Licensing

- Is data copyrightable/infringeable?
 - Keep an eye out for bills similar to H.R. 3261
 - Contract law vs. copyright
- Give and take of rights
- Implications for libraries
 - ILL
 - Printing
 - Downloading
 - Re-use or misuse of data

Potential scenarios

- Pay for articles + datasets
 - Publishers provide links to datasets
 - Publishers license access to datasets
- Bundling of datasets
- Archival access rights to datasets
- Libraries liable for forbidden re-use and misuse
- Libraries take over datasets market

Questions to take home

- How far will this shift go?
- How will library users and libraries be able to use data for ILL, copying, downloading, further manipulation, republishing?
- How can we serve as scouts, instructors and advocates of datasets on behalf of our users?
- How do we weigh specific vs. broad-appeal datasets for our users?
- What should our local infrastructure look like to support our users' datasets needs?
- How do we determine quality of datasets and account for costs beyond the content?

Questions?

Hilary M. Davis
hilary_davis@ncsu.edu

John N. Vickery
john_vickery@ncsu.edu