



## Semantic Enrichment

# The Key to Successful Knowledge Extraction from STM Literature

## Abstract

Scholarly research and the dissemination of research output are increasing at a rapid pace. This offers STM information providers the opportunity to expand their publishing, database development, and knowledge dissemination services. For producers and users of this knowledge, what does this information explosion imply? Does it really aid serious researchers in precisely discovering what they are looking for? And how can the producers of such research information enhance their offering to better meet user expectations?

This white paper identifies the key trends that triggered the information explosion and how this resulted in a paradigm shift in the way information is disseminated and used, especially in the STM domain. It also highlights the key problem areas for researchers in discovering knowledge from an avalanche of research output. Further, this white paper elaborates - with a few apt examples - on how STM publishers can use semantic enrichment in different ways to enhance their offerings to researchers.

Scope, as a leading provider of Knowledge processing services, offers a range of services to STM publishers in the area of semantic enrichment, leveraging its rich domain expertise and legacy of serving global leaders in the field.

## The Explosion of Knowledge Generation

### Journal Articles - Non-patent Technical Literature

Scholarly research and the dissemination of research output are increasing at a rapid pace. This offers STM information providers an opportunity to expand their publishing database development as well as knowledge dissemination services. Here are some statistics on how research output is turned out by the academic and scholarly societies:

- There are now approximately 5.5 million researchers worldwide
- Around 1.4 million articles are written annually by these researchers
- These articles are published in 23,000 scholarly journals
- There are about 2,000 publishers publishing these scholarly journals
- These publishers are made up of learned societies, university presses and independent publishers
- The number of articles and number of journals published each year have increased steadily by about 3% and 3.5% per annum, respectively. The growth in number of researchers is equally persistent, at about 3% per year.

*(Source: International Association of Scientific, Technical and Medical Publishers)*

### Patents

In addition to scholarly research, the knowledge space is further inundated by the growing number of patent applications filed all over the world. According to the World Intellectual Property Organization (WIPO).

- Patent applications filed across the world are estimated to be 1.76 million in 2008, representing a 4.9% increase from the previous year. The number of filings worldwide by applicants from China, Korea and the US increased by 32.1%, 6.6% and 6.7%, respectively.
- Approximately 727,000 patents were granted across the world. Similar to patent filings, patent grants are concentrated in a small number of countries. Applicants from Japan, the US, Korea and Germany received 73% of total patent grants worldwide. Between 2000 and 2006, the number of patents granted to applicants from China and Korea grew by 26.5% and 23.2% a year, respectively (average annual growth rate).

## Grey Literature

The above mentioned statistics indicate only the content available in the public domain offered by publishers and societies. But there is a wide variety of research output, lying in the corridors of academic institutions and government repositories. These publications as “Grey literature” are disseminated as conference papers, technical reports, dissertations and other government publications.

There is overwhelming evidence that grey literature is being produced at an exponential rate. According to one estimate, its rate of growth is three to four times that of conventional literature. Early estimates of the amount of grey literature available and used by the British Library Document Supply Centre (BLDSC) can be considered as an indication of the size and growth of grey literature. In 1992, the BLDSC held about three million documents that had been collected over the span of 30 years, and the growth rate at that time was 150,000 documents per year. Ten years later the BLDSC has over 17 million such documents.

## New Media (blogs, wikis)

This exploding content space is further inundated by a new publishing phenomenon that is collectively called as “social media or web 2.0”. With the advent of web 2.0, publishing has been totally democratized and attained a new dimension in the form of users as the creators of content (blogs, wikis), collective intelligence (collaborative content, forums, online communities) and new and emerging media formats like podcast, videocast etc.

## Key Trends in STM Information Dissemination and Usage

For producers and users of this knowledge, what does this information explosion imply? Does it really aid serious researchers in precisely discovering what they want? How can the producers of such research information enhance their offerings to meet user expectations?

Before examining the repercussions of this information explosion for the stakeholders, we need to discern the key trends that triggered it in the first place, and how this resulted in a paradigm shift in the way information is disseminated and used, particularly in the STM domain:

- Digital Publishing - EPS forecasts that by 2014, 95% of journals will be available in electronic format
- Increasing usage of online search and citation linkage
- STM publishing process increasingly driven by specialised knowledge and depth in domain specialisations.
- The grey literature, which does not form part of the data available in the public domain, is also becoming one of the key resource bases for the serious knowledge seekers. But, researchers face the problem of inefficient access to such grey literature since it is not subject to any metadata control and does not come under the purview of abstracting and indexing services, as in the case of the typical STM journal articles.
- With no updated statistics on the size of the grey literature and its growth, the STM content space is like an expanding universe, providing immense challenges to the knowledge seekers and disseminators in enhancing the user experience of knowledge discovery process.
- The emergence of social media in the form of blogs, wikis etc and its growing importance and usage amongst the research fraternity.

How do the users of research information benefit by these changes? This can be gleaned from these responses by academic scholars:

*"I read more literature than I did prior to e-journals. The total amount of time spent in retrieving information, on a per article basis, has decreased."*

*"I almost never go to the library. The combination of database searching with e-journals is a revolution."*

The key factors that make e-journals the emerging and dominant mode of knowledge dissemination are:

- Time savings
- Reduced number of physical trips to the library
- More time to digest content rather than find it
- Can be browsed and read more widely

## Do Users Really Discover Knowledge?

The Internet has therefore put up an information glut. Before discussing its usefulness and whether it really leads to better knowledge discovery, here are a few pointers to the patterns of reading and usage of scientific information.

Scientists read to:

- probe in new domains - web exploration
- learn, get oriented - textbook-like explanations
- position - directed searching of topic
- compete - directed searching of people
- scan the environment, stay aware - review of sources

Apart from reading, scientists also do the following activities:

Consulting - experimental resources to identify protocols, instrumentation, comparative results

Compiling - customized personal collections, laptops full of PDFs

Extracting - core knowledge base of "facts"

Building - source for database enrichment, annotation, evidence

So, mere accesses, downloads, and citation do not accurately represent any of these dimensions of online use of scientific information. It may be indicators of interest, perhaps, but not represent actual use or value to the scientists.

## So, what are the Key Problem Areas in Knowledge Discovery?

- Even as the STM information space is exploding with more and more content, the users have less and less time to read. Researchers, when faced with the mammoth volume of content available with them on a particular topic, in the form of full text journal articles, conference papers, technical reports etc, find that all their time is spent reading.
- There is a felt lack of context in keyword based searches.
- Researchers need sophisticated techniques to mobilize information according to their specific needs. The needs vary with discipline, research problems and particular research strategies, as well as with the affordability of current technology.

## Limitations of Routine Abstraction and Indexing Services

Researchers mainly depend on the numerous abstracting and indexing services to scan the literature, identify the relevant articles, and decide on their usefulness before downloading the full text articles.

The key limitations of such abstracts and indexes are:

- Many of the abstracts are indicative abstracts - they only provide what is discussed in the article, and do not elaborate on the core theme or idea discussed.
- Even with informative abstracts, the user may not be in a position to know whether the abstracts really captured the key information.
- Literature search is still dominated by a mechanism based on a Google type indexing of keyword- and metadata-based search. The process of precision search is still dependent on the intelligent use of keywords by the users.

These limitations arise primarily due to the lack of semantics in the processing of STM literature.

The key to knowledge discovery, it therefore appears, is **Semantic Enrichment of STM literature**.

*Semantic enrichment is a process whereby text within a research or scholarly document is annotated by semantic metadata. It enables free text to be converted into a database of knowledge by extracting the concepts and linking the concepts to related knowledge bases.*

Here are a few examples of how some of the leading STM publishers use Semantic Processing to enhance their product offerings.

### Elsevier's New Research Tool, Illumin8

Illumin8, Elsevier's new research tool, combines search and semantic indexing technologies to distil deep meaning, purpose and insights from the company's full-text content, scientific abstracts from 4,000 publishers, patents and billions of web pages. This research tool extracts and analyzes solutions, which are then categorized under organizations, products, technologies, approaches and experts.

Illumin8 is designed to go beyond simple keyword search, quickly finding and extracting crisp summarized answers and interrelationships that are semantically related to the context of the search query.

### **Royal Society of Chemistry's Project Prospect**

RSC Publishing's Project Prospect enhances the chemical information available from the publisher's journal articles. Such articles help researchers gain more information from research papers.

The aim of the project is to make the science within journal articles machine-readable, and to add new ways of retrieving and presenting the information within. Currently, readers using search engines have to rely on text searches to identify either chemicals or subjects of interest. But semantic enrichment allows identification of concepts and chemical substances within the text either by an exact match to the compound, or by a hierarchical classification of subject terms. To create these semantically-enriched papers, RSC editors use text-mining software to annotate compounds, concepts and data within the articles. They then link these to additional electronic resources such as biological databases.

### **Nature Publishing Group's Open Text Mining Interface**

Nature Publishing Group's Open Text Mining Interface (OTMI) aims to enable scholarly publishers, among others, to disclose their full text for indexing and text-mining purposes but without giving it away in a form that is readily human-readable.

OTMI comes in an XML format that expresses the full text via word vectors, plus "snippets" (an alphabetically ordered sentence list) for programs to do full-text search against. This way, programs can data-mine the full text of the article, but a human cannot "read" it sequentially.

## In Conclusion

Researchers are trying to cope with the avalanche of information by applying digital technologies and new strategies for knowledge discovery in ways that require increasingly fine-grained access to information in scientific articles.

With more content and less time to analyze it, users need content that reaches out in an actionable ways for decision making.

STM publishers are increasingly looking for ways to incorporate content and tools, to offer slices of content to answer questions, or to add analysis and structure to content in ways that make it useful in very specific situations. This will create a scenario where STM publishing is to be gradually transformed from publisher-centric content offerings to user-centric knowledge management initiatives.

Semantic Enrichment of STM literature provides a pivotal role in such knowledge discovery.

Semantic enrichment is still at a nascent stage, but poised to take the users of STM information to the next level of knowledge discovery. With the STM information space being further cluttered by the emergence of blogs and wikis, mining Knowledge has become more daunting, providing many challenges to both the disseminators and users of knowledge. This also creates an opportunity for the publishers to carve a niche for themselves by conceptualizing and designing many new and interesting information products and services.

This white paper is presented by Scope e-Knowledge Center - Your Trusted Partner for Knowledge Processing.

### About Scope e-Knowledge

Scope e-Knowledge Center (P) Ltd ([www.scopeknowledge.com](http://www.scopeknowledge.com)) is a leading player in the knowledge processing industry. Scope has been operating in the KPO industry for over 21 years and in the Publishing industry for over eight years, with global blue chip clients.

Scope provides its services to the high-end of the knowledge processing sector, and mainly use semantic enrichment as the key process to offer domain specific and value added abstracting and indexing services, taxonomy and ontology services and other content enrichment offerings.

Scope has been listed in *KM World* magazine's "Top 100 companies that matter in the Knowledge Management space".

Scope has a strong domain base in STM, comprising Engineers, Doctors, Life Science and Pharmaceutical Professionals.

Scope's multilingual capabilities include processing information in English, French, German, Chinese, Korean and Japanese. We are also developing capabilities in other languages including Spanish, Portuguese, Dutch, Italian, etc.

Scope was certified as an ISO 9001: 2000 company by DNV, Netherlands and successfully transitioned to ISO 27001(Information Security Management System).

### Contact us:

#### Headquarter

**R.Sivadas**  
CEO  
Tel:+91 44 24314201  
[siv@scopeknowledge.com](mailto:siv@scopeknowledge.com)

**Anindya Panda**  
Asst. Business Development  
Manager  
Tel:+91 44 24314201  
(M) +91 98846 09689  
[anindypanda@scopeknowledge.com](mailto:anindypanda@scopeknowledge.com)

#### USA

**Frank Stumpf**  
Board Advisor,  
Tel: +1 804 6905399  
[fstumpf@scopeknowledge.com](mailto:fstumpf@scopeknowledge.com)

**Ken Bozler**  
Senior Vice President of Sales  
(M) +1 813.892.4752  
[kbozler@scopeknowledge.com](mailto:kbozler@scopeknowledge.com)

#### UK

**Hector Bolaños**  
General Manager  
Tel: +44 20 7096 0493  
(M): +44 7880 557400  
[hector@scopeknowledge.com](mailto:hector@scopeknowledge.com)

#### Netherlands

**Ms. Priya Sinha**  
Business Development  
Manager  
Tel / Fax: +31 (0) 35  
7511 000  
[priya@scopeknowledge.com](mailto:priya@scopeknowledge.com)

## Disclaimer

This material is based upon information that we consider reliable, but we do not represent that it is accurate or complete, and it should be relied upon as such. Neither **SCOPE e-KNOWLEDGE CENTER PRIVATE LIMITED** nor any person connected with it accepts any liability arising from the use of this document. Investors should rely on their own investigations and take their own professional advice. Opinions expressed are our current opinions as of the date appearing on this material only. While we endeavor to update on a reasonable basis the information discussed in this material, there may be regulatory, compliance, or other reasons that prevent us from doing so.

No part of this material may be duplicated in any form and / or redistributed without the prior written consent of **SCOPE e-KNOWLEDGE CENTER PRIVATE LIMITED™**