

The Role of Knowledge Modeling In Drug Discovery Research

Edy S. Liongosari Anatole V. Gershman Mitu Singh

Accenture Technology Labs
161 N. Clark Street
Chicago, Illinois 60601
USA

E-mail: edy.s.liongosari@accenture.com

Tel: +1-312-693-6744

Fax: +1-312-652-6744

Keywords: knowledge integration, modeling, visualization, collaboration

Abstract

In this paper, we describe a knowledge platform to support drug discovery research in pharmaceutical companies. The platform uses a *knowledge model* to harvest and organize existing internal and external biomedical repositories. By providing a semantically integrated knowledge space, the platform allows the researchers to easily explore various aspects of biological data that originate from multiple disparate sources. This allows them to see the value of the platform right from the start. As a learning platform, it provides facilities for the researchers to get up-to-speed on an unfamiliar topic quickly by assembling relevant knowledge scattered across the organizations. By monitoring usage patterns, the platform can intelligently identify collaboration opportunities among the researchers. The collaboration results, any informal remarks and ideas, can be shared further through annotations. Through the use of domain-specific functional rules, the collaboration results and annotations can be analyzed to study the formation and social network of communities of practice.

1 Introduction

In the last two years, we have interviewed eighteen drug discovery researchers from several pharmaceutical companies and research institutes in Europe and North America to better understand their research tasks. These researchers are responsible for identifying new chemical compounds that have therapeutic purposes. We found that they face numerous knowledge management-related challenges in performing their daily tasks. For instance, they have to constantly deal with many disparate knowledge repositories. A simple question such as “Is there any internal expert on T4 Polynucleotide Kinase?” can pose significant challenges as no single repository may contain the answer. The answer may have to be constructed by examining the authors of various articles, reports and patent documents that maybe indirectly related to this particular kinase, such as through its proteins or biological pathways. They also have to consider the fact that the kinase may have been renamed over the years.

Another key challenge faced by some of the researchers relates to the fact that they sometimes have to provide a point-of-view on an unfamiliar topic in a matter of hours. They have to quickly find any useful references, articles pertinent to the topic, and more importantly, internal researchers who have direct or indirect familiarity with the topic. Other challenges include formulating keywords from a broad set of questions, dealing with the large search results and keeping up with their fields.

Because many of these researchers have little computer and information retrieval skills, they find systems with direct exposure to query language unattractive. Form-based user interface is discouraged as different users with diverse backgrounds will fill the form differently (Elmasri, Navathe 00). Some researchers resort to using librarians. However, this introduces other challenges such as the difficulties in articulating what they are seeking and awareness of available materials and reacting to the materials accordingly.

Many pharmaceutical companies have begun addressing some of the above issues by providing self-serve portals (Ho, Tang 01). They started by unifying the most straightforward databases, providing some query tools, building global taxonomy, and installing enterprise-wide text search capabilities (Edge 03).

In this paper, we introduce our approach to addressing these issues and a knowledge platform that embodies our answers. Our approach is based on the creation of a semantic index to knowledge contained in the underlying heterogeneous repositories. The basis for this index is the *knowledge model* that relates all major concepts in the domain. This model can be tailored further to support a specific task such as providing a quick landscape view of a particular topic. Domain and application-specific rules utilize this index to infer relationships of potential value to researchers and knowledge management professionals. The platform can be used in a variety of contexts. As an illustration, a visualization tool was developed on top of the platform to enable intelligent browsing to support research and investigation tasks by providing the ability to uncover indirect linkages among pieces of knowledge.

2 Related Work

Since there are many point solutions in this space, providing a general literature survey about them is beyond the scope of this paper. However, here are some of the more common ones. Systems like GeneLynx (Lambrix, Jakoniene 03), DiscoveryLink (Haas et al. 02) and SRS (Etzold et al. 96) for example, provide a uniform access to the disparate sources by integrating the underlying schemas of the sources into a global schema and providing a query language to query the schema. Few of them deal with the nomenclature issues and the fact that researchers may pose questions at a different level of abstraction than what is available in the underlying sources (Geffner et al. 99). Tacit's KnowledgeMail (Tacit 04), Lotus Discovery Server (Lotus 04a) and other similar tools (see Mattox et al. 99) (Ackerman, McDonald 98) are designed specifically to locate experts by mining e-mail messages or other existing repositories.

Groove (Groove 04), Lotus Sametime (IBM 04a), and eRoom (Documentum 04) are examples of group collaboration tools. Few have any knowledge about the information they are sharing, who they are sharing it with and how the collaboration results can be further reused and shared. More importantly, many of them have no specific support for identifying collaboration opportunities.

3 Knowledge Modeling

Central to our approach is a technique for modeling the kind of knowledge a researcher needs to perform his or her task (Brody et al. 99). Similar to a *learning net* (Wessner et al. 01), this model is a representation of how the researchers think about the knowledge they need. The model shown in Figure 1 contains bio-medical entities and relationships that depict, among other things, treatments for a disease, how those treatments are related to set of drugs, the chemical compounds that comprise those drugs, and how the compounds are related to various target proteins.

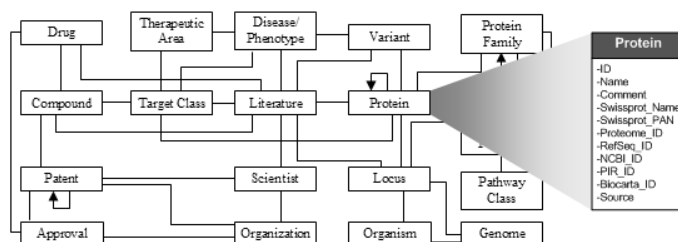


Figure 1: A knowledge model to support drug discovery researchers

Figure 2 depicts one use of the knowledge model. As part of their research activities, the researchers make some of their research work available externally and internally. This explicit knowledge stored in a wide variety of internal and external repositories is extracted, restructured and linked based on the knowledge model. This instantiated model can be used in several ways. For example, it allows the researchers to browse the entire body of knowledge as one homogeneous space of related entities.

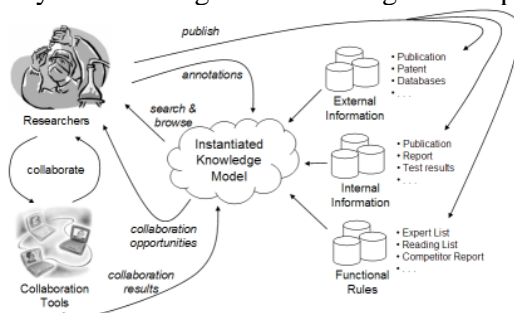


Figure 2: Using Knowledge Model to Support Research Activities

By adding domain-specific functional rules such as “Any person who has authored more than X documents on a protein is an expert in that protein”, it can be used to create a plethora of new knowledge that was not easily available before. Such rules can also be used to identify collaboration opportunities by matching people with similar profiles. The results of the collaboration can be captured informally through the use of annotation. Thus, the instantiated knowledge model becomes a highly structured organizational memory (Ackerman 98) complete with its own life-cycle.

In addition, when used with classification hierarchies, the model becomes a powerful abstraction mechanism (Geffner et al. 99). For example, when a researcher inquires about the role of G-protein in Central Nervous System (CNS) diseases, a straight text search might not reveal anything as the underlying sources may not contain the phrase “Central Nervous System” or its synonyms. However, these sources may contain references to the relationship of G-protein and Parkinson’s or

Alzheimer's disease. To answer the question, the system must utilize a disease classification hierarchy such as the one in Medical Subject Headings (NLM 03) which classifies Parkinson's and Alzheimer's as sub-diseases of CNS. As a result, the system can exploit these linkages and provide knowledge about the relationship between G-protein and higher level diseases like CNS.

4 Usage Scenario

The instantiated knowledge model is essentially a giant semantic index into the underlying sources and it can be used as a knowledge and learning platform to support a variety of applications. One such application is the Knowledge Discovery Tool (KDT) - a web-based intelligent browser developed as a Java applet that uses our semantic index and the model's inference rules to select and present the most likely relationships among the data of interest to the user.

To demonstrate some of the KDT features and the capabilities of the platform, consider the following example interaction: A drug discovery researcher who specializes in Oncology is researching a potential link between Parkinson's disease and cancer. She needs to bring herself up-to-speed about Parkinson's disease in a short time. She needs to identify experts in the area and any relevant past work.

She starts up KDT and searches for "Parkinson's" in the Disease field. KDT then displays all diseases whose names include "Parkinson's". She clicks on "Parkinson's disease" and is brought to the Landscape View for that disease as shown in Figure 3.

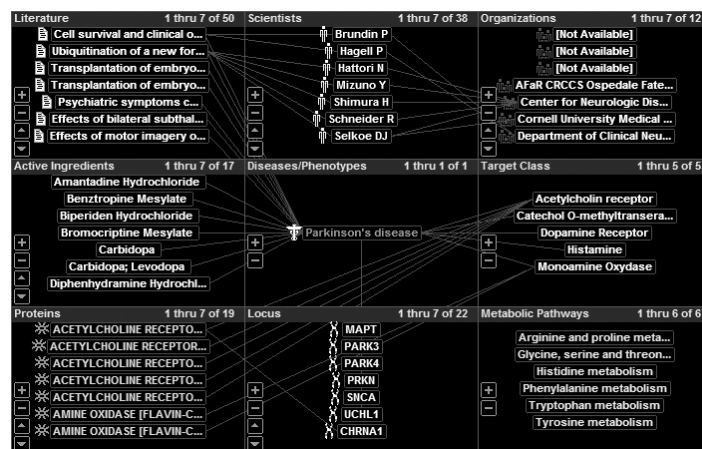


Figure 3: The Landscape View of "Parkinson's disease"

The view divides the screen into 9 (3x3) panes. The center pane shows her current focus (Parkinson's Disease). Each outer pane displays information related to the current focus as filtered by the knowledge model. She sees recent literature related to the disease in the top left pane, a list of experts on the disease in the top center pane, and a list of organizations that have published articles or own patents related to the disease in the top right pane. The center row contains a list of related chemical compounds; the disease itself; and a list of biological target classes used for treatments of the disease.

The researcher can see linkages among the items in the panes indicated by dark grey lines that are visible without cluttering the screen. The lines let users visually maintain the connectivity among entities, which seems to be intuitive to many users. Studies

have shown that exposing a large number of relationships stimulates fresh thoughts and breaks through creative blocks (Schneiderman 00).

The view facilitates fast browsing by minimizing the need for mouse clicking. For example, if the user hovers the mouse over an item, a tool-tip will pop up displaying the item's full title and attributes and it will also highlight its links. Since each item represents an index to the underlying source, she can view the source by double clicking on the item. A new window showing the item's sources will appear.

Single clicking on an item will re-orient the knowledge model, move the item to the center pane, and refresh the contents of the outer panes accordingly. Figure 4a shows how the view looks like after the user has clicked on PRKN - the third gene (locus) in the bottom middle pane of Figure 3.

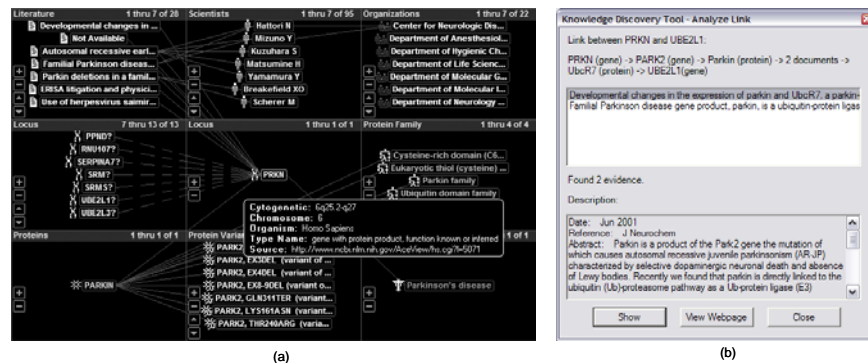


Figure 4: (a) Landscape view of gene “PRKN” (b) A window showing a path that substantiates the link between PRKN and UBE2L1

Each outer pane of the browser presents the results of a query ran against the underlying semantic index and filtered and prioritized according to the rules in the knowledge model. The user can easily customize the view in each pane by picking from a large library of predefined views. For example, the user can change the list of disease expert shown in top center pane of Figure 4a to show those people that have been published about PRKN in the past 3 years and sort the list based on their most recent publication date. A user who is an oncologist can tailor a pane to display only genes that related to PRKN and cancer-related diseases. The left center pane of figure 4a is such a pane.

The platform provides KDT with a powerful knowledge analysis capability to make it more intelligent. For example, one of the items in the left center pane UBE2L1 is a gene that has been associated with several forms of cancer including Leukemia and Breast Cancer (Ardley et al. 99). The link between PRKN and UBE2L1 is drawn as a dashed line indicating there are multiple degrees of separation between them. By clicking on the right menu over the line, the user can query the tool to show how this link is derived. The tool found two paths that substantiate the link. Figure 4b shows one of them: a path with six items as follows. Gene PRKN has been renamed to PARK2 and this gene produces the Parkin protein. Parkin is linked to another protein UbcR7 through two articles. The bottom half of figure 4b shows one of the articles that describes an interaction between Parkin and UbcR7 in rat brain (Wang et al. 01). UbcR7 in turn is produced by UBE2L1. Deriving links in this manner to expose hidden or indirect links that the users might not have thought of is a powerful capability (Schneiderman 00).

This platform can also be used as a foundation to foster collaboration among its users. As the users navigate through the web of knowledge available to them, the

platform can be used to monitor and match the users' navigation paths. Because people with similar navigation paths presumably have similar interests, this technique can be used to identify collaboration opportunities. Another way to identify collaboration opportunities is to monitor people with similar set of bookmarks.

The items in the navigation window can be linked to and from external collaboration tools such as Groove or eRoom. Through a custom interface, the platform can automatically capture the names of the participants from those tools, identify the topics of discussions and link them back to the index. Furthermore, the results of the collaboration can be posted back to the index by depositing them to the appropriate knowledge repository or by adding annotations to the index.

Some other capabilities of the platform include the ability to monitor new items and relationships around a topic of interest or to automate the repetitive tasks by developing *wizards* similar to those used in setting up in printers in Microsoft Windows operating systems.

5 Summary

As described in section 4 above, KDT uses the instantiated knowledge model as a platform for the researchers to:

- Quickly get up-to-speed on a topic by providing them a landscape view, key references and other relevant materials including a list of experts
- Allow them to easily explore thematically related information across multiple data sources as if it were a single richly connected knowledge space
- Assist them in creating hypothesis and arriving at the right questions by exposing indirect or hidden relationships among pieces of information
- Reduce the amount of documents the researchers have to read by summarizing the contents for them
- Allow them to easily share ideas by annotating the items and links in the index
- Identify collaboration opportunities by matching its users' profiles and usage
- Capture collaboration results so they can be shared further

The knowledge model transforms the contents of existing knowledge repositories into knowledge that researchers can readily use without much training. Because it behaves as an add-on component to the existing systems, it provides additional benefits without placing additional burden on the researchers. This entices the researchers to use it and to contribute further.

As described above, the platform itself can be used to easily build custom knowledge-intensive applications independent of the location, intricacies, access methods and nomenclature of the underlying sources. KDT is just one example. While communities of practice are not the focus of this effort, one can easily see how a knowledge management professional could use this platform to extract information to build social networks or study the formation of a community of practice by monitoring the amount collaboration and annotations that are being shared around a topic by adding application-specific rules.

We are currently in the process of conducting a formal evaluation of the platform effectiveness as compared to the current access methods. We are working with several pharmaceutical companies, the University of Colorado Health Sciences Center in Denver and the Integrative Neuroscience Initiative on Alcoholism – an initiative

sponsored by the National Institute of Alcohol Abuse and Alcoholism – to formally evaluate the effectiveness of platform. We have over 30 people who have signed up for our initial pilot. We plan to complete this process by the end of Spring 2004.

The researchers who have tried KDT have given us very positive feedback and expressed a strong desire to use the tool in their daily activities. In one particularly gratifying case, the senior leader of a heart failure research project tried it during a ten-minute demonstration and discovered a valuable relationship between heart failure and leukemia of which he had not been aware and was essential to his research. In another case, a post-doctoral biologist discovered a new link between Lou Gehrig's disease and genital diseases suggesting that a drug from one disease can be modified for the treatment of the other.

The knowledge modeling and integration concept has been successfully used to support research and investigation tasks in several other domains including tax fraud investigation in government agencies (Accenture 03). It also has been widely deployed internally in Accenture to support, for example, the proposal writing process by identifying key experts, credentials, similar proposals and other relevant materials (Liongosari et al. 99) (Davenport, Hansen 98).

References

- (Accenture 03) Accenture. 2003. Case Study: Ireland's Office of the Revenue Commissioners". http://www.accenture.com/xd/xd.asp?it=enweb&xd=services\technology\case\tech_revenue.xml.
- (Ackerman 98) Ackerman, M. S. 1998. Augmenting the Organization Memory: A Field Study of Answer Garden. *ACM Trans. on Information Systems*. 16 (3). 203-224.
- (Ackerman, McDonald 98) Ackerman, M. A., McDonald, D. W. 1998. Just Talk To Me: A Field Study of Expertise Location. *Proc. Computer Supported Cooperative Work*. 315-324.
- (Ardley et al. 00) Ardley, H. C., Moynihan, T. P., Markham, A. F., Robinson, P. A. (2000) Promoter analysis of the human ubiquitin-conjugating enzyme gene family UBE1LI-4, including UBE2L3 which encodes UbC7. *Biochim Bio-phys Acta*. 1491(1-3). 57-64.
- (Brody et al. 99) Brody, A.B., et al. 1999. Integrating Disparate Knowledge Sources. *Proc. Second Int. Conf. on Practical Application of Knowledge Management*. 77-82.
- (Davenport, Hansen 98) Davenport, T.H., Hansen, M.T. 1998. Knowledge Management at Andersen Consulting. *Harvard Business School Case Study*. N9-499-032.
- (Documentum 04) Documentum. 2004. eRoom Collaboration. Available at <http://www.documentum.com/solutions/collaboration/index.htm>.
- (Edge 03) Edge Science LLC. 2003. Pharmaceutical Knowledge Management Overview. Available at <http://www.edgesci.com/thinking/portalanalysis.pdf>.
- (Elmasri, Navathe 00) Elmasri, R., Navathe. S. 2000. *Fundamentals of Database Systems*. Ch 27.5 Genome Data Management, Addison-Wesley. 3rd ed.
- (Etzold et al. 96) Etzold, T., Ulyanov A., Argos, P. 1996. SRS: information retrieval system for molecular biology data banks. *Methods Enzymol*. 226. 114-128.

- (Geffner et al. 99) Geffner, S., Agrawal, D., Abbadi, T. 1999. Browsing large digital library collections using classification hierarchies. Proc. Eighth Int. Conf. on Information and Knowledge Management. 195-201.
- (Haas et al. 01) Haas, L., et al. 2001. DiscoveryLink: A system for integrated access to life sciences data sources. IBM Systems Journal, 40(2).
- (Ho, Tang 01) Ho, J., Tang, R. 2001. Towards an optimal resolution to information overload: an infomediary approach. Proc. 2001 Int. ACM SIGGROUP Conf. on Supporting Group Work. 91-96.
- (IBM 04a) IBM Corp. 2004. Lotus Discovery Server. Available at <http://www.lotus.com/products/discserver.nsf>.
- (Groove 04) Groove Networks Inc. 2004. Available at <http://www.groove.net>.
- (IBM 04b) IBM Corp. 2004. Lotus Sametime. Available at <http://www.lotus.com/products/lotussametime.nsf/wdocs/homepage>.
- (Lambrix, Jakoniene 03) Lambrix, P., Jakoniene, V. 2003. Towards transparent access to multiple biological databanks. Proc. First Asia-Pacific Bioinformatics Conf. 19. 53-60.
- (Liongosari et al. 99) Liongosari, E. S., Dempski, K. L., Swaminathan, K. S. 1999. In Search of A New Generation of Knowledge Management Applications. ACM SIGGROUP Bulletin. 20 (2). 60-62.
- (Mattox et al. 99) Mattox, D., Maybury, M., and Morey, D. 1999. Enterprise Expert and Knowledge Discovery. Proc. HCI International '99. 303-307.
- (NLM 03) National Library of Medicine. 2003. Medical Subject Headings: Annotated Alphabetic List. Technical report PB2003-964801.
- (Tacit 04) Tacit. 2004. KnowledgeMail. Available at <http://www.tacit.com/products/knowledgemail/>
- (Schneiderman 00) Schneiderman, B. 2000. Creating Creativity: User Interfaces for Supporting Innovation. ACM Trans. On Computer-Human Interaction, 7(1). 114-138.
- (Wang et al. 01) Wang M., et al. 2001. Developmental changes in the expression of parkin and UbcR7, a parkin-interacting and ubiquitin-conjugating enzyme, in rat brain. J. Neurochemistry. 77(6). 1561-1568.
- (Wessner et al. 01) Wessner, M., Torsten, H., Pfister, H. R. 2001. The Learning Net - An Interactive Representation of Shared Knowledge. Proc. ED-MEDIA 2001. 2035-2040.