OREGON HEALTH & SCIENCE UNIVERSITY

SCHOOL OF MEDICINE – GRADUATE STUDIES

Discovering Synergistic Groups of Researchers for Translational Research.

By Nathan Bahr

A THESIS

Presented to Department of Medical Informatics and Clinical Epidemiology

and the Oregon Health & Science University

School of Medicine

in partial fulfillment of

the requirements for the degree of

Master of Science

July 2009

OREGON HEALTH & SCIENCE UNIVERSITY

SCHOOL OF MEDICINE – GRADUATE STUDIES

School of Medicine

Oregon Health & Science University

CERTIFICATE OF APPROVAL

This certifies that the Master's thesis of

Nathan Bahr

has been approved

Mentor/Advisor

Member

Member

Member

Member

Table of Contents

Index of Tables

Illustrations

Acknowledgments

I would like to take this moment to thank my advisor, Aaron Cohen, and committee members Julie Earnest, Eric Orwoll, and Bill Hersh, for the time and input they contributed to this project. I am also grateful to Karen Eden, Poonam Sharma, Jeanne-Marie Guise, and Christian O'Haire for providing valuable usability feedback on my system. Also, to Vetle Torvik and Neil Smalheiser at University of Illinois at Chicago for providing access to their database of disambiguated author publications, Author-ity.

Abstract

In this study, a software application was developed that permitted investigators to browse a visual network of researchers and MeSH terms to help identify partners for future research collaboration. The intent of this was to improve the rate of translational research by suggesting researchers whom the investigator would not have normally considered or find, but shared complementary interests, as indicated by publications or grants.

The system was evaluated through use cases and surveys. The use cases guided the application's development by in that user needs were identified and then incorporated into the system. The main needs were to: have filters which showed only the most prominent suggestions; provide controls to see how a researchers coauthor and MeSH term connections changed over time; and to incorporate grants into the application's database.

The surveys were used to measure the efficacy of the system versus using traditional means, prior knowledge or the Internet. Users would make a list of researchers using the Internet and then add to that list using the application. Then, experts graded the quality of those researchers on novelty, essentialness, and appropriateness. Two surveys were run and the application was able to find an additional 50% (8/15) and 110% (11/10) set of different researchers who were graded to be of similar quality to the researchers found using the Internet.

In a follow up, the users were asked to evaluate using the Internet and the

application for this task. In using the Internet, they found that it had a lot of noise and presented many irrelevant results that had to be investigate manually. In using the application, users found the ability to explore on related MeSH terms helpful as it expanded their search spaced in a focused manner. The application provided a means of identifying researchers beyond using current Internet search tools, because it provides a focused database for that task and a means of expanding the user's search space in a relative manner.

Introduction

The purpose of this study was to discover synergistic groups of investigators for translational research because this type of endeavor demands collaboration across disciplines. Translational research is the process of translating a discovery made in the lab into an intervention which could be applied in clinical care. Translational research was described by Khoury et al.[2] as having four phases where an intervention is discovered and evaluated for efficacy over a series of clinical trials. If it proves useful, guidelines are developed and the intervention is introduced into widespread clinical use. In a provided example, mutations in the BRCA gene were linked to breast cancer, making it a likely candidate for genetic testing. Following the discovery, BRCA1/2 screening was evaluated for the possible harm and benefits it would cause, and guidelines on its usage were developed.

The continuum of translation research in human genetics; types and examples.			
Translation Phase	Notation	Types of Research	Examples
T1 (Bench)	Discovery of candidate health application.	Phase I and II of clinical trials; observational studies.	Is there an association between BRCA mutations and breast cancer?
T2 (Bedside)	Health applications to evidenced based practice guidelines.	Phase III clinical trials; observational studies; diffusion research.	What is the positive predictive value BRCA mutations in at-risk women.
T3 (Community)	Practice guidelines to health practice.	Dissemination, implementation and diffusion research. Phase IV clinical trials.	What proportion of women who meet the family criteria are tested for BRCA and what are the barriers to testing?
T4 (Public Population)	Practice to population health impact.	Outcomes research. Population monitoring of morbidity, mortality, benefits and risks.	Does BRCA testing in asymptomatic women reduce breast cancer incidence or improve outcomes?

Table 1: The four phases of translational research. Adapted from Khoury et al.[2]

For the purposes of this study, synergy was defined to be the set of complementary skills which enable researchers to work together on a translational research projects. Consider the BRCA example given by Khoury et al. In the early stages, geneticists identified the gene as a useful tool to screen for breast cancer. In the later stages, physicians and oncologists were responsible for disseminating the information to the public and developing guidelines. This demonstrates an instance where an intervention required a broad range of expertise to introduce it to a clinical setting and that these involved investigators were able to build off of one another's work due to a common interest and expertise in treating breast cancer. The impetus of translational research is to emulate this process of bringing researchers together to test and validate new discoveries such that they can be used to benefit the general populace.

The challenge with forming translational research teams is that they often require

the collaboration of researchers who are pressured work apart. Scientists are pressured to pursue NIH funding through traditional science and clinicians are expected to spend more time with patients[3]. As a result, members from either group may be constrained in seeking translational research opportunities or unaware of overlapping interests. Therefore, the focus of this study was to help construct teams of compatible researchers for translational research opportunities.

Typically, teams are assembled through acquaintances, knowledge of local researchers, and open invitation. In two research retreats at Oregon Health & Science University (OHSU), the organizers described how they identified and invited potential collaborators:

(Tobacco Retreat)
OCTRI leadership at CHR and OHSU sent out word through each organization for interested parties. Individual responders then suggested others to invite. Not much to it.
- Jeffery Fellows

(CPR Retreat)
The process was relatively informal - Rick Deyo (Program Director), Mark Spofford (Associate Program Director at Kaiser) and I (Program Coordinator) came up with an initial list based on our knowledge of local "T2" investigators. We then brought this list to the rest of the CPR team for advice on additional people who should be included. Additional names were added as CPR team members thought of them, or when investigators contacted us directly asking if they could attend (within certain limits, since we had some space limitations).
- Arwen Bunce

Figure 1: Traditional methods of identifying potential collaborators.

These methods were limited in that some resources were being under-utilized. Many publication and grant databases contain articles published by a given researcher. From here, the retreat organizers could have looked up researchers under the retreat topic to

find investigators who specialize in that area. Beyond that, they could also explore the researcher's coauthors and topics the coauthors have published under. This does not usually extend beyond the first group of authors published under a particular topic because it would be time consuming to explore those possible paths of collaboration. If the search tools supported it, however, potential collaborators could be discovered by mining the literature for researchers with skills that could be used in a research project.

In a preliminary study[4], publication information for OHSU researchers was extracted from MEDLINE publications and used to create researcher profiles. These profiles contained the publication title, coauthors, and Medical Subject Headings (MeSH) for each article a researcher published. Techniques from information retrieval and social network analysis were then used to identify complementary connections between researchers who have never worked together before but could pool their resources to collaborate on a project.

In this preliminary project, complementary connections were defined as two MeSH terms having a high mutual information score. The mutual information score measured how likely it was to see two terms A and B together rather than apart.

$$I(A,B) = \log \frac{P(A,B)}{P(A)P(B)}$$

Figure 2: The mutual information score.

Scientist A: Beer TM & Scientist C: Druker BJ		
Scientist Terms A	Bridging Terms B	Scientist Terms C
Hospice Care	Terminal Care	Retrospective Studies
Pulse Therapy	Drug Metrorrhagia	Middle Aged
Affect	Laboratory Animal Science	Mice
Calcium Channel Agonists	Dihydrotachysterol	Female
Carboplatin	Dimethyldithiocarbamate	Transfection
Oxides	Silver Compounds	Cell Line
Receptors, Calcitriol	Dihydroxycholecalciferols	Exons
Arsenicals	Dimercaprol	Phosphotyrosine

Table 2: Swanson matching between two researchers: Tomasz Beer a physician specializing in prostate cancer and Brian Druker specializing in Leukemia.The evidence shows that they are linked together through the ABC tuple: Receptors, Calcitriol; Dihydroxy-cholecalciferols (Vitamin D); and Exons. Calcitriol receptors are activated by Vitamin D, which has been shown to inhibit prostate cancer.This evidence shows that they could work together on studying the common role of Vitamin D in reducing various types of cancer.

This score was computed over the entire MEDLINE corpus for major MeSH headings. These measures were particularly useful in that they permitted linking topics that were not the same but commonly appeared together. Swanson matching[5] was then used to find indirect but complementary links between researchers based on the topics they published on.

Researchers could also be linked together based on who they worked with, i.e. their coauthors. One possible measure that was investigated was to use Small World Analysis[6] to calculate the average distance from a given author to all other authors in a coauthor network. Distance was defined as the number of acquaintances one would have to go through to meet the final person. For instance, if Bob and Sally coauthored a paper together, their distance would be 1. If Sally coauthored a paper with Tom, Bob's distance

to Tom would be 2, and so on. This measure gave an indication as to how entrenched an investigator was in the OHSU researcher network. A low value indicated that a researcher had a high level of connectivity and could contact any other researchers through a few acquaintances. This high level of connectivity implied more experience because the researcher collaborated with more individuals or a few highly connected individuals. A high value indicated that the researcher had a lower level of connectivity because they had few coauthors. This would imply that the researcher was beginning their publishing career and had less experience. This measure of connectivity was applied to the set of OHSU authors extracted from the MEDLINE data.

Computing these measures revealed interesting features about the OHSU authors. The Swanson matching over MeSH-MeSH mutual information connections provided a proof-of-concept where authors could be connected on related topics. The Small World Analysis[6] presented a coauthor network where most of the researchers were linked together in a single super cluster. It also highlighted the most highly connected researchers, most of whom appeared to be biostatisticians; the implication is that biostatisticians might have skills which make them useful in a variety of projects.

The preliminary study yielded interesting results in that it could highlight experienced researchers and suggest many possible ways of connecting them. It was not, however, clear how this information could be used to automatically construct effective, synergistic teams for translational research. These results were presented to members of the Oregon Clinical Translational Research Institute (OCTRI), supporters of this research, to get a better insight as to its possible applications. It was decided that it would

be useful area of research to develop a tool which could assist investigators in finding the collaborative resources they needed. This tool was specified to be an application where domain experts, retreat organizers, and grant writers could browse an author-mesh graph to discover researchers with the appropriate skills.

Statement of Purpose

The main question of this thesis is:

- *What information constitutes synergistic features between researchers, and does this information help investigators construct better research teams?*

The underlying problem of constructing a translational research team is that there is no clear model of how such a team should be composed. The people who regularly organize multidisciplinary teams, however, do have an intuitive sense of what resources are necessary. In this process, they rely on Internet resources and mostly acquaintances to seek out the right investigators. This study seeks to follow and enhance that process by providing a software application that presents interconnected collections of researchers. The resulting team is deemed better if organizer subjectively feels that they can identify more researchers that have features which fit the criteria of the topic at hand.

The application was reviewed through a series of use cases and comparisons of not using the application versus using the application. The use cases cover actual instances in which domain experts used the application to fill their information needs. This will include a description of the data being sought, the usage of the application, and the results. This will reveal what data they are looking for and the queries they are

attempting to find it. The comparisons will measure the quantitative differences of not using the application versus using the application to test its efficacy. This will include counting the number of additional researchers found and seeing if there are any noticeable differences in the quality of researchers found according to the user's, i.e. domain expert's, judgment.

This study drew together elements in social network analysis, information retrieval, and information visualization to build an application for assisting domain experts in finding collaborative partners. In the following sections, the implementation, evaluation, and results will be discussed in more detail. First, the literature leading up to the application's design will be reviewed. Then, the implementation and the evaluation methods will be described. Finally, the findings will be summarized and the future avenues of research for constructing teams in a principled manner will be discussed.

Literature Review

The intent of the literature review is to draw together disparate/distinct elements in the literature to explore the development of teams in translational research and the need for information tools to support this process. The main challenge of translational research is that it requires a broad breadth of resources and expertise to transform a discovery into a usable intervention which could be used in a clinical setting. Investigators, undertaking such a task, are then required to seek out these resources through personal acquaintances, public invitation, or keyword searches through online databases. This highlights a need on the researcher's part for a tool which suggests or

summarizes what resources are available so they could spend less time searching and acquire them more efficiently.

The following sections will lay the groundwork for developing such a tool. The *Translational Research* section explores the current state of translational research and the need to improve on it. The Prior Work on Collaboration Analysis section describes techniques that were used to identify significant researchers and the synergistic interests exist between them. Supplementing these sections are the Related Topics which discusses technologies related to collaboration analysis: getting clean data and visualizing the data.

Translational Research

Translational research is a complex multidisciplinary process in which biological discoveries are translated into medical interventions; the communication also flows backwards, in that the differences of the drug's action in humans can fuel further research. The interest in translational research has existed since about the 1990s [7-12], in which there was a serious interest to apply some of the newer techniques and technologies only available in labs to improve the care of humans. These interests were marshaled together in 2003, when the National Institutes of Health (NIH) established a 'roadmap'[13] and funding[14] to encourage cross-disciplinary interaction between bench scientists and clinicians. The guiding precepts[13] were to:

1. Explore new pathways to discovery
2. Develop research teams of the future
3. Re-engineer the clinical enterprise

A review of literature by Ioannidis revealed the rate of translation to be very low[15]. Only 5 of 101 papers between 1979 and 1983 had progressed to the translational stage 20 years after discovery. He stated that the basic science approach "made oversimplified assumptions that have not matched the true etiological complexity of most common diseases"[15] and later argued that "multidisciplinary collaboration with focused targets and involving both basic and clinical sciences should be encouraged"[16].

Various articles support this recommendation by discussion some of the impediments to forming a multidisciplinary team. Pober et al[3] observed that some obstacles were: "inadequate financial support, shortage of translational investigators, impediments in the academic culture to collaborate, Academic Medical Centers (AMC)* structural organization often hinders collaboration, regulatory impediments to translation, and absence of mechanisms for facilitation of translational research". Mankoff et al.[17] noted that most clinicians would be overwhelmed by the massive amounts of data while scientists would be too distracted by clinical care. Most scientists or physicians would, therefore, not have the ideal mixture of experience for translating research into practice.

These articles suggest that one important factor for accomplishing translational research is in forming a team of researchers with the right skills. As mentioned in the introduction, the possible, observed ways of identifying collaborators was through acquaintance, public invitation, or searching online databases. People are more capable at

* Academic Medical Centers (AMCs) are institutions composed of medical schools, clinics, libraries, laboratories, and administrative facilities. This unique combination allows them greater collaboration between researchers and clinicians, access to cutting edge technologies, and access to patients[1]. This gives AMCs, more than any other type of institution, a greater ability to perform translational research.

handling acquaintances and invitations while a computer is more apt at processing and displaying large amounts of data. This study, therefore focuses on the latter aspect of managing researcher information to provide suggestions on possible candidates for research projects. The following section reviews the data sources and their possible uses in this study.

Social Network Analysis (SNA)

Figure 3: Examples of social network analysis. The top images show product recommendations from (top)Amazon.com and (bottom)Newegg.com. The middle image shows a researcher's coauthors in a (left) basic network and by (right) location. The bottom image demonstrates a work sharing network, set up by Irvani et al.

Social network analysis provides a means of understanding community behavior by viewing individuals as nodes in a network who are connected by common interests. The particular reason for viewing a community as a network is that the topological features could be summarized through computational measures, e.g. who is the most connected individual, or visualized, allowing for intuitive interpretation. Perhaps the most prevalent use of social networks is in product recommender systems, such as those at Amazon.com[18] or Newegg.com[19]. Such systems form bipartite consumer-product graphs[20] in which consumers are connected to products they've purchased. Products are recommended either by those most similar to the consumer's past purchases or those that are often co-purchased together, as evidenced by other consumer's purchases; the latter is known as collaborative filtering. BiomedExperts[21] developed social network visualizations of biomedical researchers. These visualizations along with their bibliometric services, provided an interface which could be used to find experts on a particular biomedical topic. Iravani et al[22] developed a work sharing network in which telephone operators were connected by a common skill. They measured path lengths between operators to discover possible bottlenecks in skill distributions and to explore alternate training programs. These particular applications use various social network analysis techniques identify complementary relationships between nodes such that they can recommend an appropriate course of action. The following sections describe the computational and visual aspects of social network analysis in more detail.

SNA: Computational Methods

The examples presented in the previous section used a collection of computational tool to fulfill their roles and reveal patterns in the social networks they were designed around. In his work[23,24], Newman demonstrated the basic SNA tools while studying coauthor networks in mathematics, physics, biology, and computer science. He computed the cluster coefficients, shortest path length between authors, average path length, and betweenness, to describe how authors in different disciplines collaborated; the definitions of these methods are provided in Table 3.

Social Network – A graph structure in which the nodes are individuals and the edges are interests or features shared between them.
Cluster Coefficient – The degree to which the neighbors of a given node know one another. If all neighbors know each other this value is 1. If all neighbors do not know one another, this value is 0.
Path Length – The distance from one node to another. For instance, if node A is connected to node B and node B is connected to node C than the distance from A to C would be 2.
Average Distance – The average of all path lengths.
Betweenness – The number of shortest paths between two nodes that pass through a given node.

Table 3: Common techniques in social network analysis.

Some of the observations he made were: that researchers in MEDLINE had low cluster coefficients relative to other literature databases; the average distances were relatively short, ranging from 4 to 7; and the networks revealed a funneling behavior in which many paths passed through a few exceptional scientists. While Newman didn't explicitly identify an application for these methods, he did note[24] that the purpose of his

studies were to "alert other researchers to the presence of a valuable source of network data in bibliographic databases."

It was mentioned before, though, that SNA techniques had practical applications in recommender systems. Huang et als[20] computed average path lengths and average cluster coefficients in consumer product bipartite graphs. He found that these networks had longer path lengths and a greater tendency to cluster relative to randomly generated bipartite graphs. These findings suggested that consumer purchases were not random and that using collaborative filtering, i.e. suggesting what products that similar customers purchased, would increase the success of the recommendation actually being followed.

In another example that was briefly mentioned, Iravani et al[22] computed average path lengths to suggest an optimal training program for telephone operators. The setup was that a company had operators who helped callers with their problems. These operators were trained to have skills in handling a specific set of problems. If an operator was busy, they could forward the caller to another operator who had the appropriate skill and was not busy. This made up the work sharing network in which operators were nodes and common skills were edges between those nodes. In a bad training setup, where most operators were busy, the caller would be forwarded many times until they were serviced. In an optimal training setup, the caller would be forwarded a minimal number of times to be serviced. Thus, shorter average path lengths between operators suggested whether one training program was better than the next.

Though not directly related to SNA, Swanson matching[5] provided another means of identifying complementary items, similar to collaborative filtering. In Arrowsmith, an

application which performed Swanson matching, the user would submit two terms A and C and the system would then construct two sets of words co-occurring with terms A or C. Any term appearing in both sets A and C, it was dubbed a B term and used to show how the two queries might be related. In a well known example, they found, by hand, that fish oil and Raynaud's disease were related to terms for ameliorating symptoms: blood viscosity, platelet aggregation, and vascular reactivity[25]. They later developed the Arrowsmith tool for supporting this kind of discovery[5]. Swanson matching was considered useful for finding relations among individuals that were indirect and not immediately obvious from local observations.

The purpose of these methods were to measure prominent topological features and interactions between members in the network. With respect to constructing translational research teams, these nodes might be experts or principle investigators that many other researchers seek to collaborate with. These individuals are identified through a short path length to other researchers or a large betweenness path count. Interactions identify synergistic relationships, where members form a community because they share common interests or features. This might be represented tire pumps being co-purchased with bikes or a particular statistician who coauthors with a group of clinicians. These features can be revealed through the clustering or cooperative filtering that occurs in the social network.

SNA: Data Visualization

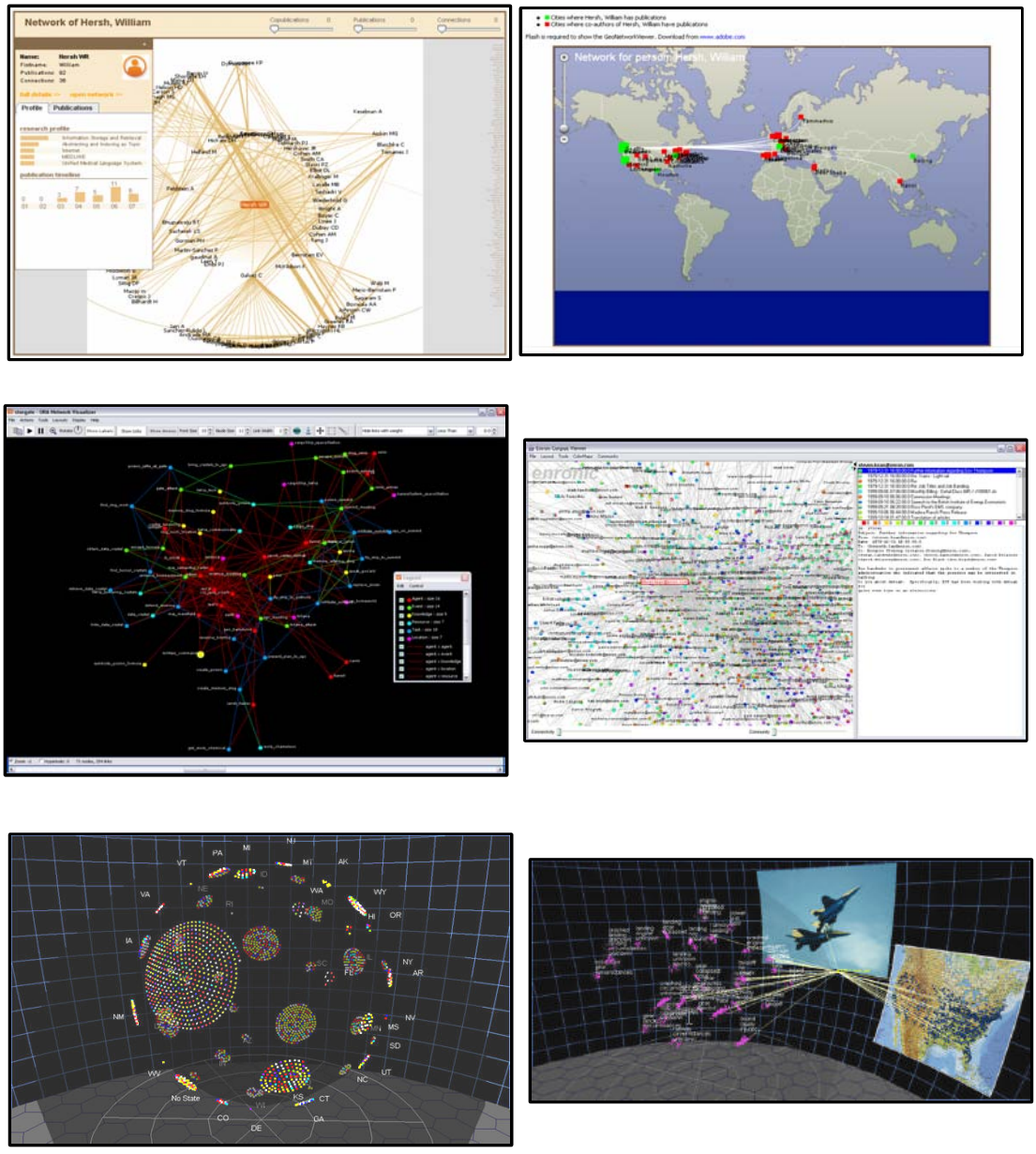

Figure 4: A collection of social network visualization tools. Top left and top right are BiomedExpert's visualization of coauthor networks. Middle left is Carniegie Mellon's Organizational Risk Analyzer. Middle right is a visualization of employees at Enron, built upon the Prefuse framework.Bottom left and middle are images of Starlight, Pacific Northwest National Laboratory's tool. Bottom right is

Another means of studying social networks is to visualize them to provide a more intuitive perspective. The challenge with SNA computations is that their meaning and application may not be readily apparent without seeing the context within which they occur, such as clustering similar individuals or coloring them by type to bring out patterns in the data. Figure 4 presents just a few tools currently available that facilitate exploring visual networks. Typically, these draw these networks according to a particular layout scheme and provide some functionality for browsing the data: moving nodes, panning, zooming, bringing up extra data, etc. This interactivity, coupled with the visual display, would then allow the user to focus on different perspectives of the data and extract meaning from it.

BiomedExperts[21] has a service which permits browsing over biomedical researchers who have a publication indexed in MEDLINE. They provide two visualizations as part of these services to browse a researcher's coauthors. The first, in the top left of Figure 4, an author is presented with all of his or her past coauthors. Edges, or lines, connect all coauthors, whether they be with the original author or not. Clicking on an author re-centers the display on that person and brings up publication statistics in a small side window. The second visualization, in the top middle of Figure 4, shows where an author and the author's coauthors have published, according to the affiliation field in MEDLINE citations. This presentation helps to quickly identify which researchers are local and which are remote.

Carnegie Mellon's Organizational Risk Analyzer[26], shown in the top right Figure 4, was intended to detect risk that personnel may pose to an organization, whether it be

from an individual's removal or malicious intent. In additional to computationally measuring risk, the tool provided a visual interface in which the user could view and manipulate an organization's structure. Changing the social structure would allow the user to observe the results of managerial decisions before actually acting on them.

Pacific Northwest National Laboratory's Starlight system[27] is a generic information system that visualizes relationships between XML objects; these relationships being: similarity, reference, co-occurrence, hierarchical, spatial, or temporal. The bottom left image of Figure 4 shows a 3d clustering of similar items which would allow a user to explore a group of related items. The bottom middle image of Figure 4 shows how the items in a network might map to a geographic location. This system has been applied[28] in the domain of national security to track terrorist attacks and public health to isolate disease outbreaks.

The bottom right image of Figure 4 shows a mapping of the Enron organization, which was displayed using Prefuse[29]. Prefuse is notable in that it is not an application but a programming library that software engineers can use to develop their own visualization tools. The library provides basic functionality in visually laying out network data and allowing that data to be manipulated. Prefuse's programmability gives it more flexibility than most systems in terms of setting up a visual representation of network data.

The computations and tools used in social network analysis were investigated because they provide a means with which to identify synergistic groups of researchers. In constructing a multidisciplinary translational research team, it may not be sufficient to gather researchers similarly specializing in a given topic. Rather, that it may be more

useful to extend the researchers' capabilities by pairing them with those who have complementary skills. Complementarity could be identified as two or more research topics occurring together or coauthors working together. Gathering researchers who have different, but related interests would then expand the abilities of a team and permit them to explore more novel avenues of research.

A combination of computational and visual methods were used to develop an application which could assist team organizers in identifying compatible partners for their projects. The shortest path, publication counts, and coauthor counts were used to identify researchers of greater experience. The method of Swanson matching was also used to list related topics or researchers. This information was then presented as a graphical network which users could interact with and explore. To guide the users in their exploration, nodes and connections with significant scores were visually embellished to help identify them more quickly. This resulted in an application that was intuitive to use and more able to provide novel suggestions.

The following section describes how the application was developed to aid investigators in browsing for complementary researchers. By providing them with more information about their scientific community, this study sought to improve the quality of research teams that were formed.

Methods & Materials

With regards to the thesis objective, an application was developed to help investigators discover prospective research partners through author-topic connections in

the literature. This was a change in direction from the prior work in that an expert driven tool was deemed to be more useful than an automated tool for constructing teams. The rationale behind this was that an automated tool would lose focus the investigators needs for the particular project they were working on. Some topics might be more relevant than others or some research candidates might have varied availability. Instead, an explorable application could help investigators build better teams by broadening the pool of available and relevant researchers.

The application was setup to permit browsing from researcher to topic as presented in the preliminary work. The user would browse a network of author and MeSH nodes, searching along some desired connection for a collaborator. Author-author connections came from co-authorships and represented synergy of working together. Author-MeSH/MeSH-Author came from an author publishing with a MeSH term and indicated knowledge of some topic. MeSH-MeSH came from two terms having a high mutual information score, as calculated over the MEDLINE corpus, and indicated a synergy in topics. The display and exploration of this network was then handled by a Java applet client that was developed iteratively to gradually fit it to the users needs.

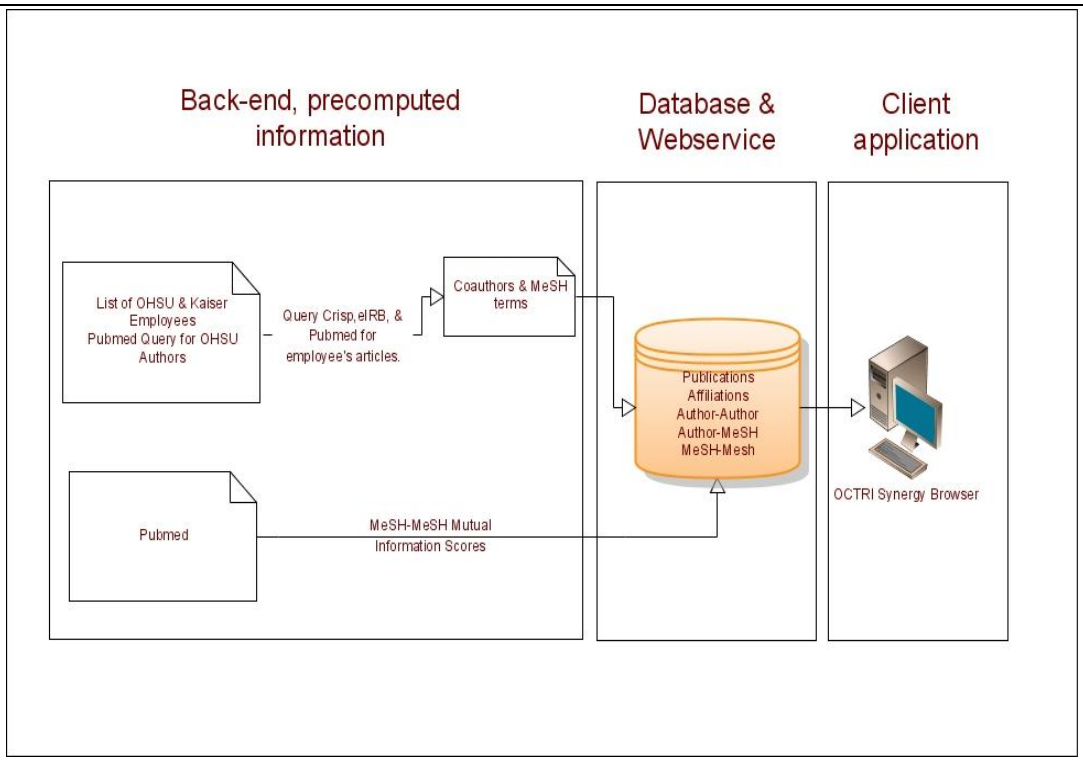

Figure 5: The architecture of the Synergy Browser Application.

The application evolved into a Java-based, client-server architecture with three components: the database, web service, and web client. Java was selected because of its portability and large number of third-party libraries which extend its functionality. The database stored connections between authors and MeSH terms, as well as additional affiliation data for the researchers. The web server-client setup provided a means of distributing and updating the application without having to worry about installing a local copy of the database or whether the user had the most up-to-date version of the application. Below, the details of these three components are described in greater detail.

Back end

The back end was setup to perform fetching, cleaning, and storing tasks that were too time consuming to perform in real-time for an application. The overall design was to have an application assist a user in exploring researchers by topic space or association with three connection types: MeSH-MeSH, author-MeSH, and author-author. A MeSH-MeSH connection was defined by the previous work as two terms having a high mutual information score or occurring together frequently in the literature. An author-MeSH connection indicated that an author published using a given MeSH term. An author-author connection indicated that two authors published together. A domain expert could then identify their needs by exploring these data.

The process for obtaining the MeSH-MeSH mutual information scores remained largely unchanged from the preliminary work. A python script iterated through the publications and gathered counts for number of times a MeSH term occurred in the literature and the number of times it co-occurred with another term in the literature. The mutual information score was then computed from these counts. The scores were computed over the entire MEDLINE corpus to obtain an unbiased representation of which topics might be related to one another.

Author-mesh and author-author connections were from MEDLINE publications and grants related to OHSU and Kaiser Northwest, or OCTRI, authors. This was done in a two step process of first compiling a list of researchers and then retrieving their grants and publications. The list was compiled because a researcher may have produced some literary work that was not necessarily tied to OCTRI. In these cases, their names could be

and were used as queries to retrieve that information.

Compiling the Name List

Initially, the list was built by simply extracting all authors or unique [Last Name] [First Initials] from all publications associated with OCTRI. This was flawed because it the affiliation field could only be reliably applied to the first author it did not account for ambiguity when the whole first name and middle initial were available. To account for the first problem, only first authors of the affiliated papers were extracted and these names were supplemented with incomplete employee lists. The employee lists, while incomplete, helped pull in authors who have never published first, or at all, but still had skills to make them reasonable candidates for collaboration.

Examples of ambiguity (not considering typos):
(JJ Johnson) – All different individuals
jodi j johnson, jennifer r johnson, jessica j johnson, john j johnson

(JV Jui) – Possible name variants
jon von jui, jonathan jui

(W Hersh, B Hersh) – Etymological abbreviations
bill hersh, william hersh

(J Janovick) – Differences in punctuation and spaces
jo ann janovick, joann janovick

(M M Bliziotes) – Missing data where compatibility is ambiguous
matthew m bliziotes, m michael bliziotes

Figure 6: Examples of ambiguity (assuming no typos).

The compiled list had 3,836 unique names, however, at this point many of those names were potentially ambiguous. For instance 'JJ Johnson' could refer to either 'Jodi J Johnson' or 'Jessica J Johnson'. In this case it is important to recognize that there are two distinct individuals, not three, and that their works should be correctly attributed to them when possible. Figure 6 enumerates the possible ambiguities that could come from having incomplete author names. To help resolve this conflict, the names were clustered according to compatibility. The rules for compatibility of two names are as follows:

1. The two names must share the same last name and first initial to be evaluated.
2. The two names must be compatible on their first and middle names.
    a) A first name is compatible if:
        - Both first names are available and equal.
        - One or both have only the first initial available and both first initials are equal.
    b) A second name is compatible if:
        - One or both are missing.
        - Both middle names are available and equal.
        - One has a middle initial and the other has a middle name and both middle initials are equal.

Figure 7: Rules for resolving name ambiguity.

The clusters were then organized into a name directory and indexed by [Last Name] [First Initial], as seen in Figure 8. A table of assigned articles was also set up to link the publications to each of the unique names in the publications.

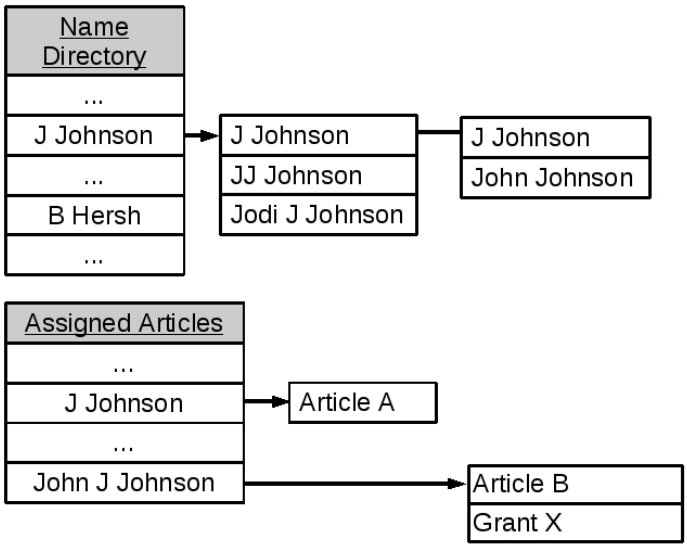

Figure 8: The name directory and assigned articles table. These two data structures helped organize ambiguous author names and assign articles as precisely as possible given incomplete information.

These clusters then allowed for either the most complete name to be composed and the incomplete names could be used to reference works that were ambiguously attributed to a given name. For instance, if there is a publication under 'JJ Johnson' it would be available for both 'Jodi J Johnson' or 'Jessica J Johnson' because there is no additional information indicating otherwise. However, if the publication was written by 'Jodi Johnson', it would be correctly attributed to the cluster of names associated with 'Jodi J Johnson'. When these rules were applied to the compiled author list, 3,451 unique authors were left.

Retrieving Grants and Publications

There were three data sources from which publication was retrieved: CRISP, eIRB, and MEDLINE. CRISP and eIRB data was provided as a tab-delimited text file of grants associated with the OCTRI. Each item in those files where then assigned to the name they were most compatible with. For the MEDLINE data, the author names were initially used to query for publication data. This, however, still resulted in ambiguity when authors shared the same last name, first initial. This required disambiguating MEDLINE publication data to determine which publications belonged to an OCTRI author and which did not.

As a first attempt, the query “Last_Name First_Initial[AU]” was used to retrieve publications from MEDLINE, since it was the degree to which MEDLINE uniquely identified authors. This quickly proved unusable as common names like “Stephens J” yielded >10,000 records, which would clearly exceeded even the most prolific individual's ability to publish papers. Heuristics were then applied, to trim these numbers down to < 100 publications. These were: 1) to use the full author name, and 2) to use some OCTRI affiliation modifiers should the first heuristic yield too many results. The publications were inspected manually by looking at the author name and article. If the article had an OHSU affiliation, was listed in personal websites by the author, or the MeSH terms were consistent with the author's work, it was scored as correct. If it did not fulfill these criteria, it was scored as incorrect. The manual inspection revealed that ~30% of the articles did not appear to be attributed to an OHSU author. The noise still present in the data demanded that a more rigorous approach or tool to acquire accurate

publication data.

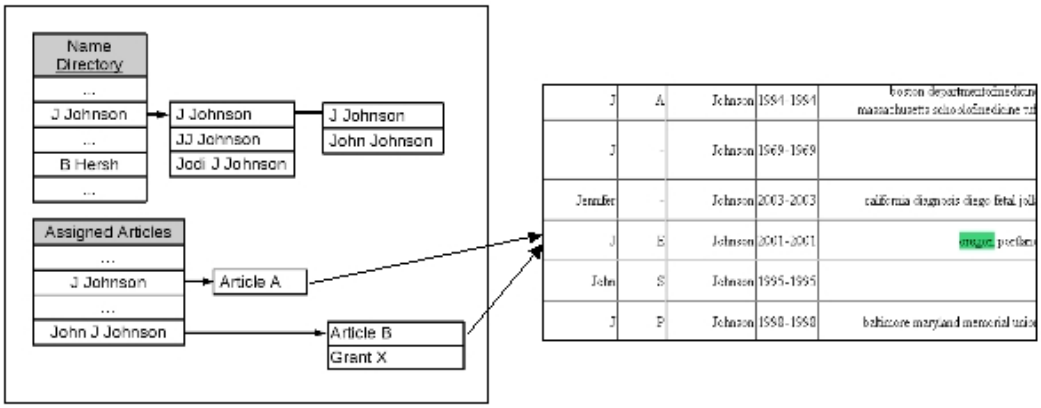

Figure 9: Articles unambiguously tied to OHSU were used to identify the Author-ity cluster which contained all of the researcher's papers.

Author-ity[30], the selected resource, provided MEDLINE publications, disambiguated and clustered by probabilistically unique authors. An author's publication data was retrieved by using papers unambiguously associated with OHSU as a pointer to the correct cluster of papers in Author-ity. This allowed the database to include articles which the author may have published independent of OHSU. Ultimately, this helped disambiguate OCTRI researchers from others sharing in the same name in other parts of the world. This resulted in 2,321 of the 3,451 researchers having some publication information about them.

The publications and grants provided information about an author's specialty and their coauthor's specialties in the form of Author-MeSH and Author-Author connections.

This information was then stored in a database with the Author-MeSH and MeSH-MeSH connections to facilitate exploration of the author-topic network.

Database & Web Service

The database was connected to a Java web service which provided remote access to the data. The database schema stored information on various entities: authors and MeSH terms, which were linked together through co-occurrences in publications. The web service then provided access to blocks of related concepts that a user would request as they were exploring the connected data. This setup was chosen because it was much more practical to maintain a central repository of researchers that was subject to change as more features were developed and more articles were published. Figure 10 provides a description of the schema and Figure 11 provides a description of the service.

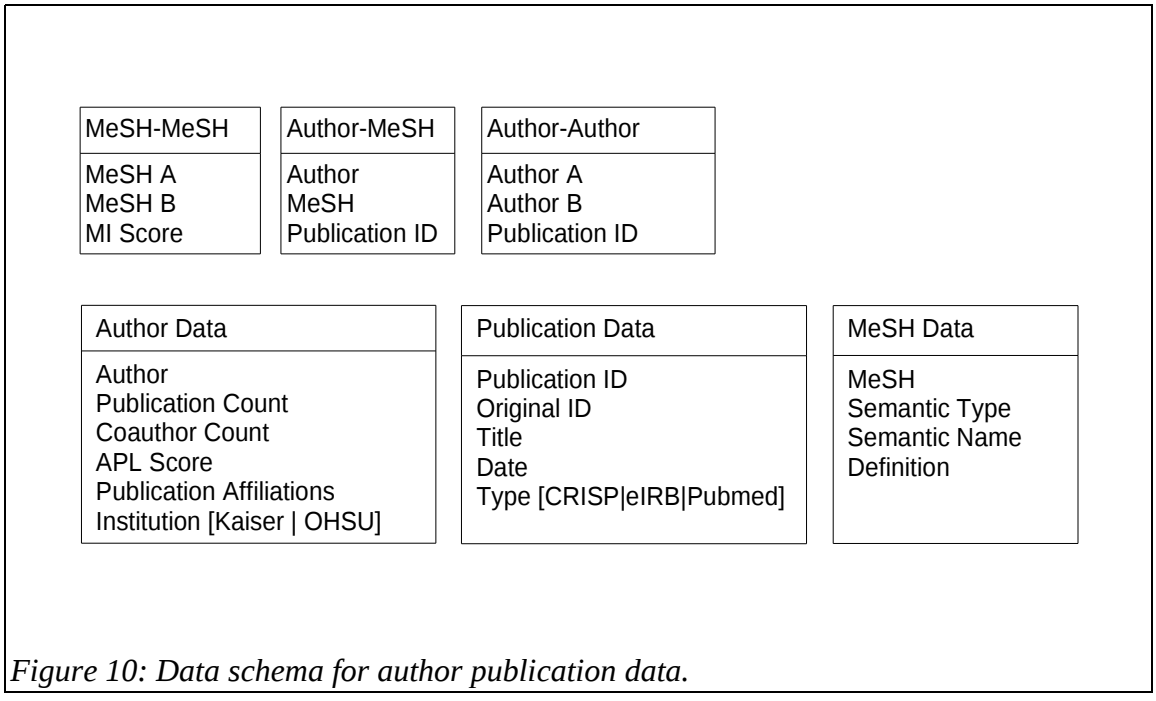

Figure 10: Data schema for author publication data.

Classes

MatchWS	Provide an interface to the service's functions.
Data	Package chunks of data to send across the server.

Functions

Search Methods

Search author	Return authors to explore.
Search MeSH	Return MeSH terms to explore.

Connection Methods

Author-Author	Return coauthors.
Author-MeSH	Return MeSH terms published by authors.
MeSH-Author	Return authors published under MeSH term.
MeSH-MeSH	Return MeSHs with high mutual information scores.

Data Methods

pub-data	Return more information about a publication.
author-data	Return more information about an author, particularly their list of publications, affiliations, and contacts.
author-author-data	Return a list of papers shared by a given coauthor.
author-mesh-data	Return list of publications in which the author used a given mesh term

Other App Specific Methods

semantic-types	Return a list of semantic types, which will be used to filter the types of mesh terms visible
semantic-MeSH-type	Return the mesh words for a given semantic type
check Author-Author	Perform lookahead on author-author connection
check Author-MeSH	Perform lookahead on author-mesh connection
check MeSH-MeSH	Perform lookahead on mesh-mesh connection.
author-MeSH-date	Return the max and min dates for a connection
author-author	Return the max and min dates for a connection

Figure 11: The web service methods.

The web service provided a number of methods which facilitated the creation and exploration of dynamic network of synergistic authors and MeSH terms. The search

methods were used to start the search by providing a list of valid authors or MeSH terms recorded in the database. Related concepts could then be expanded off of these terms by using the appropriate connection method. For instance, the Author-Author method returned a list of all authors co-published with a given author, while the MeSH-Author returned a list of all authors published with a given term. The search and exploration was constrained to authors and major MeSH terms to avoid the noise of connecting entities from an uncontrolled vocabulary. These methods were then supported with others that fetched more data for a given node or edge. Author-data and pub-data retrieved an author's affiliations, contacts, and list of publications. The edge methods, such as author-author-data and author-mesh-data retrieved the list of publications that created the connection.

The server provided data-fetch routines for most of the client's regular actions. Most of these routines focused on how the client changed from topic to topic. Changing how data was fetched would only require refactoring the server-side code. Adding new features or data-fetch routines often required that the client-side code be re-factored as well, whether to use new operations or to apply different usages of the service's objects.

Client Application

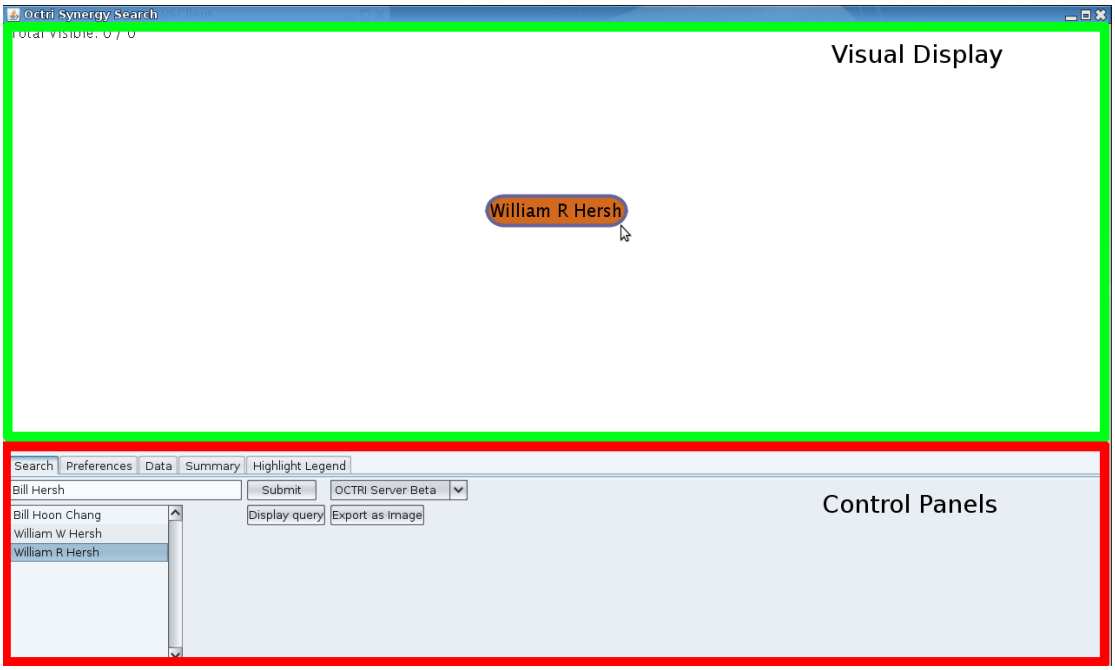

Figure 12: The OCTRI Synergy Application

The client end was a Java application implemented with JNLP[31] which allowed a user to browse over the researcher data and their corresponding MeSH and co-author connections that were stored in the database. This applet was composed of two main parts: a visual display of the researcher data and a control panel to adjust and explore the network.

Visual Display

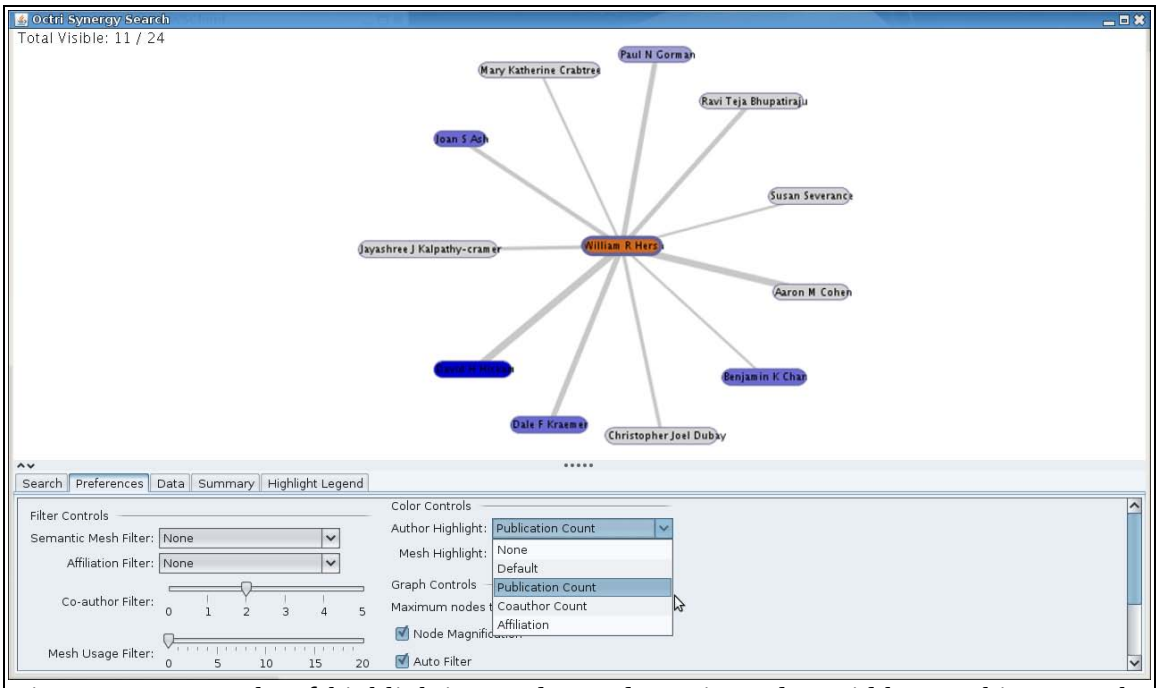

Figure 13: Example of highlighting nodes and varying edge widths. In this case, the darker nodes indicate which researchers are more prolific. The thicker edges indicate that a researcher has published more with the originating author.

The visual display, rendered using Prefuse[29], draws out a network of related concepts in the form of an undirected, acyclic graph. The nodes represented authors and MeSH terms while the edges represented some evidence connecting those items together. A user extends the network by right-clicking on an item and selecting 'Expend Author' or Expend MeSH' from a pop-up menu. Conversely, the graph is reduced by selecting 'Collapse Author' or 'Collapse MeSH' from the same pop-up menu. The nodes and edges in the network could also take on various visual properties to accentuate their importance for a specific kind of exploration. Nodes could be highlighted different colors from either being of a particular type or appearing in more publications. Edges had varying degrees of thickness to represent the strength of a connection. In Figure 13, the edges were

between to authors and thicker edges illustrated more co-authorships. This was done to provide surface level information which would attract a user's attention and draw them to more prominent researchers or frequently used MeSH terms.

Another feature that adjusted the display of information for the author was an author-filter. Over long careers, some researchers may have accumulated hundreds of coauthors and MeSH terms. In most cases, the researcher works predominantly with a small group of close associates on a specialized set of topics. The author-filters, in an attempt to reduce data overload, were applied to only show the most prevalent items. This, along with the visual cues was setup to help user find groups of synergistic researchers faster.

Control Panels

Searching provided a means of initializing the network search since authors and MeSH terms could only be expanded off of other author and MeSH nodes in the database. The search simply performs a MySQL Fulltext search on the tables a list of valid items. The user can then select an item to appear in the visual display and expand from that to explore the network of connections.

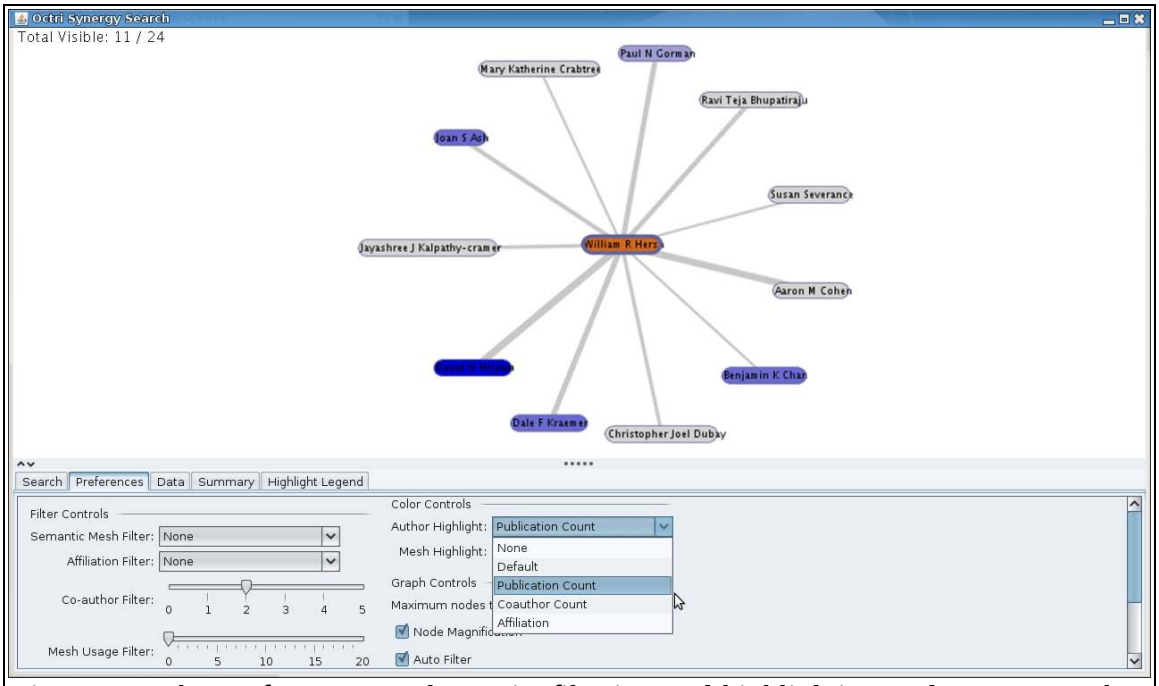

Figure 14: The preference panel permits filtering and highlighting nodes. Here, author nodes are given a heatmap to show which authors have published more, with gray being few and blue being many.

The preferences panel was used to change the filter and highlight settings for the network or nodes in the network. The filters altered the number of child nodes visible at a given threshold based on their features or strength of the edges and each node had its own settings to allow the filters to be applied locally. The feature filters, MeSH semantic and author affiliation, would only show nodes of a specific type. The edge filters co-author, MeSH usage, and MeSH mutual information only showed nodes whose connection strength was above a certain threshold. Highlighting changed the nodes' colors based on its features; this was found useful only for author nodes. Color-coded palette, a set of colors progressing from gray to blue, were applied to publication counts or coauthor counts to show the more prolific or more collaborative authors. Enumerated colors were used to show the authors affiliation of OHSU, Kaiser, or other.

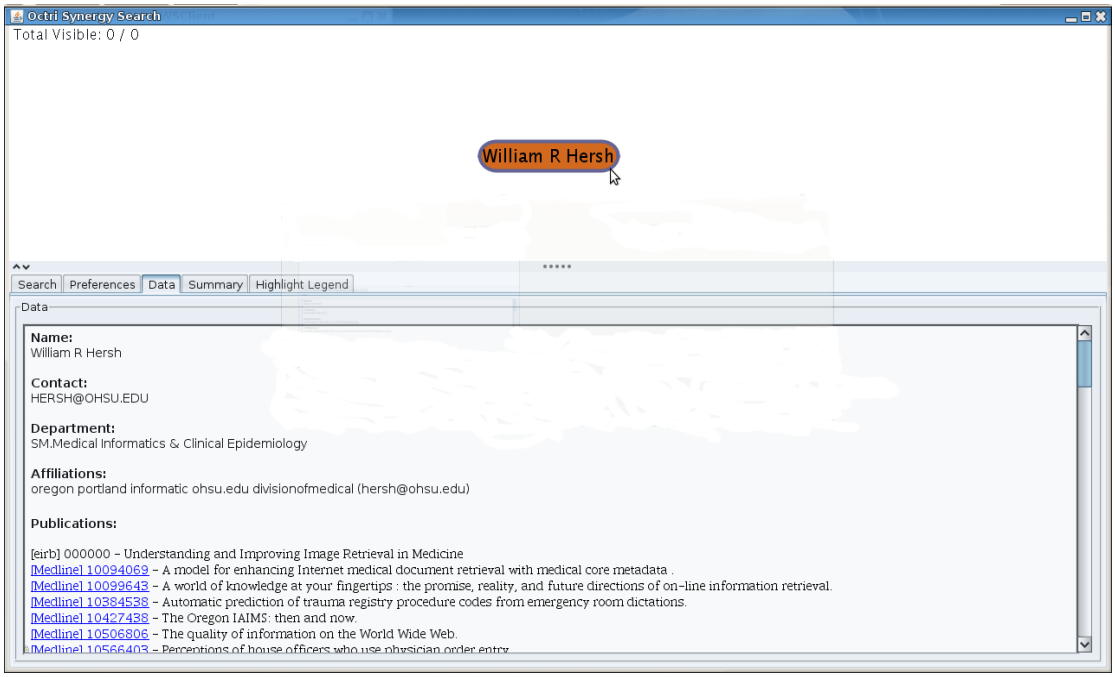

Figure 15: The data window displays extra information that could be easily embedded in the visual display.

There was some information which would have been too copious to put in the visual display. Therefore, a data panel was provided to show additional information for the last item, node or edge, that the user clicked on. Author nodes showed the authors name, contact, affiliation, and list of publications. MeSH nodes showed that terms definition. Clicking on edges showed the rational behind that connection. If it was a author-MeSH or author-author edge, the publications containing those co-occurring terms were shown. If it was a MeSH-MeSH edge, it showed the mutual information score between the two items.

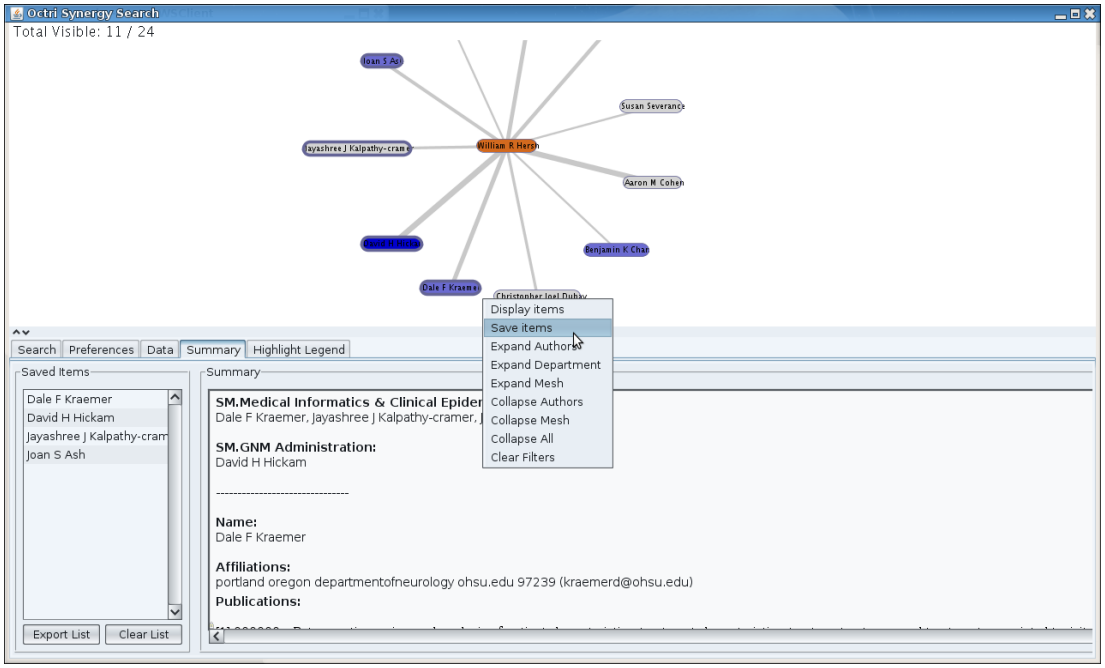

Figure 16: Nodes can be saved so the information could be re-retrieved or exported to a text file.

The summary panel stored nodes that were saved through the right-click pop-up menu. The saved items were stored in a list box and could be selected to start a new search with that item and all of the items' data were shown in a text box, which could be exported to a plain text file. The purpose of this was to make it easy to move and share large amounts of researcher publication data.

Evaluation

In order to test the validity and efficacy of the system, a set of evaluations were run on it: data validation, use cases, and surveys.

Data Validation

Much of this project relies on the accuracy with which publication data can characterize a researcher. Due to the possible ambiguity in MEDLINE data and lack of any complete profile information for a given researcher, a random sample of publication information was verified by hand. In this phase, 100 author-MeSH-publication ID tuples were selected randomly from the database. The criteria for being a true positive was if the publication was affiliated with OHSU or if the author published with the MeSH term under another paper affiliated with OHSU. Affiliation was determined by observing it in the MEDLINE record or seeing the paper recorded on an OHSU website for that researcher. The latter criteria was useful in the event that no affiliation was available for the publication but that the connection might be plausible. This step was particularly important because the publication characterized researchers and was the basis for determining whether they had synergistic qualities.

Use Cases

The use cases served to reveal user needs, drive development and demonstrate how the application could be used for its intended purpose. The procedure for a use case was relatively straightforward in that a user would present some information need and the application would be applied. This would show what users wanted and what additional features or information were needed.

Surveys

The survey measured the differences between the using traditional methods, i.e. Internet and acquaintances, and using the OCTRI Synergy Application to identify investigators for future research projects. In this evaluation, there were two sets of participants: domain experts and assistants. The assistants were tasked with identifying a lists of researchers to be invited on research projects. The experts would then validate those lists using their extensive expertise. This was organized as such due to the scarcity of most experts' time. This process was organized into four sessions:

1. Training the assistant.
2. Selecting researchers without using the synergy application.
3. Selecting additional researchers using the synergy application.
4. Scoring of the invitee appropriateness by the OHSU domain .

In the first session, assistants underwent a supervised, 30-minute training session to learn how to use the application. In the second session, the assistants identified researchers who might contribute to a research project and were permitted to use any other resources: the Internet, MEDLINE, grant databases, and prior knowledge. The organizers were then asked to record their choices and their reason for selecting a researcher on the survey document. The third session, was similar to the second session, but the organizers were asked to use the OCTRI synergy browser. They were asked to expand on the initial list by selecting more researchers and to record their thoughts on the process.

- Importance implies that a selected researcher is essential to the retreat topic at hand. Their "essential-ness" comes from years of experience or being very prolific in their field.
- Appropriateness characterizes how well a researcher's skill set and resources fit the retreat topic. A researcher may be tied to a research topic, either through association or by making a minor contribution to a paper on the topic, thus earning a co-authorship. If their expertise is low and they are still invited, they may not provide the desired spectrum of skills to participate in the research project. It is therefore important that the researcher's abilities match the roles to which they are being invited.
- Novelty indicates that a researcher is bringing a unique set of skills to the retreat. They are knowledgeable in the retreat topic but may not traditionally be included in such research teams. This group, with these novel participants, may then be able to break down traditional barriers and increase translational research.

Figure 17: The criteria for which a researcher is scored on.

In the fourth session, the domain experts were asked to review the retreat organizer's list of researchers and score the invitees for importance, appropriateness, and novelty. These features, in Figure 17, were scored using a Likert scale from 1-10 with 1 being not at all, and 10 being the maximum amount.

Results & Discussion

The following section presents the results the tests performed in the system: data validation, use cases, and surveys. It is important to note that the results were not simply the end product of a test but a guide driving the incremental development cycle. In incremental development newer versions of a system are updated based on what is learned in earlier versions of the system. For instance, the data validation revealed that

applying heuristics to disambiguate MEDLINE publications led to inaccurate researcher profiles. This led to the inclusion of the Author-ity resource in the back end, where data was collected and processed. Given this, the results are presented along with a brief discussion of what their meaning implies and how they impacted system development.

Data Validation

The data validation was performed to ensure that publications extracted from MEDLINE originated from an OCTRI author. Initially heuristics were applied to retrieve a set of clean publications and a sample of 100 randomly selected author-mesh connections were evaluated. From this evaluation it was deemed that 59 items came from OCTRI authors, 37 came from non-OCTRI authors, and 4 did not have enough information to distinguish the researcher. When applied to the extracted Author-ity clusters, 91 items came from OCTRI authors and 9 seemed to come from non-OCTRI authors. Summarizing, the heuristics seemed to provide ~60% accuracy for retrieving publications while the Author-ity data provided ~91% accuracy.

One disclaimer that should be made is that there was no gold standard list of whom the publications actually belonged to. In other words, there was some potential for mislabeling these items due to insufficient information. However, given these limitations, it was still fairly clear that the Author-ity resource provided a significant improvement in data quality.

Use Cases

The intent of the use cases was to show what user's collaboration needs were and to provide for those needs where possible. The present use cases come from two groups: the OCTRI and the Child Development and Rehabilitation Center (CDRC).

The OCTRI's research needs were more managerial in nature. They wanted to have the ability to monitor and influence collaboration activities at OHSU to enhance the rate of translational research. They specifically wanted investigate the changes in author-author and author-MeSH connections for a number of researchers between 2007 and 2010. During this time span, the OCTRI was active in connecting researchers together, and such information would show how their efforts may have expanded a researcher's social and scientific resources.

The CDRC researchers wanted to identify partners in applying for an Intellectual and Developmental Disabilities Research Centers (IDDRC) core grant. This grant provides support to centers which, as the name suggests, prevent or treat developmental disabilities. One of the eligibility criteria for a center is to demonstrate that it supports 10 or more externally funded projects on the topic. The CDRC, therefore, wanted to develop a list of all funded projects under the MeSH topic Developmental Disabilities and contact the investigators for a possible collaboration.

The following cases describe how the application was used to find these pieces of information and what was found.

Case 1

Name: Observe New Co-author Connections

Goal: Identify if a researcher has had new connections or collaborations formed over the past few years.

Summary: The OCTRI wanted to observe the formation of new partnerships between 2007 and 2010 for Lyle J Fagnan, Charles Robert Phillips, and Jonathan Q Purnell. The objective was to provide evidence of a change occurring in the researcher network and to attribute it to the OCTRI's match-making efforts.

Actors: User

Basic Course of Events:

Search for author in the application.

Expand coauthors

Export the displayed content to an image

Export the researcher's data to a text file

Move slider from 2010 to 2007.

Export the displayed content to an image

Export the researcher's data to a text file

Results:

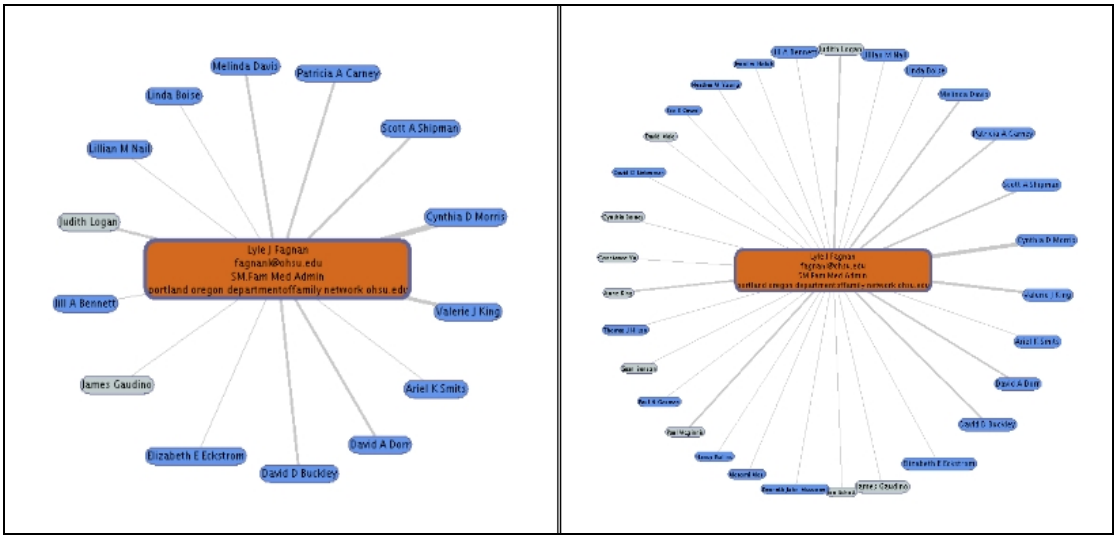

Figure 18: A demonstration of Lyle Fagnan's coauthors (left) before 2007 and (right) after 2007

Name	**# Coauthors pre-2007**	**# Coauthors post-2007**	**# post - # pre**
Lyle J Fagnan	14	30	16
Charles Robert Phillips	23	35	12
Jonathan Q Purnell	90	99	9

During the period 2007 - 2009, each of the researchers had a significant increase in new coauthors. Lyle Fagnan and Charles Phillips had around a 100% increase while Purnell, a researcher with a larger social network had a 10% increase. Interestingly enough, this sharp increase coincides period which the OCTRI was actively forming connections between researchers. While the application cannot prove that the OCTRI was responsible for providing these new coauthors, it does provide a starting point for investigating these relationships because the user is able to observe how the connections

change from year to year.

Case 2

Name: Observe New MeSH Connections

Goal: Identify if a researcher obtained new knowledge or resources over the past few years.

Summary: Like the previous use case, the OCTRI was interested in observing the changes in researcher's MeSH terms between 2007 and 2010 for: Lyle J Fagnan, David Feeny, Ann B Hill, Eric Johnson, Alison Naleway, Jonathan Q Purnell, Kathryn G Schuff, Mary P Stenzel-Poore, Gary Thomas. In this case the objective was to see if a researcher gained any knowledge or resources over the past few years due to recent collaborations.

Actors: User

Basic Course of Events:

Search for author in the application.

Expand MeSH terms

Export the displayed content to an image

Export the researcher's data to a text file

Move slider from 2010 to 2007.

Export the displayed content to an image

Export the researcher's data to a text file

Results:

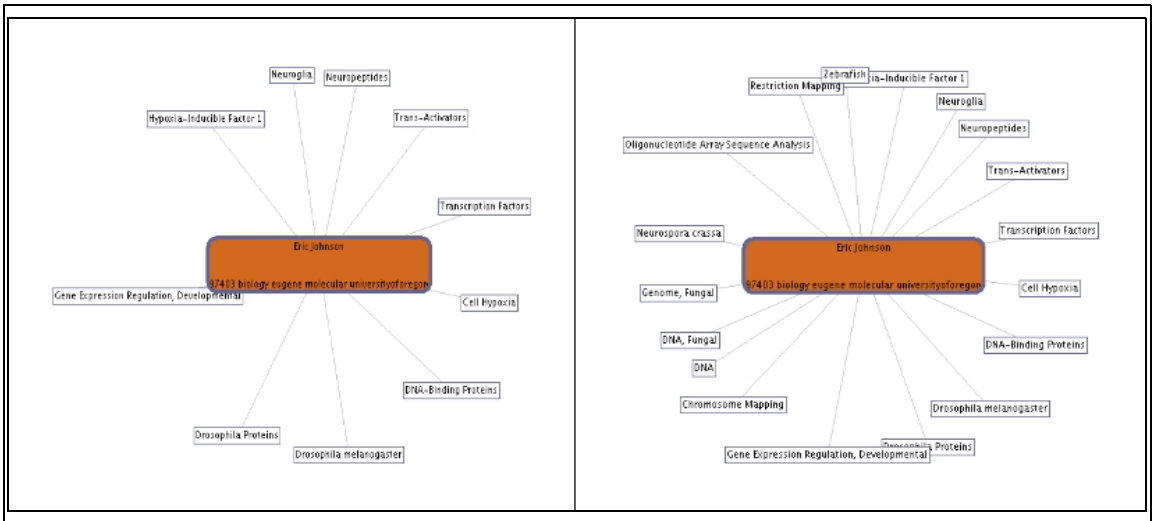

Figure 19: A demonstration of Eric Johnson's MeSH terms (left) before 2007 and (right) after 2007

Name	**# MeSH pre-2007**	**# Mesh post-2007**	**# post - # pre**
Lyle J Fagnan	13	26	13
David Feeny	0	2	2
Ann B Hill	78	88	10
Eric Johnson	11	19	8
Alison Naleway	31	33	2
Jonathan Q Purnell	147	163	16
Kathryn G Schuff	30	43	13
Mary P Stenzel-Poore	134	150	16
Gary Thomas	187	193	6

The results for the MeSH-term analysis where slightly noisier. Over the period of 2007-

2009, most of the authors showed a modest increase of MeSH terms, ranging from 2-16, with an average of 10 and standard deviation of 5. This shows that researchers actively expand the scope of their research and, consequently, need to be constantly seeking out new expertise or resources. Again, it is difficult to attribute this change in acquiring new MeSH terms to the OCTRI, but the data suggests that researchers would benefit from a service to help them find various resources. If they had such a service, they would be more likely to locate and collaborate with experts who share the same interests. This, in turn, would enable them to tackle larger projects and acquire grants.

Case 3

Name: Finding Grant Partners for Developmental Disabilities Research

Goal: Identify researchers with grants covering Developmental Disabilities.

Summary: Robert Steiner and his associates at the Child Development & Rehabilitation Center (CDRC) were seeking to apply for an Intellectual and Developmental Disabilities Research Centers (IDDRC) core grant. The prerequisite to signing up for such a grant is to demonstrate that the center provides support for 10 or more externally funded projects. The application in this case, was used to find grants under the MeSH term Developmental Disabilities.

Actors: User

Basic Course of Events:

Search for MeSH term Developmental Disabilities in the application.

Expand Authors

Click on MeSH-Author edges to see if any contain current grants.

Results:

Name	Grant Type	Grant Name
Agnieszka Z Balkowiec	CRISP	Neurotrophins and Development of Baroreceptor Pathways
Jane Squires	eIRB	The Ages and Stages Questionnaires and Children with Developmental Disabilities
Lavinia L Sheets	CRISP	Regulation of Molecular Motors in Zebrafish
Robert E Nickel	eIRB	The Ages and Stages Questionnaires and Children with Developmental Disabilities

In this case, there were 3 possible groups that the CDRC could have teamed up with to write an IDDRC grant. In performing this search, there were three observed shortcomings: 1) the publications were limited to items indexed in MEDLINE, 2) the grants only covered human research, and 3) there were an insufficient number of grants under developmental disabilities to write up an IDDRC grant. In encountering points 1 and 2, Melanie Fried-Oken of the CDRC provided a list of OHSU funded projects on Developmental Disabilities that were not indexed by CRISP or eIRB. This highlighted the need to investigate more sources publication and grant data, such that a central repository could be searched and browsed by author or topic. The third point illustrated how useful it might have been for the CDRC to expand on related concepts in their search for grant-writing partners.

Surveys

OCTRI Synergy Application Evaluation 1 Topic: Decision Making Expert: KE Assistant: PS							
Test Type	# Researchers Found	Importance Average / Stdev		Appropriateness Average / Stdev		Novelty Average / Stdev	
w/o Application	15	9.47	0.74	8.73	1.49	8.8	1.26
With Application	15 + 8	10	0	9.38	0.92	9.63	0.74

OCTRI Synergy Application Evaluation 2 Topic: Simulation, patient safety, epidemiology, fecal incontinence, constipation, pregnancy Expert: JG Assistant: CO							
Test Type	# Researchers Found	Importance Average / Stdev		Appropriateness Average / Stdev		Novelty Average / Stdev	
w/o Application	10	6.3	2.98	6.1	2.85	5.8	2.2
With Application	10 + 11	6	2	6	1.41	6.55	2.25

The survey results showed that the application was able to find additional researchers, in each case, nearly doubling the pool experts could identify collaborative partners from. These researchers were deemed to be as good as those found using traditional methods over the attributes of importance and appropriateness. Also, though not statistically significant, there appeared to be a slight increase in novelty by one point.

In addition to making and scoring researcher lists, the assistants were asked to fill out a

questionnaire regarding their experience in using the Internet, e.g. Google, Pubmed, or CRISP, versus the application. The questions simply asked what were the pros and cons of each method. Some of the notable responses were:

Internet Pros:

“The text descriptions that are underneath each of the items retrieved” where helpful.

Internet Cons:

“Lack of focus when searching. The searches resulted in a lot of items that weren't relevant...”

“Often it was difficult to weed through all of the research on a given topic to find local researchers. Also, there was a duplication of entries that hindered the search process.”

Application Pros:

“The 'Expand MeSH' and 'Expand authors' features and variety of search terms.”

“The data portion that listed the publications was particularly helpful. Also, being able to see which department a potential collaborator was in (was useful) in assessing whether or not that individual would be considered.

Application Cons:

“Often the MeSH bubbles overlapped and it was difficult to access all of the bubbles of a given branch.”

The questionnaire revealed that the application was effective because it reduced noise and permitted an expansion of the search query through related authors and MeSH terms. The

noise was reduced by creating an author-centric database of OHSU and Kaiser researchers, which had some of the ambiguities managed by combining compatible names and using Author-ity to identify the most likely collection of papers. Noise was reduced by only showing OHSU related material and redundancy was also by identifying unique authors for possibly many names.

The application's expansion of related concepts enabled a focused exploration of the search space. In a typical Internet search, the user enters a query and is returned all items similar to the query. If the user information on complementary items, they would have to enter a new query or read through the results. By providing expandable MeSH terms, either through use or mutual information, the user spends less time digging through results corresponding to the original query. These features would then allow the user to review researchers more efficiently.

Discussion

In this study, an application was developed to assist researchers in finding compatible partners at OHSU to work with. This was done by creating an author centric database and an application for browsing related information in it. The data was browsed along authors or MeSH terms through 4 connections: Author-Author, Author-MeSH, MeSH-Author, and MeSH-MeSH. Author-Author represented co-authorship between two authors. Author-MeSH and MeSH-Author indicated that an author published under a given MeSH term. MeSH-MeSH indicated that two terms co-occurred exclusively together via the mutual information score, computed over the MEDLINE corpus. The

efficacy of this setup was evaluated by having researcher-assistant pairs construct lists of people for whom they would want to collaborate with. The application was found to nearly double the candidate list by identifying researchers through co-authorships or terms related to the initial query. Though there were not enough samples to imply statistical significance, the results also seemed to indicate a trend towards finding more novel candidates when using the application.

A possible explanation for the trend in novelty is that the application encourages users to explore beyond their initial query to related topics. Typically a user searches on what they are most familiar with and remain there because they are not aware of other possibilities and search engines are designed to return a list of items most similar to the query. This can be problematic for an investigator attempting to form a multidisciplinary because they are not expanding the scope of their search. This behavior was observed in the assistants' notes when selecting potential collaborators. In the first evaluation, with PS and KE, the assistant listed 5 out of the 15 candidates because she had prior knowledge of their work. In the second case, with CO and JG, the assistant was relatively new to the department and, instead, listed researchers by publications she was familiar with. In the former case, the researcher would not be making new connections and in the latter case, the researcher would not be pushing the bounds of their research.

This behavior changed when using the application, because partners were identified through topics that were adjacent to the initial query. In the first case, the assistant identified candidates by their body of literature listed in the data panel. In the second case, the assistant identified researchers through linking MeSH terms, or

complementary skills. This allowed the application to find another set of researchers whom the assistants, initially, were not aware of.

The problem with typical search engines is that they are designed with a different task in mind: to return what the user asked for. The results are then a list of items in decreasing similarity. This presents a limitation in constructing multidisciplinary teams in that each discipline or desired skill requires a separate search to identify a corresponding research partner. If, however, there is a gap in knowledge that requires translational research, the user may overlook researchers who publish on topics that are adjacent to but not the same as those they queried on. This was addressed in the Synergy Browser by guiding users along connected nodes and showing the supplementary information in a data panel. This allows the actual search to not be disrupted.

Given these design features, the Synergy Browser was successful in finding more researchers, but the current work does have limitations which prevented it from being significantly better than a regular Internet search. One major limitation is the number of users and the time available for data collection for a masters thesis. This prevented the survey results from being statistically conclusive. Another limitation was that the publication information needed to be collected and processed from many disparate sources. Some data sets, like MEDLINE, were available through special request, while others were not available through any obvious means. BiomedExperts and Author-ity serve as additional examples of this limitation in that their services are based mostly on MEDLINE data; BiomedExperts does show the grants indexed by CRISP. This meant that the ability to identify interesting cross-disciplinary relationships was limited to what

could be extracted from MEDLINE, i.e. biomedical research. The available resources for this study were MEDLINE, CRISP, and eIRB which helped the users find human biomedical investigators.

In addition to this, there were other limitations regarding a lack of information available in an article's citation. Disambiguating ownership, as discussed in the methods section, proved to be an important hurdle in terms of accurately characterizing a researcher. The Author-ity resource only covered MEDLINE, so articles from other sources, like CRISP and eIRB, endured less sophisticated methods. In this case, the grants from those sources were small and constrained to OHSU, limiting the impact and probability of incorrect assignment due to ambiguity. If larger data sets were to be included, this would become an issue. Another piece of information that was not available in the publication data was the researcher's current status. In some cases, a researcher may have been published at OHSU and moved elsewhere. This would ultimately prove problematic when attempting to contact that researcher to form a team. In light of this, the study had to settle for using the publication it could obtain and process.

Other challenges included having a limited participant size and employing a rather experimental user interface. In having a limited number of participants, this study was unable to draw any statistically significant results, making it difficult to definitively measure the application's efficacy. Tied in with a small user base was the implementation of a rather novel interface for this task. Having a small user base would limit the number of bugs found and recommendations of desired features. An application, particularly an

experimental one, benefits from outside use as it helps evolve the system into a tool the user will need and use.

With regards to these limitations, the goals of future work will be to incorporate more literature sources, include researchers outside of OHSU, and improve the usability of the user interface. This will in turn, attract more investigative users by supporting their team building needs, and allow a larger scale evaluation.

The plans for adding more literature sources will include looking at animal grant database and other publication sources not indexed by MEDLINE. In meeting with users, one of their more frequent requests was to include animal grants. For clinical researchers, animal grants represented the application of cutting edge technologies that could enhance their research significantly. Users also noted that significant bodies of literature were missing from the domains of education and engineering. As mentioned before, MEDLINE and MeSH to not cover all scientific domains. Leaving out literature on these other disciplines may also leave out some of the more novel applications of technology to human treatment. Therefore future work will strive to include those alternate publications which OHSU researchers have published under.

In addition to this, literature from non-OHSU researchers may be included. This would be interesting to see if the Synergy Browser is deemed useful to the broader scientific community and to see what types of resources or individuals seek that they might not have at OHSU. Maybe these outside experts consist of old associates or they could demonstrate interesting cross-institutional ties. This might help identify specialized resources that facilitate research projects.

Beyond exploring alternate data sources, development on the visual interface will also continue. The notion of setting up a graphical browser over data is still very experimental and there were some minor issues that occurred, such as data overload and obscured nodes. Different layout schemes will be tested in the future to show the user the maximum amount of data efficiently.

The aims of the future work are to build on the study's current work of enhancing an investigator's ability to form research teams. By presenting users with an application of browsable author-MeSH connections, they were able to discover other researchers who have worked topics complementary to theirs. However, it was found that one of the more challenging aspects of setting this application up was in automatically retrieving high quality data. Publications are distributed across many databases which tend not to maintain unique author identifiers and have varying degrees of availability. Future work will focus on retrieving and cleaning different types of data for the Synergy Browser and work towards filling researchers' information needs for team creation.

Conclusion

This study provides a means of identifying researchers that are synergistic to the application user's needs. It has been shown to be a useful add-on to existing search methods in that it discover researchers who have worked on topics related to the query. Future work on the application will include improving the user interface and including other sources of researcher information. This will help improve the quality and formation rate of translational research teams at OHSU and the research centers.

Bibliography

1. Moses, H., Thier, S.O. & Matheson, D.H.M. Why Have Academic Medical Centers Survived? *JAMA* **293**, 1495-1500(2005).
2. Khoury, M.J. et al. The continuum of translation research in genomic medicine: how can we accelerate the appropriate integration of human genome discoveries into health care and disease prevention? *Genetics in Medicine* **9**, 665(2007).
3. Pober, J. Obstacles facing translational research in academic medical centers. *The FASEB Journal* **15**, 2303-2313(2001).
4. Bahr, N.J. & Cohen, A.M. Discovering synergistic qualities of published authors to enhance translational research. *AMIA Annu Symp Proc* 31-35(2008).
5. Smalheiser, N.R. & Swanson, D.R. Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. *Computer Methods and Programs in Biomedicine* **57**, 149-153(1998).
6. Newman, M.E.J. *Coauthorship networks and patterns of scientific collaboration*. **101**, 5200-5205(National Acad Sciences: 2004).
7. Birrer, M.J. *Translational Research and Epithelial Carcinogenesis: Molecular Diagnostic Assays Now-Molecular Screening Assays Soon?* **87**, 1041-1043(© Oxford University Press: 1995).
8. Karp, J.E. & McCaffrey, R.P. *New Avenues of Translational Research in Leukemia and Lymphoma: Outgrowth of a Leukemia Society of America-National Cancer Institute Workshop*. **86**, 1196-1201(© Oxford University Press: 1994).
9. Ki Hong, W. Retinoid chemoprevention of aerodigestive cancer: from basic research to the clinic. *Clinical Cancer Research* **1**, 677-686(1995).
10. Li, F.P. Translational research on hereditary colon, breast, and ovarian cancers. *J Natl Cancer Inst Monogr* **17**, 1-4(1995).
11. Saunders, A.M. Association of apolipoprotein E allele epsilon 4 with late-onset familial and sporadic Alzheimer's disease. *Neurology* **43**, 1467-1472(1993).
12. Tew, K.D. Glutathione-associated enzymes in anticancer drug resistance. *Cancer Research* **54**, 4313-4320(1994).
13. Zerhouni, E. MEDICINE: The NIH Roadmap. *Science* **302**, 63-72(2003).
14. Zerhouni, E.A. et al. Clinical and Translational Science Awards: a framework for a national research agenda. *Translational Research* **148**, 4-5(2006).
15. Ioannidis, J.P.A. Materializing research promises: opportunities, priorities and

conflicts in translational medicine. *feedback* (2004).

16. Contopoulos-Ioannidis, D.G. et al. MEDICINE: Life Cycle of Translational Research for Medical Interventions. *Science* **321**, 1298-1299(2008).
17. Mankoff, S.P. et al. Lost in Translation: Obstacles to Translational Medicine. *feedback* (2004).
18. Amazon Amazon.com: Online Shopping for Electronics, Apparel, Computers, Books, DVDs & more. at <http://www.amazon.com/>
19. Newegg Newegg.com - Computer Parts, PC Components, Laptop Computers, Digital Cameras and more! at <http://www.newegg.com/>
20. Huang, Z., Zeng, D.D. & Chen, H. Analyzing Consumer-Product Graphs: Empirical Findings and Applications in Recommender Systems. *Management Science* **53**, 1146(2007).
21. Collexis BiomedExperts: Scientific Social Networking. at <http://www.biomedexperts.com/>
22. Iravani, S.M.R., Kolfal, B. & Van Oyen, M.P. Call-Center Labor Cross-Training: It's a Small World After All. *Management Science* **53**, 1102(2007).
23. Newman, M.E.J. Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E* **64**, 16131(2001).
24. Newman, M.E.J. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E* **64**, 16132(2001).
25. Swanson, D.R. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect. Biol. Med* **30**, 7-18(1986).
26. Reminga, J. & Carley, K.M. *ORA: Organization Risk Analyzer*. (Tech Report, CMU-ISRI-04-106, CASOS, Carnegie Mellon, Pittsburgh PA: 2004).
27. Risch, J.S. et al. The STARLIGHT information visualization system. *1997 IEEE Conference on Information Visualization, 1997. Proceedings.* 42-49(1997).
28. Pacific Northwest National Laboratory PNNL: Starlight Information Visualization Technologies. at <http://starlight.pnl.gov/>
29. Heer, J., Card, S.K. & Landay, J.A. prefuse: a toolkit for interactive information visualization. *Proceedings of the SIGCHI conference on Human factors in computing systems* 421-430(2005).doi:10.1145/1054972.1055031
30. Torvik, V.I. et al. A probabilistic similarity metric for Medline records: A model for author name disambiguation. *Journal of the American Society for Information Science and Technology* **56**, 140-158(2005).
31. Zukowski, J. Deploying Software with JNLP and Java Web Start. at <http://java.sun.com/developer/technicalArticles/Programming/jnlp/>