

## Textual Data Mining

Textual Data Mining (TDM) is a process by which large quantities of unstructured textual records are efficiently queried to retrieve relevant information on a topic of interest. Using a set of scientific and technical data sources, a corpus of all articles, authors and research organizations relevant to the topic of interest is developed. The TDM process includes software tools that allow a research analyst to display a taxonomy that clusters the records by topical similarity and allows the analyst to visualize patterns in the database that would not otherwise be apparent. The software tools also have a robust search capability that allows the analyst to quickly and efficiently search databases for additional information including information on prolific research authors and activities working in the technology area of interest.

### Our Customers

#### Office of Naval Research (ONR)

- Office of the Director
- TechSolutions Program

#### Department of Homeland Security

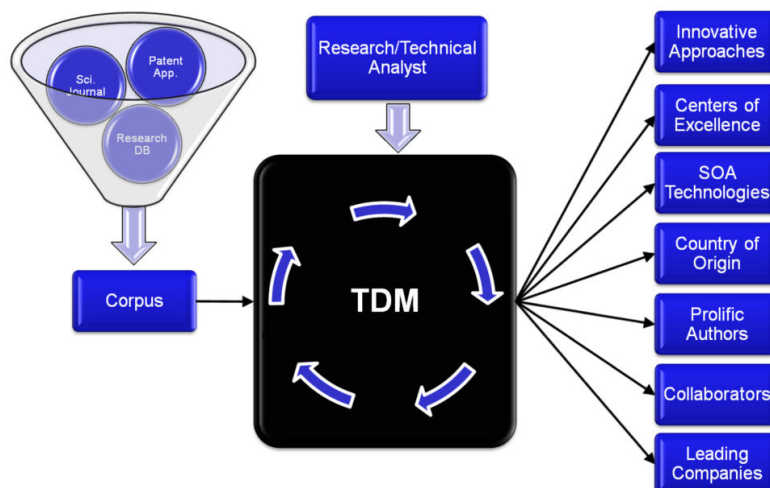
- Science & Technology Directorate

#### Department of Energy

- National Energy Technology Laboratory (NETL)

### Benefits and Applications

- Conduct technology assessments that result in actionable intelligence
- Quickly and accurately determine what research is being conducted across the globe within a technology/topic of interest and identify centers of excellence for such research
- Identify individual researchers within a specific field of interest for participation at technical reviews and workshops
- Assess and/or compare national Science & Technology investment trends for a specific country
- Deliver RFQ's and RFP's to a targeted and comprehensive list of companies or researchers working in a particular field
- Identify innovative and non-conventional solutions to specific problems



(continued)



### The TDM Process

DDL OMNI has established procedures for developing comprehensive queries to efficiently search large commercial, public and controlled access government databases for potentially relevant information on a topic of interest. A corpus of this potentially relevant information is developed, filtered and denoised using a proprietary tool, and record metrics are statistically evaluated to ascertain patterns in the data.

Based on these record metrics, clustering factors are determined using a number of attributes pertaining to the contextual usage of key phrases related to the topic of interest. These clusters enable thematic groupings of highly relevant information at any level of detail desired. (Strategic-level detail can be used to identify general technical approaches to a problem, while a more tactical focus can be used to locate specific prolific researchers in a particular area of interest.)

Our research analysts and project engineers are well versed in engineering and scientific practices and are able to augment the document clustering analysis through an iterative review cycle. This review cycle allows the development of a final report that provides concise, actionable information compiled from potentially millions of source documents.

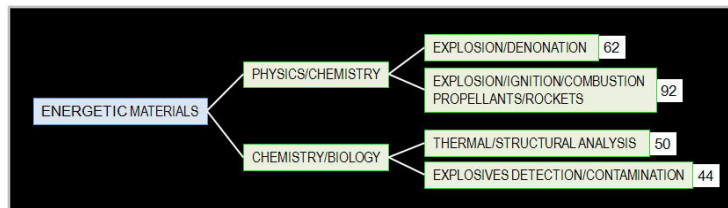
### Data Sources

The TDM Corpus is compiled from a number of data sources and can be extended to include in-house or restricted access databases. Among the standard data sources used are—

- US Patent & Trademark Office
- NIH Pubmed
- NSF Awards
- DOE Energy Citations
- DTIC for Government Contracts
- NASA Technical Reports Server
- White Paper Indices
- United Nations Bibliographic Information System
- World Intellectual Property Organization Patent Database
- Engineering Village
- Science Citations Index

### Document Clustering

Document Clustering in combination with graphical visualization is a new and unique tool that allows the user to easily maneuver within large and complex unstructured textual data. The data is arranged in a hierarchical tree that allows the users to view the thematic groupings in any level of detail they choose.



**Contact**  
 tdm@ddlomni.com