Copyright Clearance Center | RightsDirect

Powering AI Discovery with Scientific Articles

Your guide to accessing, managing, and licensing the content that fuels AI, machine learning, and data visualization projects across the enterprise.

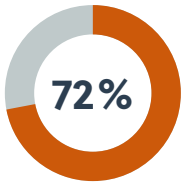
Enriching AI, Machine Learning, and Data Visualization Projects

It's no secret that the amount of information in the digital ecosystem is rapidly expanding. Nearly three million papers are published in scholarly journals annually,[1] and that's just one high value content type for R&D — intensive organizations. Factor in the need for quick and easy synthesis of patent, clinical, and other types of content — and you've got information chaos.

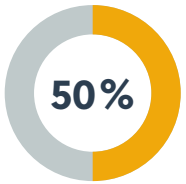
That's where natural language processing-based text mining comes in. It enables researchers to gather important insights from vast amounts of published information and draw insights from massive data sets. This form of artificial intelligence can increase productivity, efficiency, and production — and reduce the cost and time associated with bringing products to market.

Most recently, everyone is talking about Large Language Models (LLMs), like ChatGPT. These tools help accelerate research and drive operational efficiency, though users should be cautious about accepting the accuracy of output. This challenge leads some life science organizations to invest in building and training their own models to serve their business needs.

In this guidebook, we're looking at the ways R&D-intensive companies currently utilize scientific articles and text mining for a wide variety of AI projects, the challenges they're facing, and important considerations to make accessing and licensing this information easier.



72% of business leaders believe that AI will be a significant business advantage in the future, with 67% saying that AI will help them find new opportunities[2]



Nearly 50% of global healthcare companies will implement AI strategies by 2025[3]

Need to Know Terms and Definitions

Text and data mining is a process that uses software to derive high-quality information such as assertions, facts, and relationships from unstructured text (e.g., scholarly articles, internal documents, and more), and identify patterns or relations between items that would otherwise be difficult to discern.

XML: Short for Extensible Markup Language, XML is an information exchange standard designed to simplify interoperability and maintain data integrity when transferring the data from one system to another.

Semantic Enrichment: Semantic enrichment is the enhancement of content with information about its meaning, thereby adding structure to unstructured information. Semantically-enriched content has been annotated with its meaning, enabling users to more quickly find more relevant, precise content.

TDM Rights: There are a number of copyright-sensitive acts that go hand-in-hand with the text and data mining (TDM) process. Content may be copied, stored, annotated, enriched, and otherwise scanned to produce a useable research output. In many cases, commercial TDM rights are not included in standard subscription agreements from publishers.

Large Language Model (LLM): An LLM is an advanced AI program that has been extensively trained on a vast array of textual data. By learning the intricate language patterns, terminologies, and contextual nuances of life sciences, LLMs can generate coherent and relevant text responses on specific topics. Many scientists are exploring the value of using these models for tasks such as assisting in scientific research, answering queries, and aiding in the interpretation of complex data.

How Different Areas in Life Sciences Companies Use Text Mining Today

Researchers extract key concepts from literature, database pipelines, and patents to inform drug discovery including target identification and prioritization.

Drug safety teams automate some parts of adverse event detection by mining scientific literature, patient cases, and social media for faster response and action by people.

Data scientists mine scientific abstracts, conference reports, and clinical trial records to extract, structure, and visualize relationships among key opinion leaders (KOLs) in a therapeutic area to support recruiting and identify expert speakers.

Product marketers pull metadata to map the most commonly used product sheets, slide decks, and other collateral that medical science liaisons use in the market to inform future content strategies.

Medical science liaisons collect prevailing thoughts of patients, health care providers (HCPs) and KOLs through real world data feeds like social media, call center transcripts, and medical field notes to understand trending perceptions of their products.

Data scientists conduct sentiment analysis on packaging and formulations via social media feeds to understand how patients, health care providers, and the general public perceives their products in order to address potential safety concerns or misconceptions.

Medical affairs analyze customer call feeds to learn about drug switching, off-label use, or contraindicated medications among concomitant drugs to identify trends and potential promotional opportunities.

How Does It Work?

Text mining tools employ sophisticated software which uses natural language processing (NLP) algorithms to read and analyze text.

There are two basic steps:

1. The first step is identifying the entities to be searched. In a biomedical setting, these might include genes, cell lines, proteins, small molecules, cellular processes, drugs, or diseases.
2. The next step is analyzing sentences in which those key entities appear to determine how they are related.

There are two primary use cases and workflows for text mining research projects: focused and enterprise scale. A third, emerging use case is training content for LLMs.

Focused

Research Projects

Key Facts

- Discrete
- Often fixed timescale
- Focused data set is needed

Enterprise Scale

Research Projects

Key Facts

- Larger volume of content required on an ongoing basis
- Often feeds into data visualizations such as knowledge graphs

Full Text or Abstracts?

Many researchers use the summary information in article abstracts to compile a collection of records for text mining, rather than using full-text articles. Even though using abstracts seems like an easy workaround, there are major benefits that come from mining full text.

Benefits of Abstracts	Downsides of Abstracts
Short concise summaries that are easily accessible via databases such as PubMed.	May omit the richness, detail, and granularity available from the full-text papers, particularly in tables.

Benefits of Full-Text	Downsides of Full-Text
Full-text includes detailed descriptions of methods and protocols and the complete study results — ensures that researchers don't miss vital data, discoveries and assertions.	Not typically readily available from publishers in a format suitable for text mining.

Conclusion

Insights that can be found only in the full text of scientific articles undoubtedly enrich the outcomes of AI, machine learning, and data visualization projects.

According to a study published in the Journal of Biomedical Informatics, only **8% of the scientific claims** made in full-text articles were found in their abstracts.[4]

31 (37%) Relationships found in article abstracts only

53 (67%) Additional relationships found in full text of articles

Another study, conducted by publisher Elsevier, compared the use of abstracts and full-text articles to derive relevant information about drugs and proteins that affect the progression of fibromyalgia. They found 31 relationships in the literature by mining abstracts and an additional 53 relationships when they ran the same search across the full-text articles.[5]

Copyright and Licensing Challenges

Researchers typically request the right to text mine published content for commercial purposes, as this right is not commonly included in standard publisher subscription agreements. Converting PDFs intended for human consumption into a machine-readable format such as XML results in the creation of additional copies. Creating and storing those reformatted copies typically requires additional permission from the publisher. The absence of any mention of text mining in the terms of a subscription agreement does not mean that it is permitted.

When working directly with multiple rightsholders and publishers for the use of full-text XML articles, some life science companies report:

- Varying fee structures
- Inconsistent terms of use
- Time-consuming individual negotiations

Additionally, even when publishers do grant text mining rights, they may not provide the content in a format easily ingestible by machines. This requires that researchers and data scientists take the extra step of converting PDFs to a consistent machine-readable format, which introduces the potential for badly rendered XML.

Tip: Save time and effort by taking advantage of voluntary collective licensing and let someone else take on the negotiating for you. Please contact CCC for more details on how we can help. This is also an important time to involve the person or department within your company who manages subscriptions (typically a knowledge or information manager); they'll have insights into what the company currently utilizes and may already have relationships with partners that can help streamline licensing.

Thinking About LLMs

Scientific articles in well-structured and enriched XML-format are a great addition to the vast amounts of content needed to train LLMs. When the metadata is normalized, it's easier to work with large numbers of scientific articles. And when the XML-formatted content includes appropriate rights which allow mining and use of the mined output to support LLMs, it can help simplify copyright compliance. The landscape of LLMs is evolving rapidly, so it's important to understand the issues, both technological and legal. This will ensure that what you choose to do with LLMs will serve your business goals more effectively.

4 Considerations:

Beware of garbage in, garbage out. Seek high quality content to reduce the risk of poor answers. This is especially important for life sciences companies where findings can impact people's health and well-being.

If you use a third-party tool like ChatGPT from OpenAI, be sure you understand what happens to your content, your questions, and your findings. Be aware that the terms and conditions of third-party tools may give them the right to your inputs to further train their models (which may be used by your competitors).

All academic articles carry licensing restrictions — be sure to verify that what you want to do is allowable. If not, then additional rights may be needed.

If you are building your own LLM, look for validated, structured, and enriched content such as normalized XML — it can help accelerate the machine learning processes and improve efficiency.

Finding the Best Approach for Your Organization

Accessing scientific articles in flexible and normalized XML format along with a uniform set of usage rights provides you the best chance to overcome the challenges discussed above.

Consider the following:

Use a third-party search interface to create queries on a project basis, filter results, and download files relevant to your current needs.

Embed third-party functionality into your own automated tools and processes with a RESTful API.

Access XML with a data feed option that delivers specific subscribed content in XML format and offers updated content at a regular cadence.

Learn about CCC Licensing & Software Solutions

RightFind XML from CCC provides the most comprehensive coverage of licensed, full-text article content and text mining rights available on the market today. More than 50 publishers currently participate in RightFind XML, providing a consistent set of text and data mining rights with full-text XML content.

With built-in rights and normalized files from a wide variety of publishers and journals, RightFind XML helps you simplify compliance and content management so you can focus on driving science forward.

References

[1] https://ncses.nsf.gov/pubs/nsb20214/publication-output-by-country-region-or-economy-and-scientific-field

[2] https://www.pwc.com/us/en/tech-effect/ai-analytics/ai-business-survey.html

[3] https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2022-and-a-half-decade-in-review

[4] https://insidebigdata.com/2017/02/23/downside-converting-full-text-pdfs-to-xml-for-text-mining/

[5] https://www.elsevier.com/__data/assets/pdf_file/0016/83005/R_D-Solutions_Harnessing-Power-of-Content_DIGITAL.pdf

Copyright Clearance Center | RightsDirect

Learn more

Contact us at:

copyright.com/business

solutions@copyright.com

About CCC

A pioneer in voluntary collective licensing, CCC advances copyright, accelerates knowledge, and powers innovation. With expertise in copyright, data quality, data analytics, and FAIR data implementations, CCC and its subsidiary RightsDirect collaborate with stakeholders on innovative solutions to harness the power of data and AI.