The Research Mentor Ontology:
Empowering Biomedical Students to Discover Better Research Opportunities

Laura Miron and Sehj Kashyap

Although academic mentorship and resulting research projects play a crucial role in advancing any STEM field, there are currently few systems in place to help students connect with and choose between research groups. In this paper, we take the Stanford biomedical department as a case study, and show how aggregating data from publicly available sources into an ontology and front-end user application can empower students to find more potential research mentors, and more effectively choose between them. We associate faculty and students with an affiliated department, as well as research areas defined by MeSH terms and allow application users to query by affiliation or MeSH. Finally, we evaluate the success of our ontology based on a survey of graduate students doing research with selected faculty represented in the ontology, asking whether our methods mapped the faculty member to what the student evaluator defines as the correct MeSH research areas.

Background and Motivation

Academic mentorship is broadly recognized as an essential element of research training. Students who receive effective mentoring are more likely to be productive and satisfied in their career. However, the process of finding a mentor is difficult for students. At large academic centers like Stanford, the abundance of research opportunities and mentorship is both a blessing and a curse. Stanford has over 2,240 faculty members. In total, these faculty conduct more than 6000 research projects with a total budget of $1.63 billion (Fall 2018). Students search and select mentors from a large pool while matching on factors whose information is difficult to obtain.

The most common way to find a mentor is for the student to self-search eligible faculty and get to know them through conversations, lab rotations, or mini projects. The latter steps are time consuming, and students only engage in those steps with faculty whom they have pre-filtered. Stanford provides basic tools and services to help students pre-filter; however, these

tools have significant limitations and thus the most common way students pre-filter is based on word-of-mouth from other students.

To understand the limitations of current tools, we can consider the two most widely used: faculty websites and faculty directories. Faculty websites contain information about the faculty member's area of research, past projects, members, etc. Faculty directories are maintained by different academic departments and consist of a list of affiliated faculty with whom students can conduct research. A student searching for a faculty mentor in Biosciences would find over 450 faculty listed on Biosciences website. Each faculty member's website is linked; however, many of these links are either broken or contain no further information about the faculty member's research. Faculty who do have websites present different information in varied formats-- some faculty list current members while others include current, past and affiliated students (with pictures); some include links to recent publications while others describe broad research themes. Finally, some information that students consider important in searching mentors is rarely represented in these tools: for example, diversity of lab members, publication productivity of students, degree of collaborations, etc.

One factor limiting existing tools is difficulty of gathering and consolidating this diverse information. Currently, each faculty member maintains and updates their own website. There is no consistent format or vocabulary or required informational elements. It would be time-consuming to standardize this information and difficult to compel researchers to follow the standards; however, a system that automatically organizes this information and pulls from independent, diverse data sources could assist students in getting this vital information. In our project, we seek to create this system so that students can be empowered to find a research mentor.

What makes a good mentor?

There is limited published literature on how students can find and establish effective mentor-mentee relationships. This literature review spurred initial ideas for what insights our tool should provide students searching for a mentor. Some of these include identifying in mentors: record of collaboration, prior mentoring experience, strong teaching skills, and positive environment.[5] However, further improvements will be made as we test the tool with students and evaluate our tool as an aid in their mentor-finding process.

Methods

Our project has three components. First, we use *Protege* to create an ontology of biomedical research concepts. Second, we scrape Stanford faculty databases and PubMed articles for data which we process and model as instances. Finally, we evaluate the accuracy and completeness of the model using a test case of four labs in the Biomedical Department. Additionally, we build a prototype graphical user interface that performs reasoning over the model, and allows students to search for labs based on department affiliation or by research area defined by MESH terms.

Ontology

We designed the ontology by identifying the most important use cases of our user application, and letting this dictate which concepts needed to be defined. Our final application will:

1. Provide a lexicon of "research areas", namely MeSH terms within the 'Health Care' sub-hierarchy. When a user searches any of these terms, the application returns a list of Stanford faculty members affiliated with that research area.

2. Provide supplementary information about faculty results
    - All published papers by faculty member
    - Teaching experience and affiliated organizations
    - List of current students and past students who have co-authored papers with the faculty member

Based on these use cases, we model our system in OWL using *Protege* software. Our classes represent academic concepts applicable to any university. The most important classes are *FacultyMember, Student, Organization, ResearchArea,* and *AcademicPaper* (not pictured below because lacks any subclasses).

Fig. 1 - View of most important classes in Protégé. Left: Person, encompassing FacultyMember and Student, Middle: Organization, encompassing departments (e.g. Biochemistry Department), divisions (e.g. Medical School, Division of Neurosurgery), and other research centers. Right: ResearchArea, currently restricted to MeSH terms within 'health care' sub-hierarchy. Terms appear as 6 disjoint categories do to inconsistencies between the published NCBI MeSH, and the Robert Hoehndorf owl version of MeSH publish in BioPortal, used here.

*FacultyMember*s are related to *Students* via the *mentorOf/mentoredBy* inverse relationships, related to *Organizations* via the *currentMemberOf/hasCurrentMember* relationships, and related to *ResearchArea*s via the *hasResearchArea/researchAreaOf* relationship. Any *Person* (faculty or student) can be related to an *AcademicArticle* via the *authorOf relationship.*

We used a simplified version of VIVO, the current predominant ontology for modeling faculty research, teaching, grants and research organizations as the basis for our ontology, along with the MeSH "health care" sub-hierarchy as the definition of our research areas. We made adjustments including extending the *Student* class to finer granularity: *UndergraduateStudent, MasterStudent,* and *PhDStudent*. These distinctions are important to our target user; many faculty members only accept graduate students or PhD students as research assistants, and our user should be able to distinguish between these labs.

Next, we model Stanford University, specific faculty members, students, and research areas as individuals.

Data Scraping and Model Instances

We requested but never received directory information for current students and faculty from Stanford IT. Instead, we scrape faculty data and student data using *selenium* for python, from several public web sources: https://med.stanford.edu/profiles/ (med school faculty and students), https://biox.stanford.edu/ (undergraduate and PhD bioengineering students), and https://biosciences.stanford.edu/faculty/biosciences-faculty-database/ (biosciences faculty). We automatically create and add instances to the ontology using *owlready2* for python. From these sources we are able to get the first and last name of faculty and students, the degree program of all students, and at least one department/division affiliation for each faculty member.

In the next phase of data collection, we scrape PubMed for authors affiliated with Stanford who have published papers linked with all MESH terms related to the Health Care Category: Environment and Public Health, Healthcare Economics and Organizations, Health Care Facilities Manpower and Services, Health Care Quality Access and Evaluation, Health Services Administration, Population characteristics. We detected over 25,300+ Stanford researchers (including faculty, students and visiting scholars), 33,000+ publications, 15,000+ MESH terms, 3,000+ journals from the following years: 2001, 2008, 2010, 2014, 2015, 2017, 2019.

pmid	author	afft
30755192	Megan R Mahoney	Department of Medicine General Medicines Discipline ...
30683880	Esther M John	Department of Medicine, Division of Oncology and Sta...
30683880	Alice S Whittemore	Department of Health Research and Policy - Epidemiol...
30670697	Hua Tang	Department of Genetics, Stanford University, Stanford,...
30667499	John P A Ioannidis	Meta-Research Innovation Center at Stanford, Stanford...
30586832	Nona R Chiariello	Stanford University, Stanford, CA 94305, USA.
30586832	Christopher B Field	Stanford University, Stanford, CA 94305, USA.

pmid	date	journal	mesh
30755192	2019,Feb,12	BMC public health	Adolescent
30683880	2019,Feb,12	Nature communications	Breast Neoplasms
30670697	2019,Feb,12	Nature communications	Adolescent
30667499	2019,Feb,12	JAMA	Epidemiologic Studies
30586832	2019,Feb,12	The Science of the total environment	California
30586505	2019,Feb,12	The New England journal of medicine	African Americans
30586170	2019,Feb,12	Journal of surgical oncology	Aged

Fig. 2 Example data in our database of authors and publications with publication meta-data

We match authors against existing faculty and student instances in our ontology. If we find a match, we add the article as an *AcademicArticle* instance, all MESH keywords published with the article as *ResearchArea* instances, and associate the existing faculty and student instances using the *authorOf* and *hasResearchArea* relationships.

Problem-Solving Methods

We build a prototype graphical user interface using the *Tkinter* python package for the interface and *owlready2* to perform queries on the ResearchMentor ontology. Currently, the app is only able to perform queries of the form 'search for all faculty members associated with [organization OR MeSH term]'. The original intention was to use the python *RDFLib* library to perform SWRL reasoning for many of the queries in the app, however, we got better performance by using *owlready2* and precomputing several lists on load of the app (this preprocessing itself take <1s). The lists are of faculty member instances, student instance, document instances, organization instances, and MeSH term instances, and all remaining computation performed by the app is able to take place over these lists.

Evaluation

We conducted a formative quantitative evaluation of our ontology. We chose completeness and accuracy as two facets of the ontology that we wanted to explore in this testing stage. We

chose four lab groups represented in our ontology and asked one graduate student from each lab group to help in the evaluation.

- Accuracy: We wanted to evaluate whether our ontology correctly associated the right research area with the right lab groups. The graduate students were given a list of research areas (MESH terms) and asked to select which areas they felt were linked with their lab group. We then tested whether each lab-research area pairing was present in our ontology. For example, if a graduate student in Nigam Shah's lab felt that Healthcare Policy (research term) were interlinked, we checked if this Nigam Shah was linked with Healthcare Policy in our ontology. To make it easier for our testers to select research areas, we only provided them with a list of 20 MESH terms from the healthcare subtree (close to the root of the tree). The formula for accuracy thus was:

$$Accuracy = \frac{\#\ of\ reported\ pairs\ found\ in\ ontology}{\#\ reported\ research - lab\ group\ pairs}$$

- Completeness: We also wanted to evaluate whether our ontology had all the professors a Biomedical student may be searching for when looking for a mentor. We asked the graduate students which lab mentors they had either previously worked with or explored working with prior to joining their current lab. We then checked how many of these reported mentors were represented in our ontology. The formula for completeness thus was:

$$Completeness = \frac{\#\ reported\ mentors\ found\ in\ ontology}{\#\ unique\ mentors\ searched\ by\ students}$$

We planned to do a qualitative evaluation of our tool to see what aspects of our system users found insightful, helpful, unhelpful or needing improvement. We did not develop the app in time to test it with users and solicit their feedback. However, in the future, this sort of evaluation would provide additional insights beyond the quantitative evaluation.

A summative evaluation was not completed in this phase of testing as the ontology and tool are not deployed and available to students to use. This would be completed upon deployment and initial evaluation may look quantitatively and qualitatively at the use of the technology and its impact on research mentorship search.

Results

Structure of the final ontology is discussed more in depth in the methods section, but a snapshot of the entire class hierarchy is provided below. One design decision critical to our success was greatly simplifying relationships over the way they were originally defined in the VIVO ontology. For instance, in VIVO, a person cannot simply be the author of a document, they must be the member of an *Authorship,* which contains one or more authors and is itself a named instance, separate from all authors and the document itself. This level of granularity provides no useful functionality to our application, so we eliminated it.

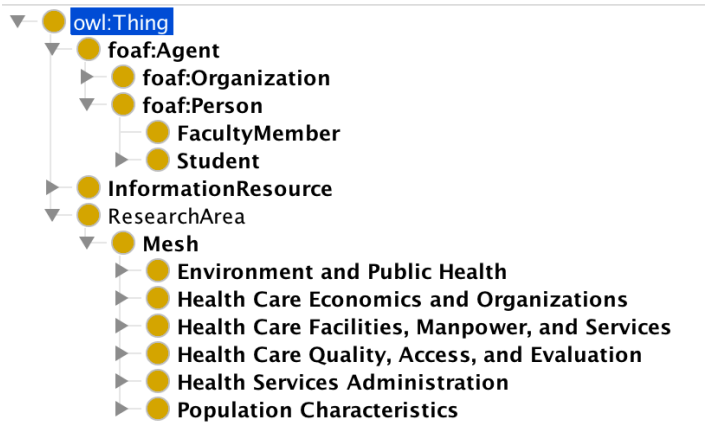

Fig. 3 – Partially expanded view of entire ontology

The table below summarizes the numbers of instances of various types we were able to scrape from the websites discussed above. Given the number of student and faculty instances we pulled, we were not able to link them to more document and MeSH terms because 1. The majority of the PubMed authors we pulled were affiliated with the medical school, and we did not scrape medical school faculty (although we plan to in a later iteration of the ontology) and 2. We only created instances for MeSH terms within the "Health Care" subcategory, due to computational limits and the size of the MeSH ontology.

DL Query	**Num. Instances**
FacultyMember	440
Student	611
AcademicArticle *(associated with min 1 author)*	615
MeSH *(associated with min 1 researcher)*	189
FacultyMember and authorOf min 1 AcademicArticle	144
Student and authorOf min 1 AcademicArticle	10
FacultyMember and hasResearchArea min 1 MeSH	120

Fig. 4 Number of instances scraped and added to ontology for various queries

The figures below depict some of our instances. We were able to scrape similar information for all of our faculty instances despite differing formatting across websites. For students however, only medical students have publicly available profiles. We created PhD and Undergraduate instances from https://biox.stanford.edu/person-group/***, and associated them with mentors based on the listed faculty advisors. Students of any kind are associated with faculty mentors if they have ever co-authored a paper, however, the information we have on medical students versus non-medical students generally differs.

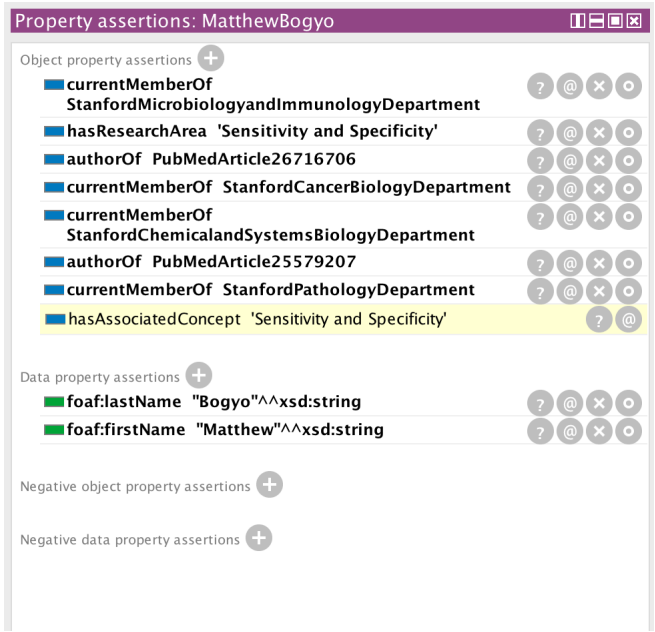

Fig. 5 Example FacultyMember instance

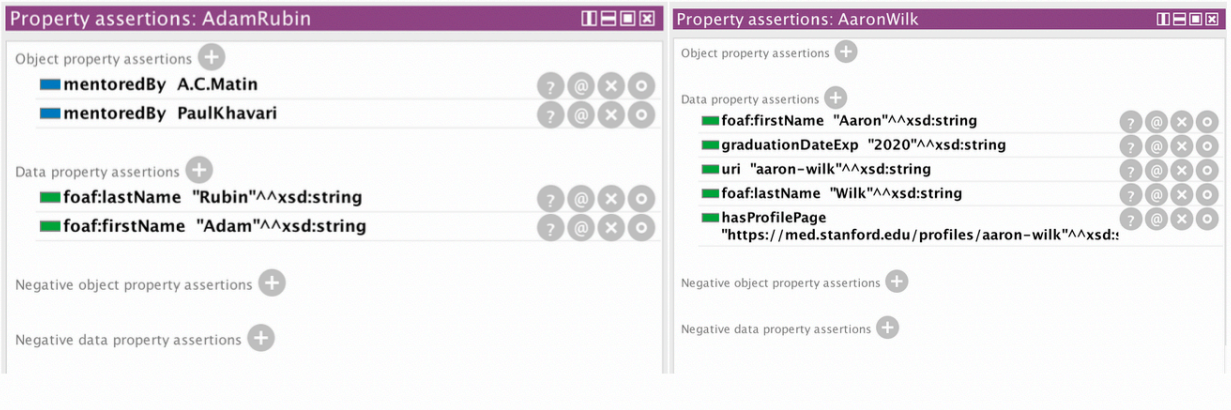

Fig. 6 Left: Example DoctoralStudent instance, Right: Example MedicalStudent instance

We have completed a beta version of a front-end application, with the intention that the final version can be used by Stanford students to explore faculty mentors with whom to do research. The app is built using *Tkinter* for python, and currently provides the ability to search for faculty based on affiliation (department, med school division, or research center), or by MeSH term. MeSH search in the beta version finds only faculty associated with specified term, but the final version will provide the option to also search based on the subtree of the specified term.

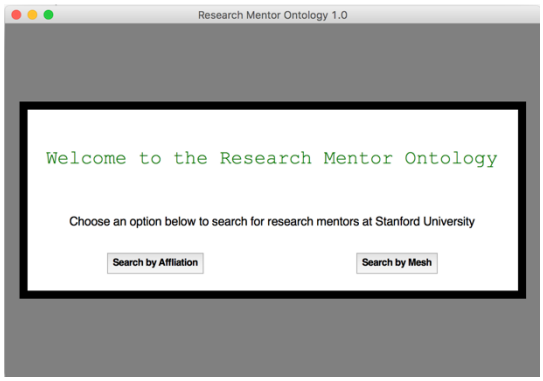

Fig. 7 Application home page

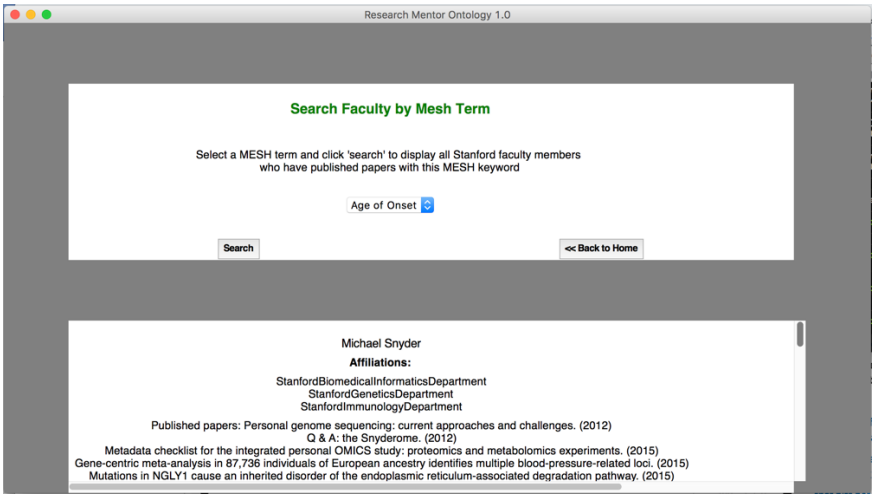

Fig. 8 Example search for faculty by MeSH term

Evaluation

Accuracy

Graduate students reported 38 pairs of research terms and lab groups. Out of these 38, 13 pairs were found in our ontology. This corresponds to an accuracy of 38%. This may be low for two reasons – our approach of associating research terms with faculty (by the MESH terms of their publications) needs improvement, and/or our evaluation method could have been improved. To describe the latter, the graduate students reported that the provided list of MESH terms were not applicable or vague in some cases. So a term was chosen but it did not necessarily accurately represent the research areas that they wanted to link to their lab. If we had provided the full list of MESH terms to the testers, their choices may have better aligned

with our method of associating research terms with professors. Additionally, searches based on research terms may need to return results from MESH terms from super or sub-ordinate classes of MESH terms to improve search results. This approach may return a lot of other researchers, but we may have to think of ranking based return approaches.

Completeness

In total, graduate students reported searching for 10 mentors. All 10 (100%) were found in the ontology. This suggests that our method of creating instances of faculty from scraping web directory may allow us to get complete set of faculty/lab groups. Of note, none of these 10 reported lab groups were multi-faculty lab groups and a separate evaluation may need to explore whether lab groups headed by multiple faculty are represented in our ontology.

Discussion and Future Work

In this project we show that publicly available data can be used to construct an ontology to help students search research mentors. We associate students and faculty with research areas based on MeSH terms associated with their publications (indexed in PubMed). We build a web application using the ontology to allow students to search research mentors by research areas or affiliation. There are multiple strengths of the approach we took. We merge two existing ontologies, one to represent research areas (MeSH ontology) and the other to model relationships between faculty, students and departments. This allows us to reason about research areas in a way that is already considered standard and accepted. We primarily use publications versus researcher's stated interests (on their individual websites) to parse their research areas. This not only saves us from parsing many distinct websites each with their own formatting it also uses publications as a truer signal of research interest. Finally, unlike other approaches that might try to parse research areas from text by relating it to ontologies like SNOMED, we rely on the MeSH indexing completed by PubMed. MeSH indexing in Pubmed is primarily manual and automated methods for associating research areas to publications have not been adopted yet. This supports our method of using existing indexing of researchers in their publications.

The single easiest improvement for our ontology and application would be to procure more complete instance data. If we eventually get student and faculty from Stanford IT, we can easily add all relevant faculty/student instances, without dealing with the inconsistent formatting and incomplete data of web-scraping. Student data in particular is not readily available online.

Moreover, due to time constraints we scraped only PubMed articles, but a more advanced iteration could scrape multiple academic article publication databases. Finally, due to computational constraints we restricted the MeSH terms we added to the 'Health Care' sub-hierarchy, but an ideal application would be able to search the full hierarchy.

We have considered available literature on the most important factors contributing to the success of a research mentor-mentee. However, due to time constraints, we have prioritized collecting information that is quantitative and easily available. Many sources suggest that a student and mentor sharing the same preference for mentorship style (i.e. the degree of hands-on help vs independence) is an important factor in research success, however, such data is difficult to obtain. A more advanced iteration of our tool could solicit ratings on aspects of mentorship style from current and past mentees of each faculty member, and display these as past of the data returned for each faculty member returned by a query. This is one area of future work. The other area is to allow lab groups to edit information like current members or recent publications. Our approach associates mentors and students based on co-publication. However, many students might work with a mentor and not co-publish; or new members would not be represented in our ontology since they would not have had time to publish yet. Thus, a separate login for the tool where a lab manager could edit certain information would be warranted.

We learned several things from completing this project. Firstly, there is a desire and need for supports to help people find mentorship. The existing tools for finding mentors at Stanford are limited. Yet public data and existing ontologies enable us to represent this information in a scalable way. Representing researcher information through ontologies of research areas and faculty/student relationships allows us to reason through queries like finding researchers of a particular research areas or by certain affiliation.

Division of Labor

Laura – creation of ontology in protégé, importing of instances using owlready, front-end app, contribution to the write-ups.
Sehj – contributed to data extraction by pulling Pubmed data, conducted the evaluation and contributed to the write-ups.

References

1. Sorkness CA, Pfund C, Ofili EO, et al. A new approach to mentoring for research careers: the National Research Mentoring Network. BMC Proc. 2017;11(Suppl 12):22.
2. Straus SE, Johnson MO, Marquez C, Feldman MD. Characteristics of successful and failed mentoring relationships: a qualitative study across two academic health centers. Acad Med. 2013;88(1):82-9.
3. https://facts.stanford.edu/academics/faculty/
4. https://facts.stanford.edu/research/
5. Fleming M, Burnham EL, Huskins WC. Mentoring translational science investigators. JAMA. 2012;308(19):1981-2.
6. Burnham EL, Fleming M. Selection of research mentors for K-funded scholars. Clin Transl Sci. 2011;4(2):87-92.
7. https://biosciences.stanford.edu/faculty/biosciences-faculty-database/
8. https://uit.stanford.edu/service/web
9. Mitchell, Stella & Chen, Shanshan & Ahmed, Mansoor & Lowe, Brian & Markes, Paula & Rejack, Nick & Corson-Rikert, Jon & He, Bing & Ding, Ying. (2011). The VIVO Ontology: Enabling Networking of Scientists.
10. Musen, M.A. The Protégé project: A look back and a look forward. AI Matters. Association of Computing Machinery Specific Interest Group in Artificial Intelligence, 1(4), June 2015. DOI: 10.1145/2557001.25757003
11. Nakikj D, Weng C. Extending VIVO ontology to represent research and educational resources in an academic biomedical informatics department. *Stud Health Technol Inform*. 2013;192:1206.
12. Lamy JB. Owlready: Ontology-oriented programming in Python with automatic classification and high level constructs for biomedical ontologies. Artif Intell Med. 2017;80:11-28.
13. Robert Hoehndorf Version of MeSH. Last uploaded: April 21, 2014. Accessed: http://bioportal.bioontology.org/ontologies/RH-MESH

Appendix

All code and owl files submitted on canvas.