

# Text & Data Mining (TDM)

Dr. Haralambos Marmanis  
CTO / VP, Engineering  
Copyright Clearance Center

# What do we do?

## Rightsholders



600+ million rights from

- Publishers
- Authors
- Creators

- 
- **Licensing Solutions**
  - **Rights Management**
  - **Content Delivery**
  - **Copyright Education**
- 

## Content Users



- 35,000 companies
- Employees worldwide
- Users in 180 countries

# Overview of TDM

- What is it? How is it done?
- Why people care?
- An example: Drug discovery process
- Practical challenges
- CCC Solution Walkthrough

# What is TDM?

The Discovery of (new) knowledge  
from a set of natural language  
resources (e.g. articles, books) and  
other structured or unstructured data

# How is it done?

1. Identify a set of resources that are relevant to a particular research objective
2. Analyze and extract information specific to the research objective
3. Develop and explore the various relations between extracted objects of interest

# Why people care?

Biomarker discovery  
Drug repurposing  
Drug safety  
Competitive intelligence  
Sentiment analysis

.....

Legend: positive negative hsd positive hsd negative hsd possible

Current Document: pr003.txt Document Sentiment Score: 0.392726331949234

NewsMax.com Names Lazaro Gonzalez 'Hero of the Year' WEST PALM BEACH, Fla., Jan. 1 /PRNewswire/ -- The online news agency NewsMax.com has named Lazaro Gonzalez as Hero of the Year for 2000. NewsMax.com editors selected Gonzalez as Hero of the Year because, like other great heroes, he made "a great sacrifice and by his actions changed the course of history." When Elian Gonzalez was plucked from the ocean on Thanksgiving Day in 1999, Lazaro Gonzalez had little idea the impact he would have on American politics. An immigrant who barely spoke English, with humble financial means, he seemed no match for the power of the Justice Dept., the federal government and many in the major media. Still, he held his ground and tried to keep his nephew on free soil. Elian was eventually returned. But Gonzalez's actions had significant repercussions. One was that George Bush was elected to the White House because thousands of Democratic-leaning Cuban-Americans voted for Bush. NewsMax.com says that Lazaro Gonzalez is an "unusual hero." "Today, we call someone a hero that is famous, well liked, or a celebrity, but they aren't really heroes," NewsMax.com editor Christopher Ruddy explained. "Heroes are ordinary people who do extraordinary things." Wired News recently reported that NewsMax.com's "popularity has skyrocketed during the last couple of months." NewsMax.com has already been rated #1 by Deja.com. The Wall Street Journal Business Report calls it one of America's leading alternative Web sites, and Talkers magazine says it's the number 1 source for radio and TV producers for story ideas. For more information, go to: www.NewsMax.com Contact person at

# “Drug Discovery” Process

Goal: Develop new treatments for diseases through hypothesis formation.

Methodology:

Keyword/Database Searching

Review Literature

Find relationships

Generate hypotheses

Test

Product development

Etc.



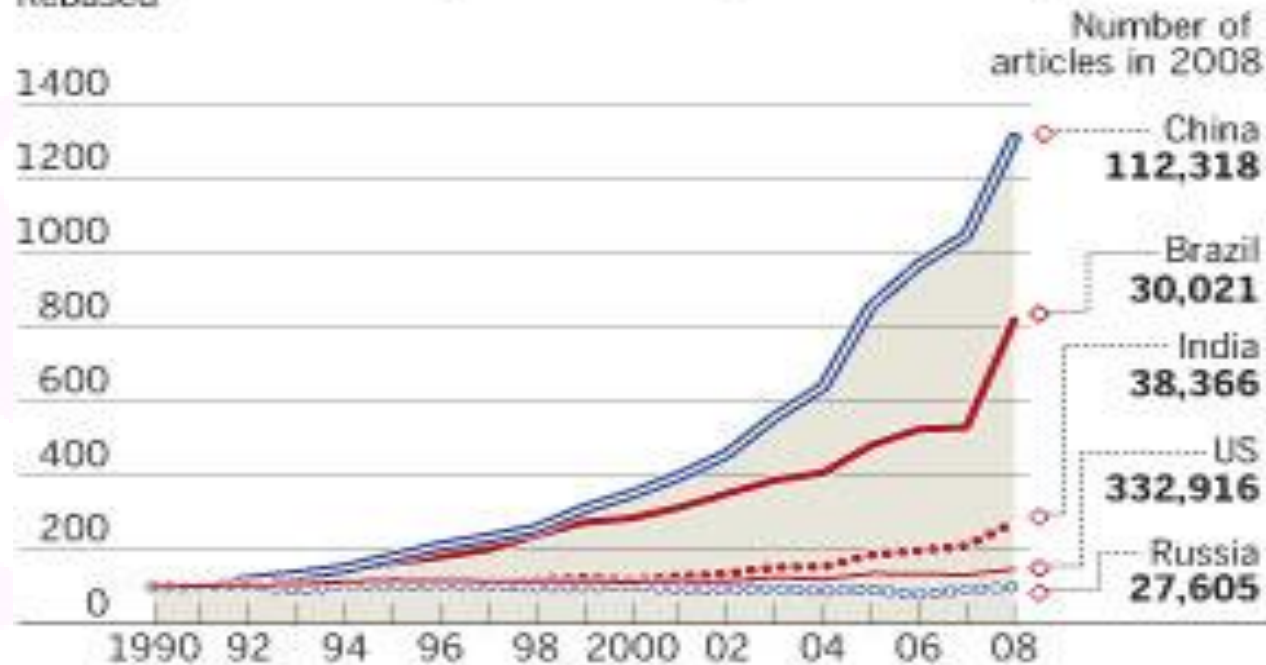




# Too Much Research

## Growth of articles published in peer-reviewed journals

Rebased



Sources: Thomson Reuters; Web Science Database

53M Records in Scopus

800,000 Journal Articles published per year

# Even within a single research area

Lots of disorders ...

[Angina](#)

[Acute coronary syndrome](#)

[Alexia](#)

[Anomic aphasia](#)

[Aortic dissection](#)

[Aortic regurgitation](#)

[Aortic stenosis](#)

[Apoplexy](#)

[Apraxia](#)

[Arrhythmias](#)

[Asymmetric septal](#)

[hypertrophy \(ASH\)](#)

[Atherosclerosis](#)

[Atrial flutter](#)

[Atrial septal defect](#)

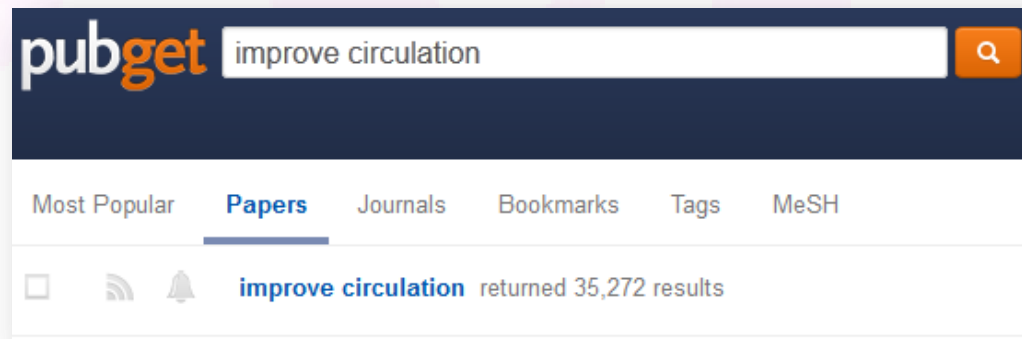
[Atrioventricular canal  
defect](#)

[Atrioventricular septal  
defect](#)

[Avascular necrosis](#)

Lots of documents...

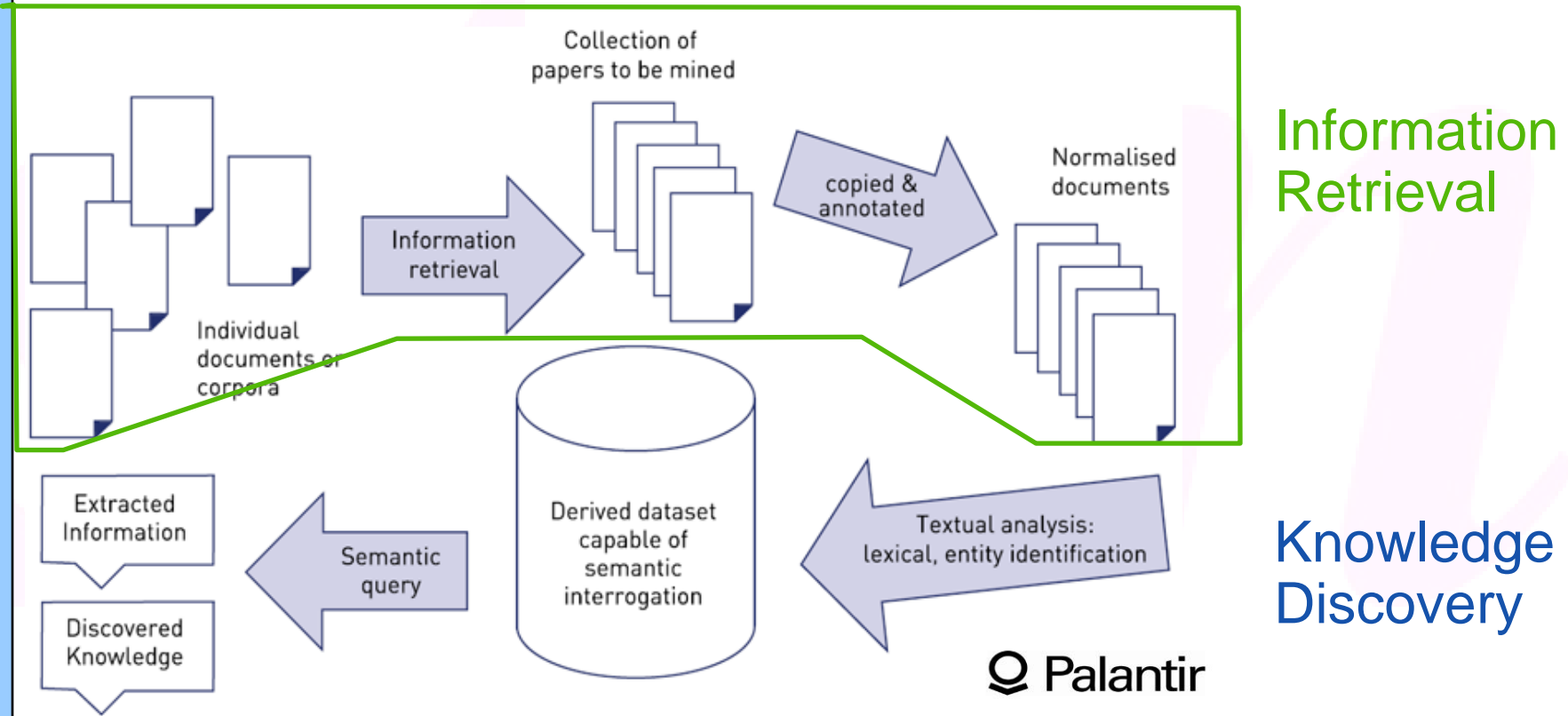
- 35,000+ on Improve Circulation
- 7,000+ per disease area



# More practical challenges

- **Many sources** of content
- **Many formats**
- Difficult to obtain **full-text** in XML
- Difficult to **integrate** content into TDM software.
- Hard to negotiate and manage **licenses** and feeds from all publishers.

# Data Processing Workflow



# Solving the retrieval problem

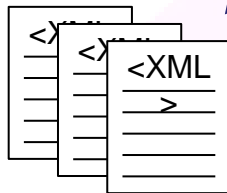
## A single retrieval platform:

- **Speed up** time to obtain properly **licensed** content for text mining
- **Discover** and **download full-text in XML**, not just abstracts
- **Main corpus** includes Subscribed and Not-Subscribed content
- **Normalize** XML format across many publishers
- Provide a **Web UI** and **RESTful API** services

# Publishers



1. Publishers provide **content** and **rights**



**DirectPath  
TDM Service**

2. Researchers create content sets by using search or other discovery criteria



3. Researchers slice and dice results and identify an appropriate corpus for their project

4. XML corpus can be imported into various TDM tools

**TDM  
Software**

**XML  
Article  
corpus**

## ADMIN PORTAL

Subscriptions

Digital Library

Reports

Approval Requests

Agreements

Subscribed Titles

Platforms

Search:

All titles

SEARCH

Add Title

Upload Title List

View upload history

Download List

Download List

1 - 100 of 13329

Previous 1 2 3 4 5 Next

Journal	ISSN	Platform	Start Date	End Date	Months Back	Embargo (months)	Mechanism	Date Added	Actions
Advanced Emergency Nursing Journal	1931-4485	HeinOnline	19 Feb 2012	31 Dec 2014	none	none	Manual Upload	Invalid date	Edit   Delete
Advanced Emergency Nursing Journal	1931-4485	American Association of Neurosurgeons			none	none	Manual Upload	Invalid date	Edit   Delete
Advanced Emergency Nursing Journal	1931-4485	AGU			none	none	Manual Upload	Invalid date	Edit   Delete
Advanced Emergency Nursing Journal	1931-4485	AIP			none	none	Manual Upload	Invalid date	Edit   Delete
Journal of Cardiopulmonary Rehabilitation and Prevention	1932-7501	American Association of Neurosurgeons			none	none	Manual Upload	Invalid date	Edit   Delete


Content &amp; Rights

XML for Mining

 Your Projects

## YOUR PROJECTS

Create Project

 1 - 10 of 10  Delete

<input type="checkbox"/>	Creation Date ▾	Project ⇅	Description	Actions	Status ⇅	Results
<input type="checkbox"/>	19 Nov 2014	<a href="#">Her2 Breast Cancer</a>		<a href="#">View options ▾</a>	Completed	<a href="#">View</a>
<input type="checkbox"/>	19 Nov 2014	<a href="#">NPC1</a>		<a href="#">View options ▾</a>	Completed	<a href="#">View</a>
<input type="checkbox"/>	18 Nov 2014	<a href="#">testing build on 11/18 - cancer</a>		<a href="#">View options ▾</a>	Completed	<a href="#">View</a>
<input type="checkbox"/>	13 Nov 2014	<a href="#">Breast cancer demo</a>		<a href="#">View options ▾</a>	Completed	<a href="#">View</a>
<input type="checkbox"/>	10 Nov 2014	<a href="#">demo test cancer</a>		<a href="#">View options ▾</a>	Open	----
<input type="checkbox"/>	07 Nov 2014	<a href="#">NPC1 - second try</a>		<a href="#">View options ▾</a>	Preparing download	----
<input type="checkbox"/>	03 Nov 2014	<a href="#">Hemagglutinin</a>	Project preview says 140 results	<a href="#">View options ▾</a>	Ready for download	<a href="#">View</a>
<input type="checkbox"/>	03 Nov 2014	<a href="#">Cancer w date range</a>	added date range	<a href="#">View options ▾</a>	Completed	<a href="#">View</a>
<input type="checkbox"/>	23 Oct 2014	<a href="#">Alzheimers</a>	A project to create a description that describes how this is a project	<a href="#">View options ▾</a>	Preparing download	----
<input type="checkbox"/>	17 Oct 2014	<a href="#">Diabetes</a>		<a href="#">View options ▾</a>	Completed	<a href="#">View</a>

1 - 10 of 10



# CREATE PROJECT

Project name:

Polio

Project description: (Optional)

Find all documents with mention of Polio (Poliomyelitis)

## SEARCH RETRIEVAL METHOD ?

- Search query analysis
- Nearest neighbor analysis
- Article ID list

Enter your search query:

[? Search Tips](#)

Full Text

Poliomyelitis

AND

- Full Text
- Document title
- Authors
- Journal title
- Journal ISSN
- Abstract
- Introduction
- Materials and Methods
- Conclusion
- Citations

[Show limiters](#)

[Search citations ?](#)

Do you want to view a preview of the results before requesting the full text?

[View Preview](#)



OR, do you want to go ahead and get the full text associated with the matches to your query?

[Get Full Text](#)

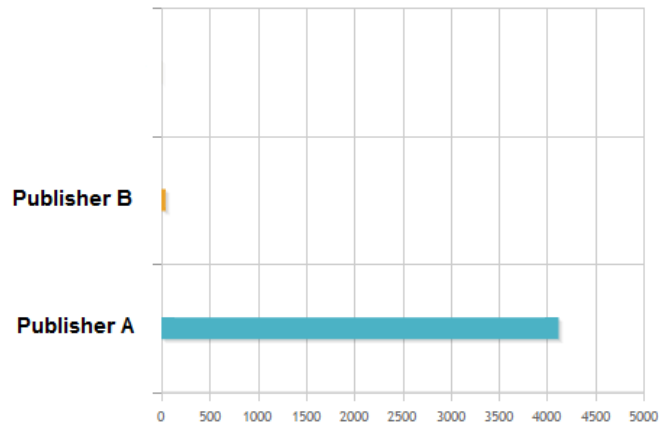
# PROJECT PREVIEW

Project name: Diabetes  
Criteria: Full Text: diabetes  
Number of results: 4165

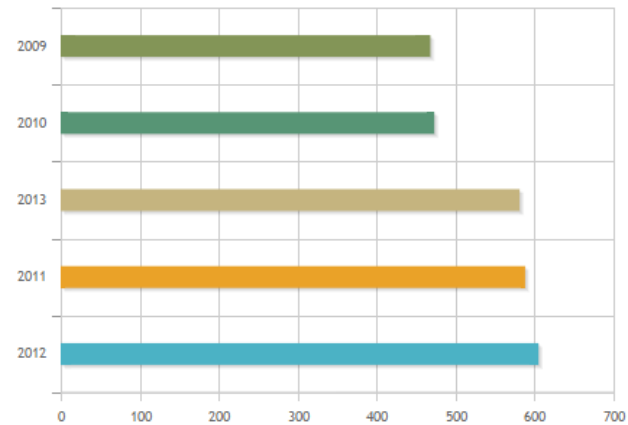
Get Full Text

Revise search  Save project for later  Cancel

### Number of Results by Publisher



### Number of Results by Year



### Number of Results by Journal

(Top 41)

Journal	ISSN	Rightsholder	Number of results
Future Cardiology	1479-6678	Rightsholder 1	364
Pharmacogenomics	1462-2416	Rightsholder 2	352
Aging Health	1745-509X	Rightsholder 3	332
Clinical Lipidology	1758-4299	Rightsholder 1	275
Diabetes Management	1758-1907	Rightsholder 2	268
Regenerative Medicine	1746-0751	Rightsholder 3	265
Clinical Lipidology	1758-4302	Rightsholder 3	254

Filter your results: Project: **Diabetes test 2** [View project details](#)

Publication Years

- 2013 (293)
- 2014 (78)
- 2012 (15)
- 2011 (7)
- 2010 (6)

Publisher

- DEFAULT\_TDM (399)

Through subscription

- Unsubscribed (399)

QUERY: Full Text: Diabetes


Add Additional criteria to refine these results further

 Search

 Clear

Create Download

Results: 1 - 100 of 399

100 Results/page  Previous 1 2 3 4 Next

### Age and psychological influences on immune responses to trivalent inactivated influenza vaccine in the meditation or exercise for preventing acute respiratory infection (MEPARI) trial

 Not subscribed  
[View article](#)

Hayney, Mary S; Coe, Christopher L; Muller, Daniel; Obasi, Chidi N; Backonja, Uba; Ewers, Tola; Barrett, Bruce, *Human Vaccines & Immunotherapeutics*, 2014 Jan 01, Vol. 10 Issue 1, pages 83-91

ISSN: 2164-5515

Publisher: DEFAULT\_TDM

DOI: *n/a*

Language: (English)

the groups is noteworthy. The parent study, from which the current data were derived, showed that both exercise and MBSR reduced the number of days sick from all-cause acute respiratory infection (ARI) illness.<sup>23</sup> The incidence, duration and global severity ... [View more](#)

### Direct inhibition of hexokinase activity by metformin at least partially impairs glucose metabolism and tumor growth in experimental breast cancer

 Not subscribed  
[View article](#)

Marini, Cecilia; Salani, Barbara; Massollo, Michela; Amaro, Adriana; Esposito, Alessia Isabella; Orengo, Anna Maria; Capitanio, Selene; Emionite,

DirectPath

Text &amp; Data Mining

 Your ProjectsFilter your results: Project: **Diabetes test 2** [View project details](#)

## ☑ Publication Years

- 2013 (293)
- 2014 (78)
- 2012 (15)
- 2011 (7)
- 2010 (6)

## ☑ Publisher

- DEFAULT\_TDM (399)

## ☑ Through subscription

- Unsubscribed (399)

QUERY: Full Text: Diabetes

Add Additional criteria to refine these results further

[Create Download](#)


Results: 1 - 100 of 399

100 Results/page ▾ Previous 1

[Age and psychological influences on immune responses to trivalent inactivated influenza vaccine](#)  
[the meditation or exercise for preventing acute respiratory infection \(MEPARI\) trial](#)

Not su  
Vie

# Time's Up!



## About your speaker:

**Name: Dr. Haralambos Marmanis**

**Company: Copyright Clearance Center**

**Tel: +1 978.646.2723**

**Email: [hmarmanis@copyright.com](mailto:hmarmanis@copyright.com)**

**Social Media:**

**<https://www.linkedin.com/in/marmanis>**

**<https://twitter.com/marmanis>**

**<http://www.amazon.com/H.-Marmanis/e/B002BOA806>**