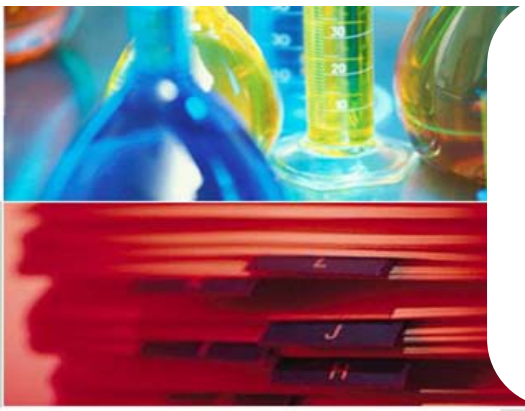

Project Catapult: Data System for Evaluation and Visualization of Relationships between Technologies

Final Report

August 2008

EB Reference Number:
06-526-OD-OIR-OTT

Laurel L. Haak, PhD
Discovery Logic, Inc.
1375 Piccard Dr., Suite 325
Rockville, MD 20850

Table of Contents

Project Catapult: Data System for Evaluation and Visualization of Relationships between Technologies

Abstract

The NIH Office of Technology Transfer (OTT) evaluates, protects, markets, licenses, monitors, and manages the wide range of NIH and Food and Drug Administration (FDA) inventions. The sheer volume of inventions, combined with the interdisciplinary nature of work at the NIH and FDA make it difficult for the technology licensing specialists at OTT to extract the disparate information residing in OTT's proprietary information tracking system (TechTracS®) or to evaluate its potential. Synapse℠, the data system underlying Project Catapult, was created to research and visualize relationships between technologies, patents, and the marketplace. The Project Catapult prototype was released in August 2007, and integrated with TechTracS® in July 2008. Synapse/Catapult provides the means for OTT to push NIH and FDA technologies by searching for potential commercial matches, and to pull the needs of the marketplace by matching commercial needs to available NIH and FDA technologies. Synapse/Catapult relationship mining will also be important for understanding how technologies can be bundled in a context-driven manner.

Background

OTT sought support from the Office of Evaluation for the design and development of Project Catapult, to further its mission to disseminate information and efficiently market NIH technologies. Project Catapult is the next phase of Synapse℠, a system developed by Discovery Logic for evaluating, analyzing, synthesizing, and visualizing the intellectual property portfolio of the National Institutes of Health (NIH) and the Food and Drug Administration (FDA) intramural research programs. For Project Catapult, Discovery Logic was requested to customize currently available text mining and visualization tools to create a system to identify relationships between and among the thousands of technologies within the NIH and FDA portfolios, as well as relationships between those technologies and other NIH and external sources.

Exploring relationships among these diverse sources is vital if NIH is to extract and leverage the knowledge necessary to advance technology transfer at OTT and to accelerate the translation of basic discovery to medical innovation. Much of the scientific community's explicit information is found in unstructured text documents. The inability to integrate information from a multitude of information sources is a serious problem not only for OTT, but universally across NIH. One approach is to couple data visualization tools with text mining tools. Text mining uses sophisticated algorithms to extract relevant data from unstructured text. The technology relies on finding patterns, not single facts. The mined data then must be synthesized so that the results of the data analysis can be comprehended. A significant challenge is to find ways to effectively integrate data between sources to support the visualization of relationships revealed by text mining.

Results: Features of the Data System

Data

Technology databases

The NIH and FDA Intramural Research Program (IRP) Intellectual Property Portfolios – an OTT internal database containing information about intramural inventions, managed using OTT's proprietary version of the TechTracS® recordkeeping software platform. Synapse provides two views to the IRP portfolios: a full view of all listings, and a restricted view of disclosable technologies. The full view is drawn from the technology tables in the TechTracS® database. The restricted view was created as part of this evaluation: a separate set of interrelated tables were created in TechTracS® to list disclosable technologies. For more on methods, see *Integration with TechTracS®* below.

Non-profit Technologies – disclosable technologies provided by universities and research institutes. Currently includes a feed from the Massachusetts Technology Portal[1] and from Yeda Research and Development Company, representing technologies from the Weizmann Institute of Science.

Complementary sources

CRISP (Computer Retrieval of Information on Scientific Projects) –federally funded biomedical research projects, for 1946 to present. Includes projects funded by NIH, SAMHSA, HRSA, FDA, CDCP, AHRQ, and OASH.

RaDiUS (Research and Development in the U.S.) –federal research and development contracts and grants, compiled and maintained by RAND, for 1994-2006.[2]

USPTO (U.S. Patent and Trademark Office) – issued patents and patent applications in the Chemistry subclass, for 2001 to present.

Medline –biomedical journal articles, for 1865 to present. Includes listings of funding sources that can be linked to CRISP entries

Integration with TechTracS®

To improve the ability to track the list of technologies available for licensing, TechTracS® was modified to add tables and screens to allow for data entry of this information. Prior, Synapse pulled technology abstract data from TechTracS® but relied on episodic manual refreshes of newly licensable technologies from the OTT Web site. Additionally, technology updates were manually entered into the Web site database. Substantial effort was required to synchronize information and versions across databases. By pulling these data into one location, Discovery Logic has been able to create a single system of record

[1] The Portal lists technologies from Beth Israel Deaconess Medical Center, Boston Biomedical Research Institute, Boston University, Brandeis University, Caritas St. Elizabeth's Medical Center, Boston Children's Hospital, Dana Farber Cancer Institute, Harvard University, Joslin Diabetes Center, Massachusetts Eye and Ear Infirmary, Massachusetts General Hospital, McLean Hospital, Northeastern University, The Schepens Eye Research Institute, and Tufts University.
[2] RAND discontinued updates to this database in 2007.

for disclosable technologies, namely TechTracS®, a clear benefit to the management of licenses and marketing of technologies.

After TechTracS® was modified, Discovery Logic implemented an export routine to move licensable technologies to a central database, which supports the OTT Web site. In addition, Discovery Logic modified an existing extract, transform, load (ETL) module that pulls data from the TechTracS® database using an ODBC connection to include the new licensable technologies data and load them into Synapse/Catapult. This module runs nightly.

Multi-database search

Synapse provides the user the ability to search each data source by field, keyword, or by mining text for statistical similarities with the search string. Synapse/Catapult includes a multi-database search feature, which allows a user to perform a federated search across all or a selection of databases. The results are returned as ranked lists per database. Users can continue to drill into the data by performing additional searches on the results list.

Visualizations

Synapse provides both NIH and its private sector partners with a tool capable of integrating and analyzing vast amounts of data, displayed using a grid-based user interface. While lists of information can be useful, they do not provide the user a satisfactory appreciation of linkages between data sources or even data from the same source. One of the primary objectives of Project Catapult was to create a visualization system directed towards identification of relationships.

Discovery Logic implemented a first phase of visualization features: mapping and time-series plots. *Mapping* provides the user an appreciation of geographic location of publications, patent assignees, inventors, or grant sponsoring organizations (see **Figure 1**). Users can drill into the result by clicking on a state or individual data point to learn more.

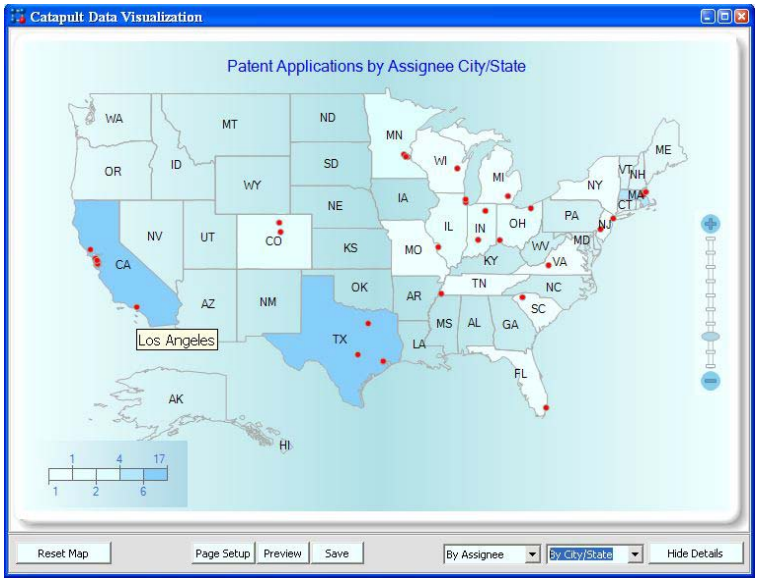

Figure 1: Mapping of patent applications by assignee city/state. Each red dot represents the geographic location of a patent application defined by the search query term(s) against a specific database. The 'tool tips' feature in the system identifies the city (in this case, Los Angeles). The drop-down menu (lower right) optionally allows a user to view search results by zip code.

Time series plots provide a means to display rank of the search result over time. **Figure 2** displays a plot of technology results by similarity to search string and by year. This visualization can provide insight into when a research or patent area started developing, indicated by more dots per year with higher match scores. A user can drill into the data by drawing a box around a set of papers or by clicking on an individual paper for more information.

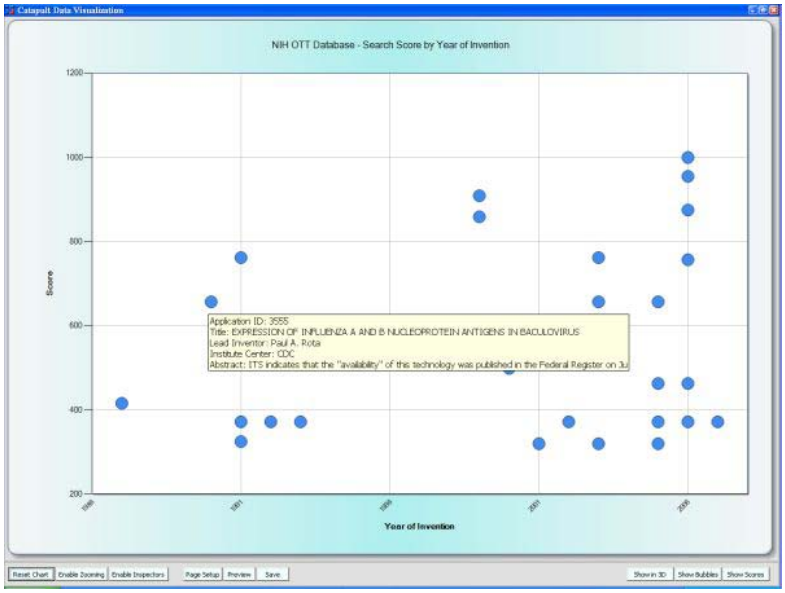

Figure 2: A scatter plot representation of retrieved query hit scores vs. year. This particular example shows NIH/FDA technologies related to the term "influenza". Each data point represents a matching technology; while 'tool tips' offer summary information for initial inspection. Selection of a data point or a region containing multiple data points provides detailed information.

Findings

Project Catapult supported the development of substantial enhancements to the Synapse system. The capabilities of the prototype tool to analyze and integrate data are significantly greater than at the onset of the project. Among system capabilities of note are: (i) multi-database searching, (ii) visualization tools, (iii) addition of non-profit data, and (iv) integration with TechTracS®. The ability to search across multiple databases greatly enhances the functionality of the system. In addition, new search/sort fields are now available for RaDiUS (Funding Agency, Bureau, Program, and Project) and CRISP (PI). Available mapping and charting features offer the ability to graphically represent retrieved query results. Catapult databases have been geocoded, and a map feature added so that a user can view query results by city and state. Such charting features form the basis for visual identification of patterns and trends in technology development or research area focus.

Synapse was designed to analyze large amounts of data very rapidly. Ultimately, this may speed the translation from research to product. In addition to its primary utility in searching filed and issued patents to find those related to NIH and FDA technologies, through the multi-database search, Synapse/Catapult supports exploration of potential relationships between NIH-funded research and NIH intramural inventions and identification of innovation hot spots at NIH. It can also be used to find similar work being done at non-profit research centers.

Conclusions and Recommendations

The strength of Synapse/Catapult is its ability to integrate information from biomedical databases and extract meaningful information for the user. Project Catapult should make it possible for OTT to reach a wider business market by quickly and reliably matching its intellectual property portfolio to the research interests of biotechnology and pharmaceutical companies. OTT is able to proactively market and provide individualized and targeted information to companies that were unaware of the scientific possibilities available to them at NIH and the FDA.

Challenges remain largely due to the quality of data in the databases—not only are data missing, there are issues of cross-walk (e.g., fields have different meanings between and within databases). In addition, there are incomplete and/or inaccurate data within these repositories. For example, while some databases store city and state information in discrete fields, Medline stores addresses in a single free-text field, making it necessary to run an extraction program to populate discrete city and state fields. Nevertheless, the system should be of substantial use to OTT and NIH Institutes.

To reach the next level, Synapse/Catapult needs network charting capabilities. A force-directed placement layout of relationships between technologies based on keywords, inventor, assignee, or related patents would provide a solid use case to develop toward. A second step would provide the user a means to interactively explore these relationships.