

Advancing Text and Data Mining

Libraries, Researchers, and Elsevier
working together

Gemma Hersh

9th June 2015

Outline

- What is text mining?
- How does text mining work?
- Elsevier and TDM
- Addressing concerns
- Industry Initiatives

What is Text Mining?

What is text and data mining (TDM)?

- Process of deriving information from large amounts of text or data content
- Computers automatically extract information like trends, word patterns and other types of relationships
- Uses programs which learn about the text in order to extract the right information (known as natural language processing)
- Potential to lead to new discoveries and resources



Why researchers use TDM

Two main use cases:


1. Answering a specific research question

- How long does it take for concepts in STM literature to reach general media?
- What is the relationship between the research and consulting commitments of economics and finance professors?
- What are the characteristics of subjects in social psychology experiments?

2. Building a new data resource for the community

- An HIV mutation database for which mutations found in literature are mapped to the underlying sequence database
- A database on growth and alimentation of fishes, and a fish classification to identify new species for aquaculture
- A database with the electrophysiological properties of diverse neuron types

[illegible]


[About](#)
[Neuron Types](#)
[Electrophysiology Properties](#)
[Articles](#)
[FAQs/Methods](#)
[Contribute](#)
[API](#)
[Contact](#)

Listing of neuron types in the database

Electrophysiology values across neuron types obtained are obtained from literature text-mining.
(neuron types are mostly from [Neurolex.org](#))

Show entries
 Search:

Neuron type	Number extracted electrophysiology values	Number articles
Hippocampus CA1 pyramidal cell	197	34
Neocortex basket cell	142	21
Neostriatum medium spiny neuron	104	14
Dorsal root ganglion cell	97	11
Neocortex pyramidal cell layer 5-6	92	17
Dentate gyrus granule cell	84	15
Neocortex pyramidal cell layer 2-3	57	11
Ventral tegmental area dopamine neuron	57	8
Neocortex Martinotti cell	50	7
Neocortex layer 4 stellate cell	47	6

<http://neuroelectro.org>

How to text mine

TDM can automate the process of discovering relationships between interesting things

CONTENT



INTERESTING THINGS...

Enzyme

Response

DNA

Toxin

RELATIONSHIPS...

Enzyme

DNA

Response

Toxin

Pre-processing

Name entity recognition

Relationship extraction

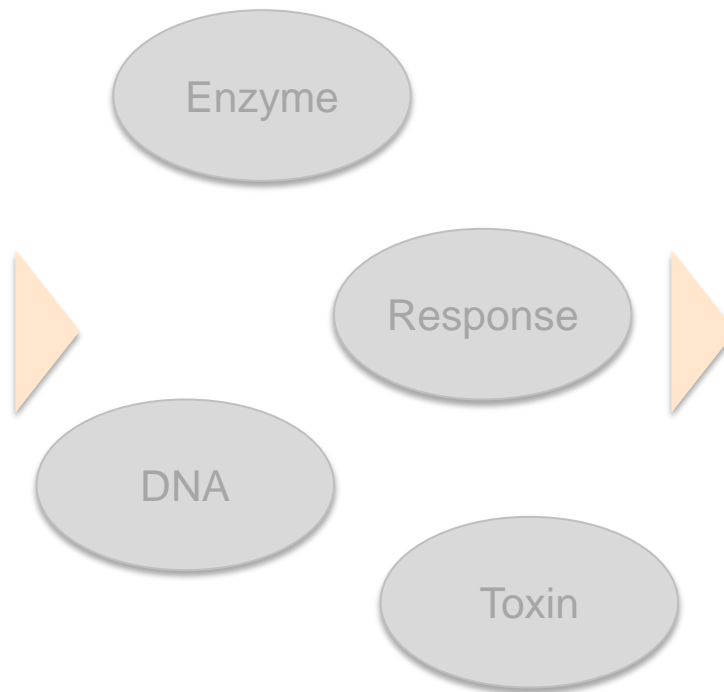
First you want good, clean content...

CONTENT



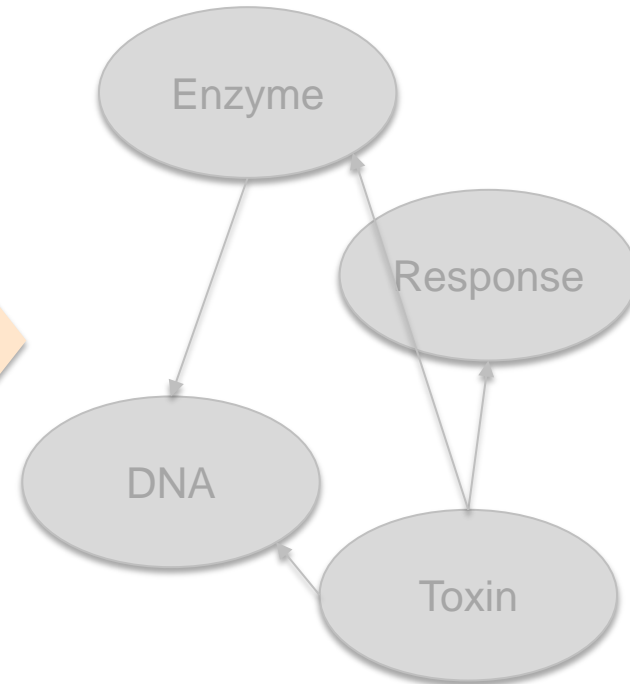
Pre-processing

INTERESTING THINGS...



Name entity recognition

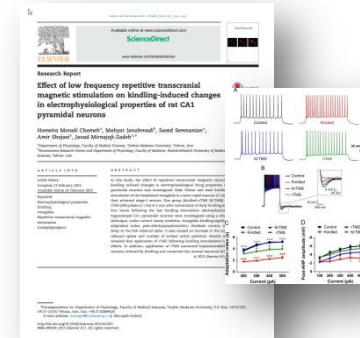
RELATIONSHIPS...



Relationship extraction

Published literature?

- The information source that researchers seek already exists.
- Researchers can build on existing results published in scholarly articles.
- It's a curated and trusted source of information.
- There is a lot of it, it's cross discipline and it covers archives as well.



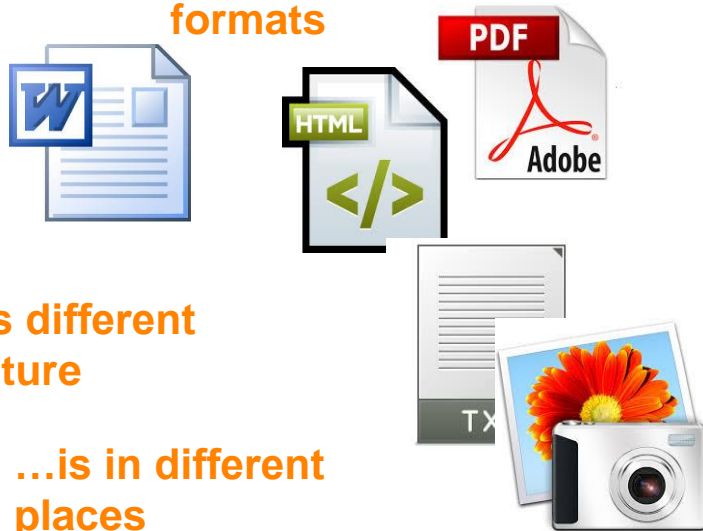
Peer review



But online, data is fragmented and of varying quality...

Content on the web ...

...has different
formats

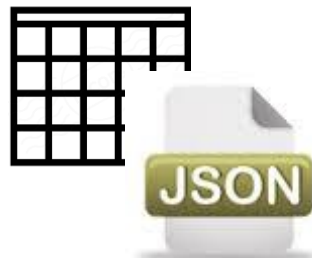


...has different
structure

...is in different
places



...is of variable
quality



This is where publishers can help



More on this later...

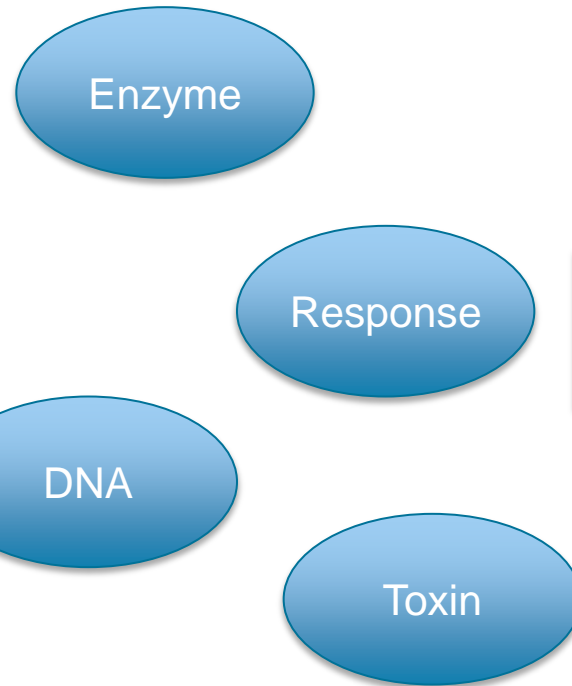
Next comes identifying the interesting things

CONTENT



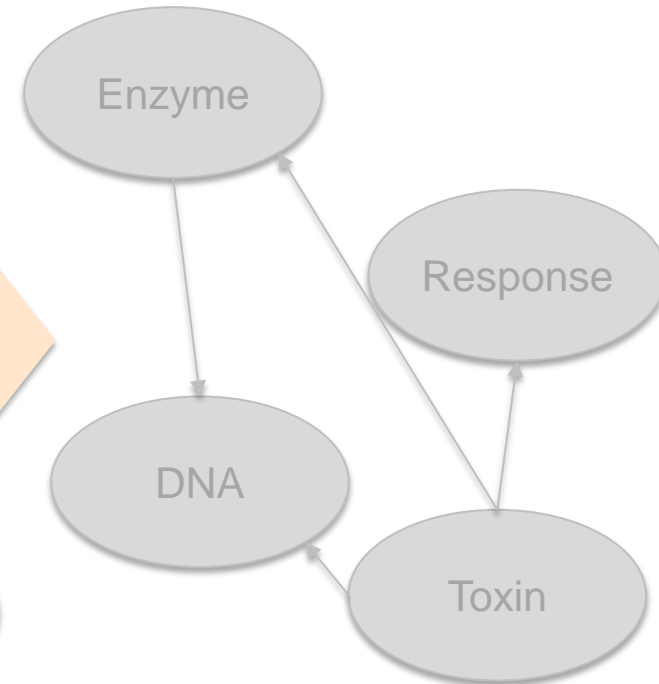
Pre-processing

INTERESTING THINGS...



Name entity recognition

RELATIONSHIPS...



Relationship extraction

Named entity extraction

“The **Airbus** **A380-800** has been a
first to use new technology over
the skies of **London**.”

The diagram illustrates named entity extraction for the sentence "The Airbus A380-800 has been a first to use new technology over the skies of London." The entities are: "Airbus" (Manufacturer, red box), "A380-800" (Model, blue box), and "London" (City, green box). A purple box labeled "Plane" is positioned above "Airbus" and "A380-800".

but even with a simple example its not that easy...



Alternative names...

Airbus A380-800

Airbus A-380-800

Airbus A 380-800

Airbus A380/800

Airbus a380

A380-800

...

Abbreviations, irregular spelling, mis-spelling...

Arbus A380-800

Airbus A380800

...

Finally we pull out the relationships...

CONTENT



INTERESTING THINGS...

Enzyme

Response

DNA

Toxin

RELATIONSHIPS...

Enzyme

DNA

Response

Toxin

Pre-processing

Name entity recognition

Relationship extraction

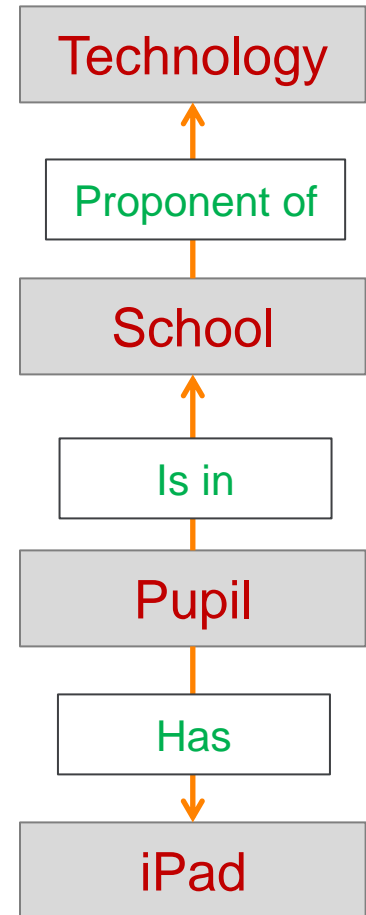
...though it looks easy, it's very difficult...

“The **school** has been a **proponent** of using **technology** in lessons every **pupil** **in** the school **has** their own **iPad**...”

proponent (school,
technology)

has (pupil, iPad)

in (pupil, school)



A typical text mining workflow

Define the problem

What are you wanting to learn?

What relationships do you need to know?

What content do you want to cover?

Define your approach

Choose your tools:

- Off the shelf (open source)
- Outsource to a specialized company

Write an algorithm that will extract what you want

Test the algorithm to ensure it works

- Use corpus and tree-banking to help

Access & extract

Gain access to the published literature

Download all the content

Retest your algorithm

Run your tools and analyze the error reports

Re-run if needed

Analyze & build

Process extracted data

Analyze to see if relationships exist

Write your results in a paper

Build the database of extracted values

Elsevier and Text Mining

Supporting TDM at Elsevier

2006 Started supporting *ad-hoc* TDM access requests

- Low - but increasing level - of interest from early adopters

2012 First content mining policy published

- Developed new APIs & rolled out better TDM technical solutions

2013 Pilot (~30 academics)

- Discover two main user cases for TDM and input to refine our TDM technical solutions

2014 Publish updated TDM policy

- Automatically give access to subscribed users
- Launch a developers portal to facilitate access

2015 Launch STM Open Access Corpus

- Valuable open access tool to help researchers refine and develop their tools in order to TDM more effectively

TDM Pilot Learnings – Research Challenges

Technical

- Obtaining necessary infrastructure
- Having to deal with different formats from content providers
- Sourcing and understanding TDM technology

Functional

- Fine-tuning pipeline, curating output, representing output meaningfully

Logistical/Legal

- Gaining access to the needed content
- Gaining permission to mine the content

TDM Pilot Learnings – Library Challenges

Expertise

- Understanding specific TDM-based projects well enough to assess implications & offer advice to patrons

Legal

- Understanding and tracking what is allowed for what resources
- Negotiating permissions with multiple providers
- Ensuring academic freedom is protected

Financial

- Concerns about any additional costs
- Understanding how TDM affects usage figures for the library

Elsevier's TDM policy

- Researchers at academic institutions can text mine subscribed content on ScienceDirect for non-commercial purposes via the ScienceDirect APIs
- Access is granted to faculty, researchers, staff and students at the subscribing institution

Text mining output can be shared publically under these conditions

- May contain "snippets" of up to 200 characters of the original text
- Proprietary notice
- Must include DOI link to original content

Open access content

- Text and Data mining permission are determined by the author's choice of user license.
- This information is detailed in the individual articles

How to get access?

What do researchers need to do?

1. Register their project at <http://dev.elsevier.com>

ELSEVIER Developers

Chris Shillum | Logout

My Projects

The My Project page provides you with a list of your applications and registered websites.

My Registered Websites/Applications		
Registered on	Website URL	Actions
01 Apr 2014	http://www.elsevier.com	View API Key Edit Delete

Register a new site

Register your website or application here!

[Register a New Site](#)

My Text Mining Projects		
Registered on	Project name	Last message

[Manage Vocabularies >>](#)

Register a new text mining project

You can register new text mining projects here.

[Register a New Text Mining Project](#)

How to get access?

What do researchers need to do?

1. Register their project at <http://dev.elsevier.com>
2. Accept a simple registration form

The screenshot shows the 'Register a New Text Mining Project' page on the Elsevier Developers website. The page has a header with the Elsevier logo and a navigation menu. The main content area is titled 'Register a New Text Mining Project' and includes a brief explanation of the service. Below this, there is a form with the following fields and options:

- Project name:** A text input field containing 'My TDM Project'. A red asterisk indicates it is a required field. A note below the field states: '* 100 character maximum (do NOT use = and | characters)'. The field is highlighted with a red oval.
- Project Type:** A dropdown menu with two options: 'Hosted' and 'Local'. The 'Local' option is selected. A link '(What's this?)' is next to the dropdown.
- Agreement:** A checkbox labeled 'I agree with the Text Mining Agreement.' is checked. This checkbox and its label are also highlighted with a red oval.
- Buttons:** At the bottom of the form, there are two buttons: 'Register Project' and 'Cancel'.

The background of the page is a light gray, and the form itself is a white box with a thin border. The overall layout is clean and professional.

How to get access?

What do researchers need to do?

1. Register their project at <http://dev.elsevier.com>
2. Accept a registration form
3. Obtain an Elsevier API Key

The screenshot displays the 'ELSEVIER Developers' interface. At the top, there's a navigation bar with 'My Projects' and a user profile 'Chris Shillum | Logout'. Below this, the page is titled 'Edit Text Mining Project'. Under the 'Project Information' section, there are three fields: 'Project Name' with the value 'My TDM Project' and a note '100 character maximum (do NOT use = and | characters)'; 'API Key' with the value 'e51ee66ea608b2879c260448ce379603', which is circled in red; and 'Project type' with the value 'level'. At the bottom of the form, there are three buttons: 'Save Changes', 'Return to My Projects', and 'Delete Project'.

TDM Access: What do Institutions need to do?

- TDM access clause will be part of standard ScienceDirect subscription agreement for new academic customers and upon renewal
- For existing agreements, an add-on contract amendment is available – just contact your Elsevier Account Manager
- After signing institutional agreement/amendment, access to our API key registration page for your researchers will be enabled for your institution's IP address range
- We facilitate access for researchers at non subscribing institutions too

Addressing concerns

What is an Application Programming Interface (API)?

- An interface for software programs that enables interaction with other software.
- APIs make the TDM process more efficient for vast quantities of data
- They enable publishers to maintain site performance, availability and reliability for their (human) users.
- Deliver extracted text in an optimal format, preferred by text miners

APIs are standard across industries:

- Wikipedia: http://en.wikipedia.org/wiki/Wikipedia:Database_download
- Twitter: <https://twitter.com/tos>
- PubMed Central: <http://www.ncbi.nlm.nih.gov/pmc/about/copyright/>

Practicalities

1. Registration form

- Automatic process
- No additional liability
- Aligned with institutional e-amendment
- Provides guidelines on reuse
- Can offer one-to-one support
- Removal of project description in response to feedback



Practicalities

2. Mining images

- We always have done on request
- Now do automatically
- Terms of reuse, as we aren't always the rightsholder



Practicalities

3. TDM outputs

- No copyright claim on original research findings
 - We do have output requirements when our material is quoted as follows, in line with standard citation practices:
- To distribute the TDM Output externally, which may include a few lines of query-dependent text of individual full text articles or book chapters which shall be up to a **maximum length of 200 characters** surrounding and **excluding the text entity matched (“Snippets”)** or **bibliographic metadata**.
 - Where Snippets and/or bibliographic metadata are distributed, they should be **accompanied by a DOI link** that points back to the individual full text article or book chapter.
 - Where images are used you should clear the rights for reuse with the relevant copyright owner and/or rightsholder.
 - TDM Output should include a proprietary notice in the following form:
“Some rights reserved. This work permits non-commercial use, distribution, and reproduction in any medium, provided the original author and source are credited.”
 - Open access content usage terms defined by license



Practicalities

4. Flexibility

- We consider all requests and help all researchers with any queries
- Continued evolution in response to feedback: project descriptor, license outputs, images

Elsevier.com > About Elsevier > Policies > Text & data mining

Text and data mining

As the volume of scholarly output increases, we recognize that researchers are increasingly interested in using tools such as text and data mining (TDM) to explore patterns and trends across large databases of content. Elsevier has for many years actively collaborated with researchers and institutes to facilitate text and data mining, both by enabling access and by investing in our platforms to support these initiatives. Our text and data mining policy is intended to be a set of guidelines to help researchers understand how to easily access and mine content published by Elsevier.

We are always happy to support and accommodate researchers who have specific text mining needs and for any questions please email universalaccess@elsevier.com

Feedback

Conditions of text
mining

How to gain access

Policy

Feedback

Elsevier's updated policy is a reflection of our ongoing collaboration within the research community. We encourage anyone with feedback and ideas to contact us via universalaccess@elsevier.com.



Developing NLP Tools

- Elsevier has provided a freely-available and redistributable corpus.
- For each article in the corpus we provide:
 - XML source
 - A plaintext version for easier text mining
 - Several versions with different annotations. (speech tags, sentence breaks, simple noun and verb phrases, root forms of words, syntactic constituents parse trees, Wikipedia concept identification, and discourse analysis
- Kick-started the process of manually creating test sets,
 - Commissioned a *treebank* of 10 full-text articles to be used as a default test set.

110 articles

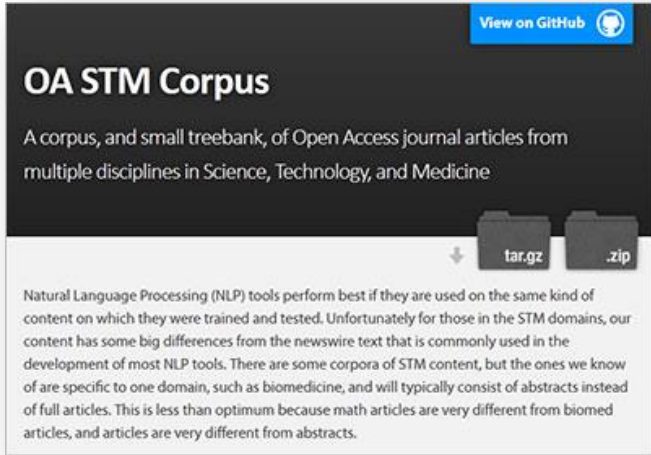
Open access with a CC-BY license

10 domains

agriculture, astronomy, biology, chemistry, computer science, earth science, engineering, materials science math and medicine

10 test sets

<http://elsevierlabs.github.io/OA-STM-Corpus/>



The screenshot shows the GitHub repository page for 'OA STM Corpus'. At the top right is a 'View on GitHub' button with the GitHub logo. The title 'OA STM Corpus' is prominently displayed. Below it, a description reads: 'A corpus, and small treebank, of Open Access journal articles from multiple disciplines in Science, Technology, and Medicine'. There are two download buttons labeled 'tar.gz' and '.zip'. A paragraph of text explains that NLP tools perform best on similar content, and this corpus provides a good example of domain-specific text (STM) that differs from general newswire text.

View on GitHub

OA STM Corpus

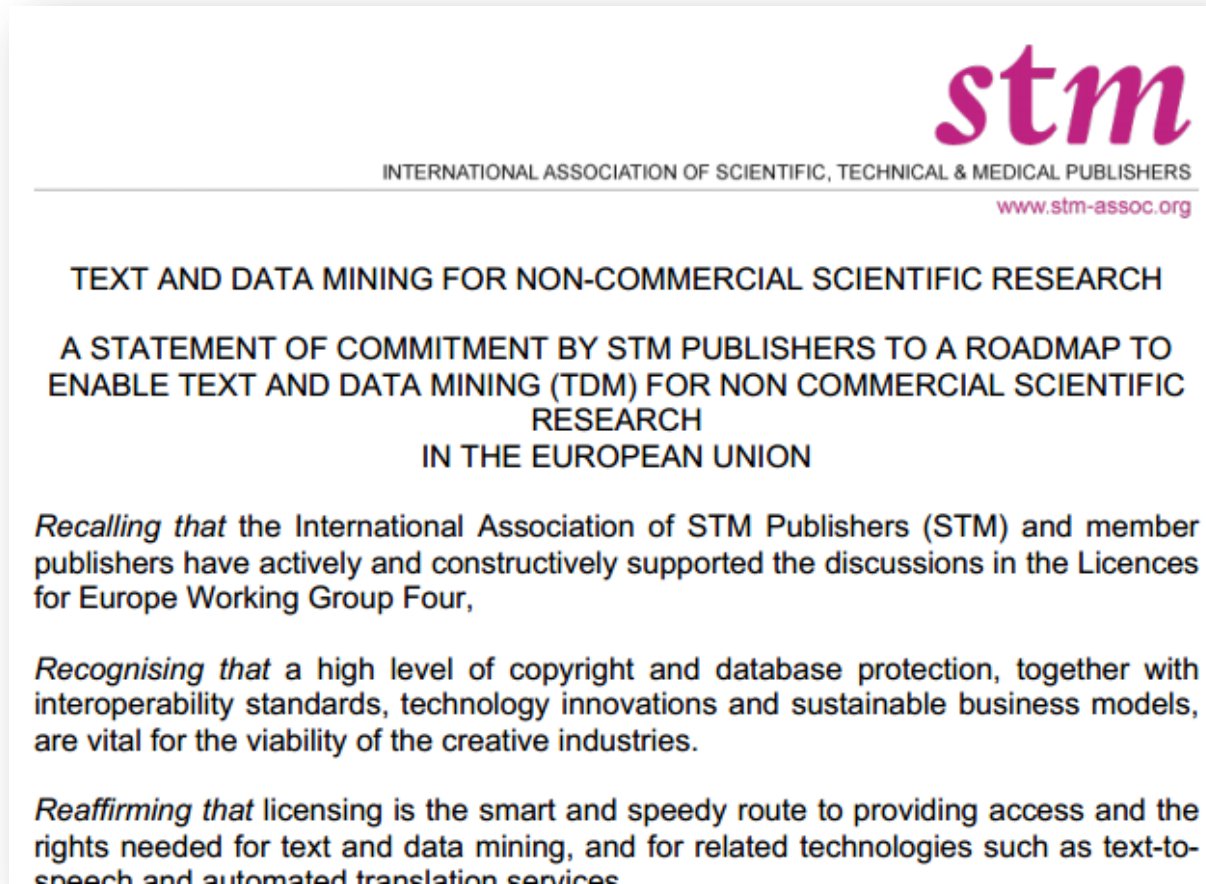
A corpus, and small treebank, of Open Access journal articles from multiple disciplines in Science, Technology, and Medicine

tar.gz .zip

Natural Language Processing (NLP) tools perform best if they are used on the same kind of content on which they were trained and tested. Unfortunately for those in the STM domains, our content has some big differences from the newswire text that is commonly used in the development of most NLP tools. There are some corpora of STM content, but the ones we know of are specific to one domain, such as biomedicine, and will typically consist of abstracts instead of full articles. This is less than optimum because math articles are very different from biomed articles, and articles are very different from abstracts.

Industry initiatives

Aligned with recent STM Declaration on TDM



Supporting cross-publisher TDM

Addressing two issues:

1. Problem: Researchers want to be able to access full text content from multiple publishers' sites for OA or subscribed content in a consistent way.

Solution: Common API (protocol) for requesting machine readable full text from many different publishers

2. Problem: Researchers want to know whether text and data mining is allowed, and if not, get permission.


Solution: Licensing information embedded in article metadata and a registry for supplemental text and data mining terms and conditions (licenses).



Welcome to CrossRef's click-through service for text and data mining. Here you can review publishers' terms and conditions for text and data mining. You have the opportunity to accept or reject these agreements or postpone your decisions until later. When you agree to a publishers' terms and conditions this action will be logged against your API token. You can then use this API token to request full text from the publisher (subscription permitting).

Common agreements used by multiple publishers

Agreement name ▲	Publishers	Status	Action
Publisher-Specific Agreements			
Publisher Name ▲	Terms and Conditions	Status	Action
	Example License from CrossRef	Not yet reviewed	view
	Journal of Zhejiang University Science	Not yet reviewed	view
	Elsevier Text and Data Mining Service Agreement	Accepted	view



<http://www.crossref.org/tdm/index.html>

Status of CrossRef TDM

- BETA version is live

Participating Publishers

- AIP
- IoP
- HighWire
- Walter de Gruyter
- Elsevier
- KAMJE
- PLOS
- Journal of Zhejiang University Science
- Wiley
- Springer
- APS
- BMJ

13.5 Million

Articles with full test links
and license information

Capacity for

36 Million+

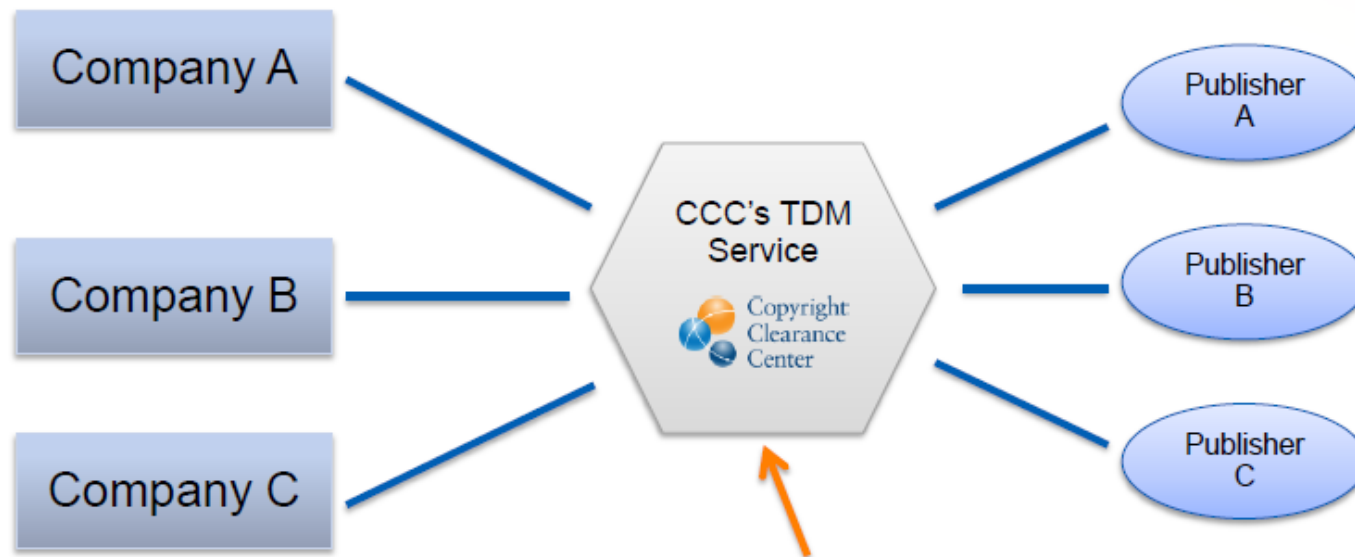
DOI records by mid 2015

FREE

Public service

Copyright Clearance Center Text and Data Mining Pilot

- Pilot service to help researchers access and download XML content in a standardized way.
- Commercial service aimed at researchers working in corporate settings



The service is designed to be used in conjunction with existing enterprise text mining software and methods.

Elsevier and text mining

- Actively collaborating with researchers & institutes to facilitate TDM
 - 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
 - Sponsoring academic conferences, e.g.
<http://www.clips.uantwerpen.be/clin25/home>
- Enabling access with an updated policy framework
- Providing separate channels for machine-to-machine and human access to content.
- Investing in our platforms to support researchers to text mine and launched a developers portal
- Providing technical support and services beyond the basic content access.
- Working on Natural Language Processing tools; Providing resources such as an open access STM corpus and tree-banking
<http://www.elsevier.com/tdm>

Thank you!



g.hersh@elsevier.com



[@gemmahersh](https://twitter.com/gemmahersh)