Semantic Discovery in Biomedical Research

Grants, Patents, and Publications

Chris Asakiewicz, PhD.

Affiliate Professor of Information Management,
Wesley J. Howe School of Technology Management, Stevens Institute of Technology
Castle Point on Hudson, Hoboken, New Jersey 07030

Introduction

Within biospecimen sciences, semantic discovery tools could be used to mine knowledge [2] (both structured and unstructured), collaborate [3] (both formally and informally), and discover [1][4] (both within and across a research space) – directly impacting Pharmaceutical Companies, Biomedical Organizations, Medical Centers and Research Hospitals, as well as Healthcare Organizations.

The first step toward mining knowledge across these stakeholder boundaries is to leverage semantically aware tools to increase overall understanding of what is being done in an area of research and who is doing it – knowledge which can be surfaced by mining grant, publication, and patent data.

Most government agencies have developed portals to assist researchers with access to tools and information to facilitate their grant development and submission activities. However, since grants may be available from a host of different sources, some governmental, some philanthropic, some institutional getting access to the right information and adhering to the correct application process may be difficult.

Semantic discovery tools could be used to mine this information on a global scale, with a key focus on highlighting what type of research is or isn't being done, who's doing it, what kind of results are they seeing or not seeing. Tools which could help organizations in shaping their research proposal process as well as enhancing their research portfolio development process.

Methodology

For the purposes of this pilot grants.gov was used as the primary knowledge base, although comprehensive, it only covers US government research funding. Publications cover a wide range of potential repositories, since grant funding from the National Institutes of Health (NIH) requires publication in PubMed, it was chosen as the principal knowledge base for the pilot. The patents knowledge base used in the pilot, only contains US filings and only in select areas related to biomedical and scientific research.

The ability to combine all three opens up the opportunity to answer some of the following questions:

- What is the role and success of different funding mechanisms and philosophies in advancing research (as seen in resultant publications and patents) – *what is successful and what isn't*?
- How are resultant publications and patents distributed across authorship and co-authorship networks and how are they correlated with grants funding – *who is doing it*?
- How are patents and other scientific knowledge domains evolving and interacting with each other – *where is it taking place*?

Results

The pilot discussed in this poster used a Latent Semantic Analysis (LSA) search engine. This engine ensured that the semantic search returns relevant information related by concept and not just key word. While semantic search engines are not uncommon, most contain limitations – including lack of transparency and user control – which can ultimately undermine the overall value of results. For example, they typically do not show precisely how search results are generated and the user must simply trust that the right relevance between the original query and the semantic application are, in fact, appropriate to the intent of the searcher.

The semantic search engine used in this pilot, from Pure Discovery, however, overcomes such challenges to accomplish the following:

- **Transparency:** Each query is enhanced by machine intelligence and shown to the user for their complete understanding and engagement.
- **Increased control:** Not only is the semantic search transparent, but users are in control with the ability to add, delete, increase or decrease the importance of all query words (concepts) in a unique visual query interface.
- **Scalability:** Although the knowledge bases of information used in this pilot are in the public domain, the semantic search engine platform can associate semantic searches to virtually any index, whether it resides internally or on the web.

Radial Tree View for Grants.gov

Shown here is a map of funded scientific grants from the grants.gov data base. This two-dimensional visualization comprises 7,870 grants.

The visualization provides a (somewhat) comprehensive view of federally funded biomedical research in the United States, and allows scientists to examine what projects are being funded.

Folder Tree View for Grants.gov

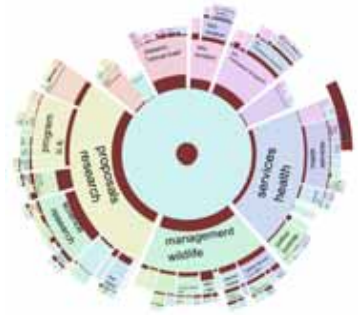
Shown here is a folder view of the grants broken into specific category folders.

Concept Search Criteria and Results

Shown here is a semantic query seeking to highlight research related to "biological specimen banks" or "biobanks".

This particular query highlights all concepts associated with the query term and their weights. Weights can be adjusted altering the results of the overall query. Search results highlight 1333 documents out of a 7870 document knowledge base. Individual documents can be selected and displayed.

Shown here is a semantic query result highlighting research related to "biological specimen banks" or "biobanks".



Conclusions

Scientists and researchers can benefit from a better understanding of the landscape of scientific funding and its correlation with resultant publications and patents. Yet the scale of published literature, patents, and their corresponding funding portfolios are too large to be able to be easily understood. Semantic search and visualization tools can help working scientists and researchers navigate this bewildering landscape.

The first step toward mining knowledge across these stakeholder boundaries is to leverage semantically aware tools to increase overall understanding of what is being done in an area of research and who is doing it – knowledge which can be surfaced by mining grant, publication, and patent data.

One area of significant interest involves investigating how to perform combined analyses over multiple knowledge bases (such as grants, publications, and patents); as well as, performing quantitative historical assessments to evaluate the role and success of previous funding mechanisms and philosophies in advancing research (as seen in resultant publications and patents).

Finally, visualizations facilitate awareness of how knowledge domains change over time. Providing insight into current and past intellectual borders, and ideally, from this developing some understanding of where research and development is heading.

Literature cited

1. Jaffe, Adam and Trajtenberg, Manuel. (2002). **Patents, Citations, and Innovations a Window on the Knowledge Economy**. *MIT Press*. Cambridge, MA.
2. Herr II, Bruce, et.al. (2009). **The NIH Visual Browser: An Interactive Visualization of Biomedical Research.** *IEEE Computer Society.* 13th International Conference on Information Visualization.
3. Neirynack, Thomas and Borner, Katy. (2007). **Representing, Analyzing, and Visualizing Scholarly Data in Support of Research Management**. *IEEE Computer Society.* 11th International Conference on Information Visualization.
4. Kutz, Daniel. (2004). **Examining the Evolution and Distribution of Patent Classifications.** *IEEE Computer Society.* 8th International Conference on Information Visualization.

For further information

Dr. Chris Asakiewicz
IT Strategy & Management Consulting, LLC
55 Van Doren Avenue
Chatham, New Jersey 07928

Phone: (973) 255-9063
Email: Chris.Asakiewicz@gmail.com
Web: www.drchrisasakiewicz.com