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.*

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 model by surveying \*\*\*\*\*\* 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).

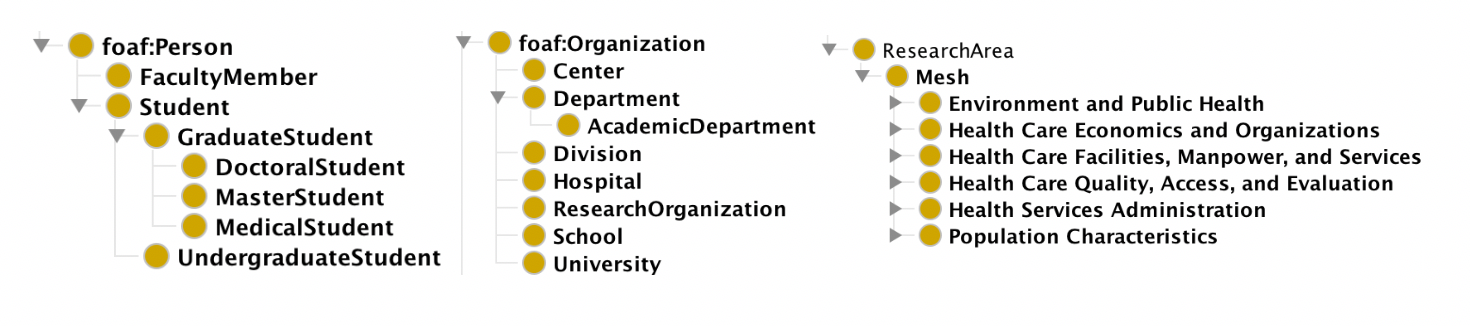


Fig. - View of most important classes in Protégé. Left: Person, encompassing FacultyMember and Student, Middle: Organization, encompassing departments (e.g. Biochemistry Department), divisions (e.g. Medical School, Division of Neurosurgery), and other research centers. Right: ResearchArea, currently 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 Collection***

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/](https://med.stanford.edu/profiles/browse?affiliations=capMdStudent) (med school faculty and students), [https://biox.stanford.edu/](https://biox.stanford.edu/person-group/undergraduate-fellows) (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 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***

Our project proposes to help students find research mentors by allowing them to find faculty mentors based on research interests and see and filter on supplementary information like publication data, current students, diversity etc. We will evaluate our ontology on a few different domains using an evaluation scheme for each one:

* Accuracy of research based search: We will select 3-5 labs and share a survey with their lab members soliciting which research terms (out of a bag of research terms in our lexicon) would best describe the lab to a prospective member searching for their lab. Then we will see if what proportion of those terms returns the target research labs that we solicited the information from.
* Qualitative Evaluation: We will conduct a qualitative evaluation of our tool with students who are seeking research mentors to see what aspects of our system they found insightful, unhelpful or needing improvement.

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. – 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.

|  |  |
| --- | --- |
| **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. 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/\*\*\*](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.

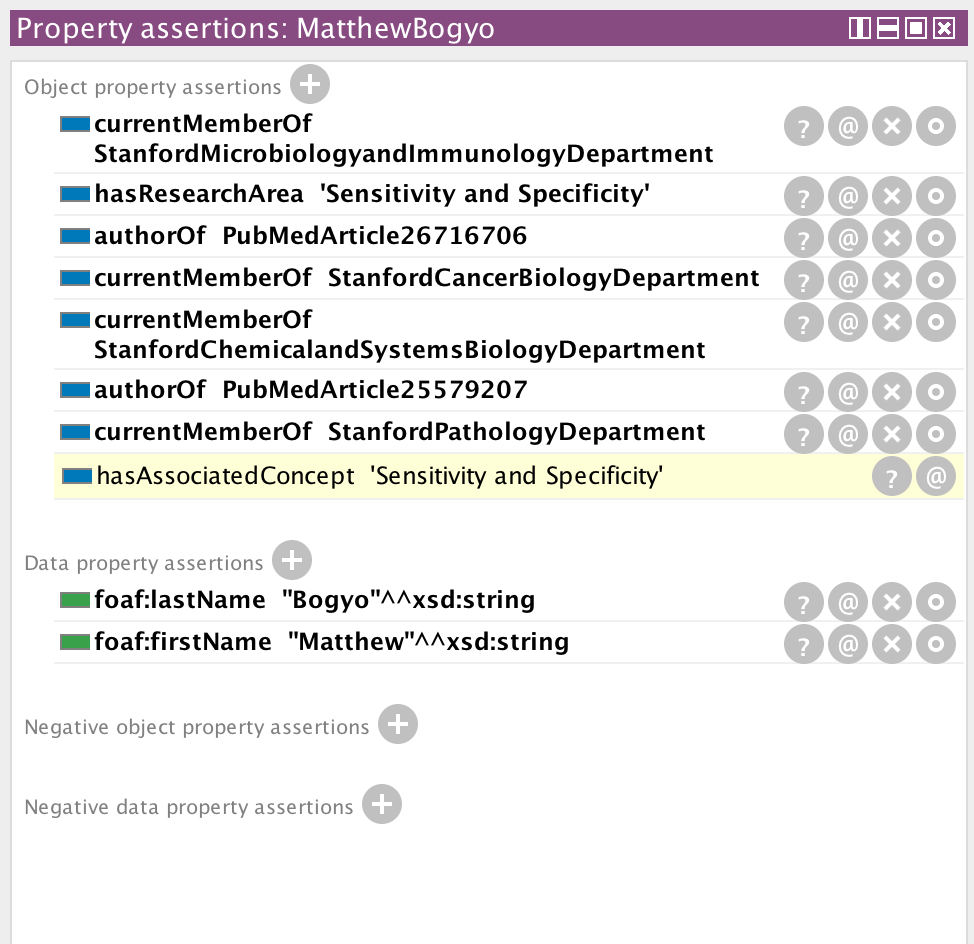


Fig. Example FacultyMember instance

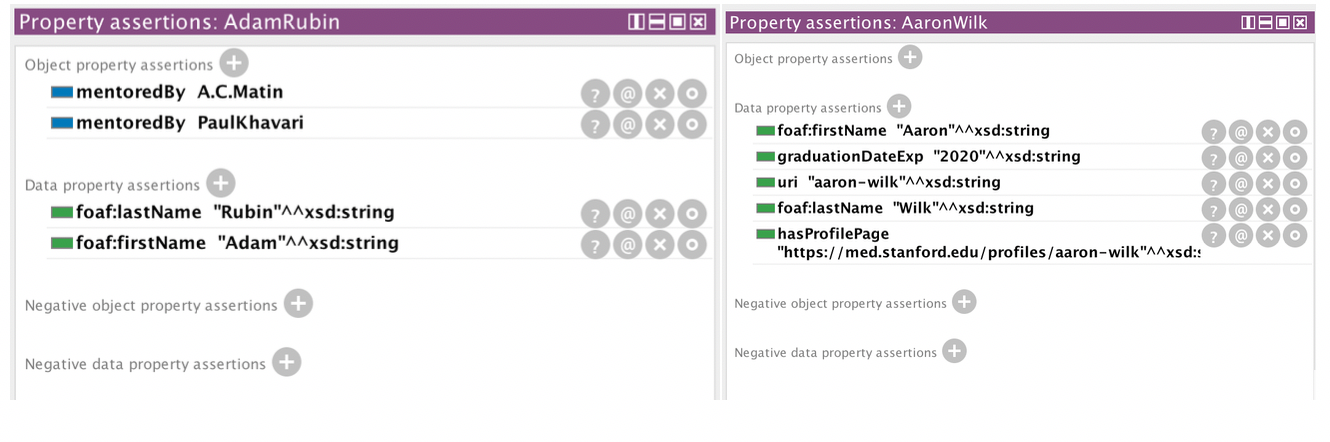


Fig. Left: Example DoctoralStudent instance, Right: Example MedicalStudent instance

\*\*\*\*\*Although they are not yet added to the ontology, we have completed PubMed scraping.  We have 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.





*Fig. 6  Example data in our database of authors and publications with publication meta-data*

Discussion and Future Work

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.

Division of Labor

Laura – creation of ontology in protégé, importing of instances using owlready, front-end app, contribution to final writeup

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](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4883684/). 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