Research Intelligence

Text & Data Mining at Elsevier

Dr Nick Fowler
Managing Director, Research Networks

Elsevier has a unique vantage point on the world of research

Each year

• 1.4 million article manuscripts received by ~2,000 journals (all offer Open Access options)
• 400,000 new articles published, in addition to 11M existing articles
• 2,000 new books published

• ScienceDirect: 800M digital article downloads
• Scopus: 55M records, 21,900 titles, 5,000 publishers, 700M citations
• SciVal: 75 trillion metrics values
• Pure: current research information system: >200,000 researchers supported
• Mendeley: 5M users globally
• Grants: 7,000 sponsors, 20,000+ active opportunities, ~5M awarded grants
• Patents: >93m records, 100 patent offices

Primary publishing

Derived and aggregated data
Elsevier is increasingly the partner of choice for global research leaders

National research assessment and benchmarking reports

- HEFCE: sole citation provider for REF 2014
- US NSF Science & Engineering Indicators
- ERA (Australia)
- FCT (Portugal)
- VQR (Italy)

Global University Rankings

- Times Higher World University Rankings
- QS rankings
- Maclean’s
- US News rankings (Arab Region)

EU partnerships / initiatives – selected examples

- FP projects (e.g. U-Multirank, THOR)
- Stem Cell Research (with DG Research)
- Research Collaboration & Mobility (with Science Europe)
- Brain Science (with DG Research & DG Connect)
- Ex-post evaluation of FP7 (with DG Research)
- Brexit Resource Center
- JRC collaboration on TIM
- Gender in Global Research Report (with JRC)

Brexit Resource Centre

Free access to research, data and analysis
Elsevier facilitates TDM
Text mining Elsevier published content

Researchers at academic institutions can text mine whatever they have access to via their institution at no additional cost for non-commercial research purposes.

Elsevier

- Converts all our content into text mining friendly formats (XML)
- Invests in improving the user experience for miners
  - Dedicated developers portal
- Makes it easier to mine across publishers
  - Collaborate with CrossRef’s Text and Data Mining API
- Works with the community to help develop text mining tools
  - OA corpus

http://www.elsevier.com/tdm
...and we collaborate with researchers and institutes

Funding TDM research

- Funding postdocs at the University of Bologna
- Funding interns at the University of Melbourne
- Joint research grant with the University of Manchester
- Funding an NLP Centre of Excellence at Indian Institute of Technology, Bombay

Partnerships

- Established partnerships with Humboldt University to better understand and address TDM challenges

Conferences & Events


http://www.elsevier.com/tdm
TDM with bibliometric data
What is Scopus Data?

Scopus data is the highly structured content that is searchable through Scopus.com

**JOURNALS**

- 21,568 peer-reviewed journals
- 361 trade journals
- Full metadata, abstracts and cited references (ref’s post-1995 only)
- Funding data from acknowledgements
- Citations back to 1970

**CONFERENCES**

- 90K conference events
- 7.3M conference papers
- Mainly Engineering and Computer Sciences

**BOOKS**

- 531 book series
- 30K Volumes / 1.2M items
- 119,882 stand-alone books
- 974K items
- Focus on Social Sciences and A&H

**PATENTS**

- 27M patents
- From 5 major patent offices
  - WIPO
  - EPO
  - USPTO
  - JPO
  - UK IPO

That has been enhanced with

- **Deep citation linking** for all articles 1970-present
- Authoritative **Author & Affiliation Profiles** for all records 1823-forward
- Additional **enhanced metadata**, ex. Medline & other index terms, Funding Acknowledgements, etc.
European research collaboration follows geopolitical, historical & linguistic ties

Key:
- Thickness of line & proximity of countries together shows collaboration strength
- Highest impact countries nearer the centre of the map
TDM with Patent Data
European comparator countries’ research is well-cited in the global patent literature

Figure 7.4 — Share of global patent applications for UK and comparators, 2007-2011.
Figure 7.7 — Share of 2007-11 patent citations to articles published 2007-11 for the UK and comparators. Each data point corresponds to journal articles published in the year shown and citations to these articles from patent applications and granted patents in the period 2007-2011.
TDM with Funders’ Data
Brain Research: strong alignment on disorders research between FP7 & NIH projects

### Table 3.10

Top 10 concepts that occurred in brain and neuroscience research articles relating to disorders from document sets A, B, and C, based on the sum of term frequency-inverse document frequency (tf-idf) of the concept in the document set that it belonged to. Figures in parentheses are the frequency with which the concept occurred in the document set. Highlighted in violet are concepts that appeared in the top 10 disorder-related concepts in all three document sets, reflecting common areas of focus. Highlighted in magenta are concepts that only appeared in Set A and Set B. Concepts that are not highlighted were those unique to each document set, indicating different areas of focus in disorder-related concepts in brain and neuroscience research.

<table>
<thead>
<tr>
<th>Set A</th>
<th>Set B</th>
<th>Set C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brain and neuroscience articles from Scopus</td>
<td>Brain and neuroscience funded grants awarded from NIH RePORTER</td>
<td>Brain research project synopses supported by the European Commission</td>
</tr>
<tr>
<td><strong>Stroke (21,404)</strong></td>
<td><strong>Alzheimer Disease (1,366)</strong></td>
<td><strong>Stroke (6)</strong></td>
</tr>
<tr>
<td>Depression (21,668)</td>
<td>Stroke (1,044)</td>
<td>Parkinson Disease (7)</td>
</tr>
<tr>
<td>Neoplasms (25,047)</td>
<td>Schizophrenia (1,219)</td>
<td>Schizophrenia (5)</td>
</tr>
<tr>
<td>Alzheimer Disease (14,522)</td>
<td>Pain (817)</td>
<td>Memory Disorders (3)</td>
</tr>
<tr>
<td>Pain (16,719)</td>
<td>Parkinson Disease (869)</td>
<td>Vision Disorders (2)</td>
</tr>
<tr>
<td>Schizophrenia (13,752)</td>
<td>Depression (910)</td>
<td>Alzheimer Disease (4)</td>
</tr>
<tr>
<td>Parkinson Disease (11,366)</td>
<td>Neoplasms (829)</td>
<td>Myasthenia Gravis (1)</td>
</tr>
<tr>
<td>Wounds and Injuries (13,414)</td>
<td>Glioma (526)</td>
<td>Hearing Loss (3)</td>
</tr>
<tr>
<td>Syndrome (13,258)</td>
<td>Pervasive Child Development Disorders (626)</td>
<td>Alkalosis (1)</td>
</tr>
<tr>
<td>Multiple Sclerosis (9,275)</td>
<td>Bipolar Disorder (499)</td>
<td>Pain (1)</td>
</tr>
</tbody>
</table>
Earlier this year, DG Research* used SciVal for bibliometric analyses for the ex-post evaluation of

Graph 12: Share of the priorities’ publications in top 1% and top 5% highly cited publications (2007-2015)
Source: SciVal based on Corda-Sesam-Respir

Source: ex-post evaluation of FP7 (webpage)
The Seventh Framework Program shows high impact throughout all disciplines

Source: SciVal (Scopus data up to 19/10/16); FP7 publications; time window: 2008-2015

Average: 2.55
TDM with Social Media
How are we able to report on the gender of authors?

Scopus®

genderize.io

NamSor

SOCIAL MEDIA

Sociolinguistics
Gender in Research: Proportion of women and men authors per comparator and period

[Bar chart showing the proportion of men and women authors for different countries and periods.]
How Elsevier uses TDM to help researchers
Big Data to support the World of Research

it will all be about networks
Mendeley, a Social Network for Researchers

Stephane Berghmans
DVM, PhD
VP EU Academic & Research Relations
Elsevier

Research Interests
Science Policy
Science Strategy
Open Science

About
I am responsible for building and further developing relationships with research organisations and EU representatives aimed at creating a better understanding of Elsevier’s role in the advancement of research and science. Deeply engaged in topics that are high on the EU agenda, such as Open Science, gender diversity and research evaluation, my aim is to help stakeholders in the European scientific and political landscape understand how modern
Zebrafish based assays for the assessment of cardiac, visual and gut function - potential safety screens for early drug discovery


Mendeley Profile - Suggests

Based on all the articles in your library

- The zebrafish reference genome sequence and its relationship to the human genome.
  Howe K, Clark M.D, Torroja C.F, Torrance J, ...
  Nature (2013)
  + Save reference
  - Get full text at journal
  ➤ Share reference with followers

- Based on all the articles in your library
- Most recently added to your library
- Popular in your discipline
- Trending in your discipline
- Upload articles that reflect your current research
- Set research interests in your profile

Based on your research interests

Science Policy  Science Strategy  Open Science

Open science resources for the discovery and analysis of Tara Ocean
When publications lead to products: The open science conundrum in new business models
Curing TB with open science
Conclusions

• Text & Data Mining is important for:
  • The World of Research
  • Policy-Making and Decision-Making
  • Elsevier

• We facilitate TDM for researchers, the public sector and the private sector

• Elsevier is already partnering in the EU and we are interested in further collaboration
Thank you

For further information: N.Fowler@elsevier.com