Big Data and Info Pros: Let’s Own the Conversation!
Data collection everywhere!

I left headphones in hotel

“Report it at iLeftMyStuff.com”

What a gold mine of info!
REPORT A LOST ITEM

Tell us about your item(s)

* Date Lost: [Select Date]

* Location the Item was lost:
(Please specify Room # or other location)

* Type of Item
-- Select an Item --

Description:

Color:

Size:

Brand:

Additional Information:
(Other Identifying Features)

Would you like to add a photo?: [Add Photo]

Hotel Location

Hyatt Regency Crystal City
2799 Jefferson Davis Highway
Arlington, VA 22202

Contact Details

* Your Name:

* Your Email:

* Your Phone Number: [ ] ext [ ]
Where’s our role?

We’re not looking at that type of personal info

We can ID, compare, evaluate data sources
Evolution of information

- Print indexes, catalogs
- Bibliographic databases, Boolean logic
- Full-text databases
- Text and data mining of “semantic triples” (a.k.a. info bits)
TDM = extraction of info bits

Structured info forms a **semantic triple**

Subject — predicate — object

“the sky is blue” = “sky — has_the_color — blue”

```
genid:ARP40722 | http://abc.com/color | hex:#0000FF
```
TDM = extraction of info bits

Semantic triples from Wikipedia article:
  shingles — is_also_called — herpes_zoster
  shingles — is_caused_by — varicella_zoster_virus
  varicella_zoster_virus — is_treated_with — acyclovir
  immunosupression — is_risk_factor_for — shingles
TDM = extraction of info bits

Semantic triples from a bib cite:

Article_X — has_author — Doe, John
Article_X — published_in — Heredity
Doe, John — has_affiliation — Drexel_University
Article_X — funded_by — grant_123
Article_X — has_subject — Alzheimer’s_Disease
Google Books Ngram Viewer

Millions of books digitized
Each word & sentence parsed

*The school nurse treated the boy.*

Adjective “school” modifies noun “nurse”
Subject “nurse” conducting “treated”
action on object “boy”
Google Books Ngram Viewer

Can search for ANY adjective modifying “librarian”
Strange syntax: book_INF, _START_ women...

books.google.com/ngrams
Linked open data

Linked data enables meaningful connections across content
Normalized data = enhanced discovery beyond full-text

Full-text content + linked OPEN data + APIs = WOW!
Big data use cases

Springer Nature Journal Suggester
You provide manuscript title & abstract; it recommends where to submit MS

journalsuggester.springer.com
Enter your manuscript details to see a list of journals most suitable for your research.

**Manuscript title**
A Meta-analysis of the Impact of Aspirin, Clopidogrel, and Dual Antiplatelet Therapy on Bleeding Complications in No

**Manuscript text**

OBJECTIVE:
The aim of this study was to determine the bleeding risks associated with single (aspirin) and dual (aspirin + clopido versus placebo or no treatment in adults undergoing noncardiac surgery.

SUMMARY OF BACKGROUND DATA:
The impact of antiplatelet therapy on bleeding during noncardiac surgery remains controversial. A meta-analysis was associated with single and DAPT.

METHODS:
A systematic review of antiplatelet therapy, noncardiac surgery, and perioperative bleeding was performed. Peer-review from relevant societies were queried. Studies without a control group, or those that only examined patients with coro
<table>
<thead>
<tr>
<th>Journal Name</th>
<th>Impact Factor</th>
<th>First Decision (Average)</th>
<th>Acceptance Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cardiovascular Drugs and Therapy</td>
<td>2.82</td>
<td>27 days</td>
<td>22%</td>
</tr>
<tr>
<td>Current Atherosclerosis Reports</td>
<td>2.731</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Cardiology and Therapy</td>
<td>-</td>
<td>17 days</td>
<td>83%</td>
</tr>
<tr>
<td>American Journal of Cardiovascular Drugs</td>
<td>2.768</td>
<td>27 days</td>
<td>57%</td>
</tr>
</tbody>
</table>
Big data use cases

Help searchers gauge article impact
API that looks up article DOI, calculates # of citations to that article

Build an internal open-access image repository
Monitor open access journals for specific type of image
Big data use cases

ID high-quality vs. predatory conferences

Chart # of institutions represented by speakers, citation & reference metrics of speakers
Big data use cases

Ensure more comprehensive searches

API that looks up search terms for MeSH equivalent, appends all terms w/in that concept

“opioid dependence” → Opioid-Related Disorders, Heroin Dependence, Morphine Dependence
Big data use cases

Create dashboard for business intel
Monitor key publications
What institutions are publishing research
Who are the most cited researchers at an institution
What institutions are receiving grants?
Linked open data examples

Springer Nature SciGraph
View visual patterns in large datasets
See relationships across disciplines, formats

springernature.com/gp/researchers/scigraph

BatesInfo.com

20
You searched for 'alzheimers' 2577 matches note: max 50 results per type shown

<table>
<thead>
<tr>
<th>Type</th>
<th>URI</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Article (1282)</td>
<td>articles:0abd2c094e78f5b5402e748431607020</td>
<td>Article: Alzheimers Lebenswerk</td>
</tr>
<tr>
<td>CategorizationEvent (752)</td>
<td>articles:7fa3852f94bda986352eb138436648d1</td>
<td>Article: Evaluating the Predictive Power of Multivariate Tensor-based Progression via Convex Fused Sparse Group Lasso.</td>
</tr>
<tr>
<td>Grant (468)</td>
<td>articles:18795ac13d63679b2c262ddee3dc7839</td>
<td>Article: Benzodiazepines in Alzheimers: increased risk of pneumonia</td>
</tr>
<tr>
<td>BookChapter (47)</td>
<td>articles:480df95136bea730d61ad28841d1ee4</td>
<td>Article: Fourth International Conference on Alzheimer's Disease and Related Disorders</td>
</tr>
<tr>
<td>Affiliation (19)</td>
<td>articles:35f053e9992412b714ab611b5a9ce0c8</td>
<td>Article: Inhibition of cytosolic human forebrain choline acetyltransferase phosphomonoester that accumulates during early stages of Alzheimers disease</td>
</tr>
<tr>
<td>Subject (4)</td>
<td>articles:4d8aebecc012e16769c9a157ee407f6b3</td>
<td>Article: Alzheimers disease is not associated with the hypertension gene beta3, either independently or interactively with apolipoprotein E.</td>
</tr>
<tr>
<td>BookEdition (3)</td>
<td>articles:a2e62328e07b551429b54f86ef3818fe</td>
<td>Article: Lack of genetic association of cholesteryl ester transfer protein (CETP) and Alzheimers disease.</td>
</tr>
<tr>
<td>BookSeries (1)</td>
<td>articles:8a35c49d645c5a422829e118139d9447</td>
<td>Article: Alzheimer's disease</td>
</tr>
</tbody>
</table>
| Nonprofit (1)       | articles:7b710ee922987bd87f72730240356a33 | Article: Amyloid Beta Peptide Processing, Insulin Degrading Enzyme, Alzheimer's disease. }
Article: Adherence to a Mediterranean diet and Alzheimer’s disease risk in an Australian population
Linked open data examples

PubChem open chemistry database

Use API to expand search for all names for a substance

See links from PubChem to SpringerNature articles
Aspirin

**PubChem CID:** 2244

**Chemical Names:** Aspirin; ACETYSALICYLIC ACID; 50-78-2; 2-Acetoxybenzoic acid; 2-(Acetyloxy)benzoic acid; Acetylsalicylate

**Molecular Formula:** C₉H₈O₄ or CH₃COOC₆H₄COOH or HC₆H₇O₄

**Molecular Weight:** 180.159 g/mol

**InChI Key:** BSYNRYMUTXBSQ-UHFFFAOYSA-N

**Drug Information:** Drug Indication, Therapeutic Uses, Clinical Trials, FDA Orange Book, FDA UNII

**Safety Summary:** Laboratory Chemical Safety Summary (LCSS)

Aspirin is the prototypical analgesic used in the treatment of mild to moderate pain. It has anti-inflammatory and antipyretic properties and acts as a cyclooxygenase which results in the inhibition of the biosynthesis of prostaglandins. Aspirin also inhibits platelet aggregation and is used in the prevention of blood clots.
### 13.6 Springer Nature References

<table>
<thead>
<tr>
<th>Title</th>
<th>Journal or Book</th>
<th>Year</th>
<th>PMID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dissolution and pharmacokinetics of a novel micronized aspirin formulation</td>
<td>Inflammopharmacology</td>
<td>2011</td>
<td>22057729</td>
</tr>
<tr>
<td>Aspirin Desensitization and High-Dose Aspirin Therapy in Aspirin-Exacerbated Respiratory Disease</td>
<td>Practical Medical and Surgical Management of Chronic Rhinosinusitis</td>
<td>2015</td>
<td></td>
</tr>
<tr>
<td>Non-compliance is the predominant cause of aspirin resistance in chronic coronary arterial disease patients</td>
<td>Journal of Translational Medicine</td>
<td>2008</td>
<td>18759978</td>
</tr>
<tr>
<td>Actions and interactions of acetylsalicylic acid, salicylic acid and diflunisal on platelet aggregation</td>
<td>European Journal of Clinical Pharmacology</td>
<td>1984</td>
<td></td>
</tr>
<tr>
<td>Rapid Aspirin Challenge in Patients with Aspirin Allergy and Acute Coronary Syndromes</td>
<td>Current Allergy and Asthma Reports</td>
<td>2016</td>
<td>26758864</td>
</tr>
</tbody>
</table>
Linked open data resources

Library of Congress Linked Data Service
Datasets of subject headings, names, etc.
(id.loc.gov)

DBpedia.org
Structured data extracted from Wikipedia
Info pros’ role in big data

We can think creatively about info
We understand uses of structured data
We know what metadata is needed
We care about quality of taxonomies
We understand our clients’ info needs, search behavior
We can evaluate, curate datasets

Develop quality, cost checklist

ID what’s missing from open access

Monitor govt agencies, open-data initiatives
We can evaluate, curate datasets

Monitor dataDryad.org – repository of research data

Use re3data.org (Registry of Research Data Repositories) to ID sources to monitor

Use Google Dataset Search (toolbox.google.com/datasetsearch)
Our challenge?

IDing where TDM would help “backstage”
Inconsistently indexed content (Motor Neuron Disease is “Lou Gehrig’s disease” in US)
Enhancing content with metadata (consistent geo coding, image tagging)
Conduct strategic reference interviews

Clients only ask for what they think we can get

We understand the client’s use case

We ask “What’s essential? What’s nice to have?”

We think creatively about finding answers
Example

Univ wants to calculate value of college degree to show ROI
What data sets do they need?
Starting & mid-range salaries, by profession
Cost of living data
Student loan payment data
[...]

Source: The Accidental Data Scientist (2015)
Example

B2C company notices product sales down
Info pro builds API to monitor news, IDs
dataset of weather, imports to Tableau,
creates dashboard enabling user to
understand trends
More big data roles

We understand copyright issues
Greater discoverability = greater demand for content
Ensure license compliance, avoid lockouts

We leverage information for greater ROI
We know how to promote resources internally
Higher ROI for online content subscriptions
More big data roles

We collaborate with other groups
  Big data projects involve teams from collections development, cataloging/metadata, IT, outreach

We see the bigger picture
  We think beyond the tool or resource to delivering insight
Lead the big data discussion

Create internal libguide re: big data
Licensed sources & licensing restrictions
Internal subject experts to consult
Best practices
TDM tools & tutorials
Try googling: inurl:libguides (“data mining” OR “big data”)
Build TDM checklist (& info center’s role)

✓ ID questions being addressed
✓ ID data elements, functionality needed
✓ ID best resources for the budget & project
✓ Retrieve, extract & enhance the data
✓ Build tools to query the data
✓ Analyze the results
✓ Archive/curate the data
Lead the big data discussion

Create a big-data sandbox
Build relationships with publishers, your legal dept
ASK clients what they’re working on
Engage! Our clients need our help
Questions?