The Research Mentor Ontology (Milestone): 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[1][2]. 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)[3][4]. 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[5][6]. The latter steps are time consuming, and students only engage in those steps with faculty whom they have

[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.

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[7]. 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[8]. 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

[7] https://biosciences.stanford.edu/faculty/biosciences-faculty-database/ Accessed Feb 27 2019
[8] https://uit.stanford.edu/service/web Accessed Feb 27 2019

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 will build a simple graphical user interface that performs SWRL reasoning over the model, and aggregates data to help students find and evaluate research groups.

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”. When a user searches any of these research areas, the application returns a list of all 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
    - Current students doing research with faculty member, degree program (undergraduate/masters/PhD) and contact information of each student
    - (Possible reach goal): Estimated likelihood that faculty member would accept a new researcher of specified degree program, based on past student researchers

3. Allow the user to filter and sort faculty mentors on metrics that quantify important mentor qualities:
    - Publication productivity
    - Number of collaborators and students on publications
    - Racial and gender diversity of mentor’s research group

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, AcademicPaper,* and *ResearchArea*. We model Stanford University, specific faculty members, students, and research areas as individuals. We use VIVO, the current predominant ontology for modeling faculty research[9], teaching, grants and research organizations for the majority of our ontology. However we make 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.

Data Collection

We have requested directory information for current students and faculty from Stanford IT; however, pending request, we scraped faculty data from the Stanford Biosciences web directory using *selenium* for python, and automatically created and added instances to the ontology using *owlready2*. From the Biosciences site, we scrape name, department, and a list of recent publications. We use the departments as some of our *ResearchArea* instances.

We define more granular *ResearchArea* instances, and locate further publications by scraping PubMed. We scrape all 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. First, we add MESH terms as *ResearchArea* instances, and articles as *AcademicPaper* instances. Finally, we associate scraped PubMed publications with existing *FacultyMember* and *Student* instances. We associate the MESH terms of each publication with the authors of that publication (both faculty and students) as a *ResearchArea*.

Problem-Solving Methods

We will build a graphical user interface using the *Tkinter* python package for the interface and *owlready2* to perform SWRL queries on our ResearchMentor ontology. The user can input

[9] "(PDF) The VIVO Ontology: Enabling Networking of Scientists." 26 Jan. 2019, https://www.researchgate.net/publication/228934646_The_VIVO_Ontology_Enabling_Networking_of_Scientists. Accessed 1 Mar. 2019.

queries according to the use cases described above, see results displayed, and optionally further filter the results. SWRL is necessary to perform many of the user queries because statistics such as ‘publication productivity over past 5 years’ are not directly stored in our model, but all published papers with publication dates are stored, and linked to all of their authors, so average rate can be calculated. [This section will be expanded in final report with specific actual queries we write].

Evaluation

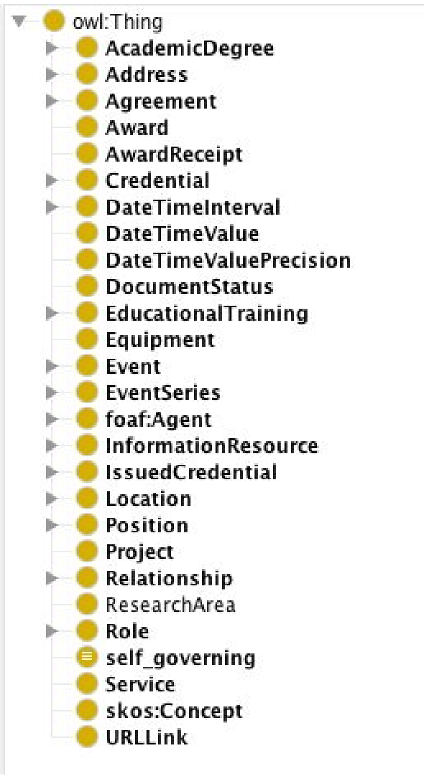
Our project proposes to help students find research mentors by allowing them to find faculty mentors based on research interests and see and filter on supplementary information like publication data, current students, diversity etc. We will evaluate our ontology on a few different domains using an evaluation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

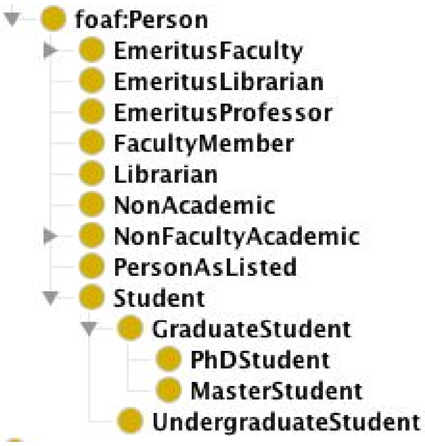
- Qualitative Evaluation: We will conduct a qualitative evaluation of our tool with students who are seeking research mentors to see what aspects of our system they found insightful, unhelpful or needing improvement.

Results

So far, we have built our ontology in *Protege*, and imported all Stanford biosciences faculty (listed on https://biosciences.stanford.edu/faculty/biosciences-faculty-database/) into the model. The figures below provide a snapshot of pieces on the ontology:



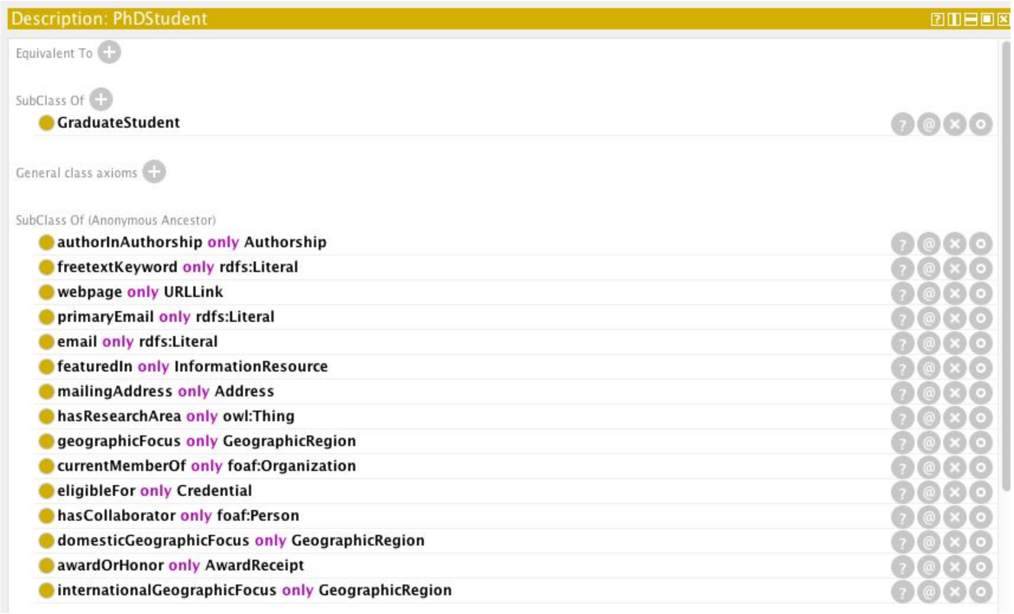
- owl:Thing
  - AcademicDegree
  - Address
  - Agreement
  - Award
  - AwardReceipt
  - Credential
  - DateTimeInterval
  - DateTimeValue
  - DateTimeValuePrecision
  - DocumentStatus
  - EducationalTraining
  - Equipment
  - Event
  - EventSeries
  - foaf:Agent
  - InformationResource
  - IssuedCredential
  - Location
  - Position
  - Project
  - Relationship
  - ResearchArea
  - Role
  - self_governing
  - Service
  - skos:Concept
  - URLLink



- foaf:Person
  - EmeritusFaculty
  - EmeritusLibrarian
  - EmeritusProfessor
  - FacultyMember
  - Librarian
  - NonAcademic
  - NonFacultyAcademic
  - PersonAsListed
  - Student
    - GraduateStudent
      - PhDStudent
      - MasterStudent
    - UndergraduateStudent

Fig. 1 Full Ontology in Protege, contains many VIVO classes that we do not use

Fig. 2 Expanded 'Person' subtree, important classes are FacultyMember and Student + subclasses



Description: PhDStudent

Equivalent To

SubClass Of

GraduateStudent

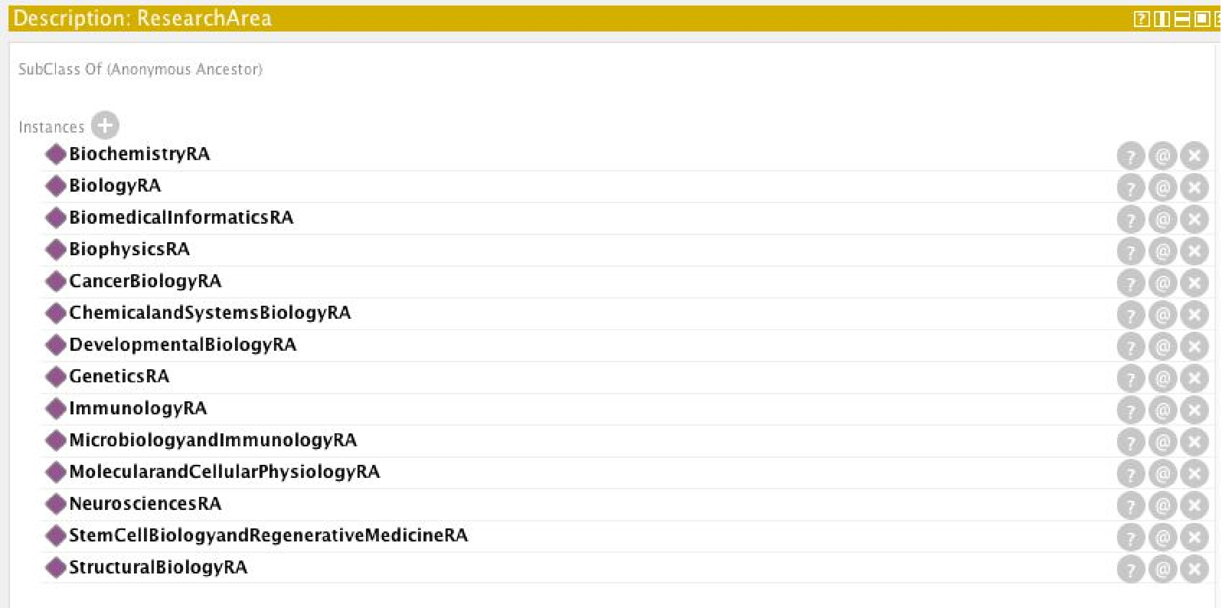
General class axioms

SubClass Of (Anonymous Ancestor)

authorInAuthorship only Authorship
freetextKeyword only rdfs:Literal
webpage only URLLink
primaryEmail only rdfs:Literal
email only rdfs:Literal
featuredIn only InformationResource
mailingAddress only Address
hasResearchArea only owl:Thing
geographicFocus only GeographicRegion
currentMemberOf only foaf:Organization
eligibleFor only Credential
hasCollaborator only foaf:Person
domesticGeographicFocus only GeographicRegion
awardOrHonor only AwardReceipt
internationalGeographicFocus only GeographicRegion

Fig. 3 Example class definition - PhDStudent

The figures below depict some of our instances. So far we have added faculty members and research topics from the biosciences website, but we have not imported and linked the PubMed articles and MESH terms.



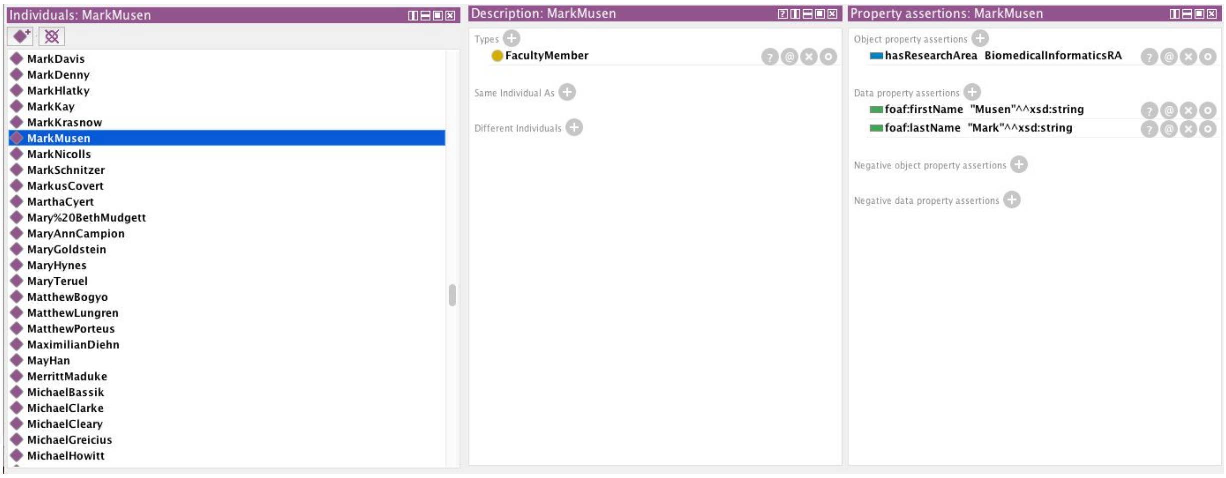
Description: ResearchArea

SubClass Of (Anonymous Ancestor)

Instances

- BiochemistryRA
- BiologyRA
- BiomedicalInformaticsRA
- BiophysicsRA
- CancerBiologyRA
- ChemicalandSystemsBiologyRA
- DevelopmentalBiologyRA
- GeneticsRA
- ImmunologyRA
- MicrobiologyandImmunologyRA
- MolecularandCellularPhysiologyRA
- NeurosciencesRA
- StemCellBiologyandRegenerativeMedicineRA
- StructuralBiologyRA

Fig. 4 List of ResearchArea instances, scraped from biosciences website



Individuals: MarkMusen

- MarkDavis
- MarkDenny
- MarkHlatky
- MarkKay
- MarkKrasnow
- MarkMusen
- MarkNicolls
- MarkSchnitzer
- MarkusCovert
- MarthaCyert
- Mary%20BethMudgett
- MaryAnnCampion
- MaryGoldstein
- MaryHynes
- MaryTeruel
- MatthewBogyo
- MatthewLungren
- MatthewPorteus
- MaximilianDiehn
- MayHan
- MerrittMaduke
- MichaelBassik
- MichaelClarke
- MichaelCleary
- MichaelGreicius
- MichaelHowitt

Description: MarkMusen

Types
FacultyMember

Same Individual As

Different Individuals

Property assertions: MarkMusen

Object property assertions
hasResearchArea BiomedicalInformaticsRA

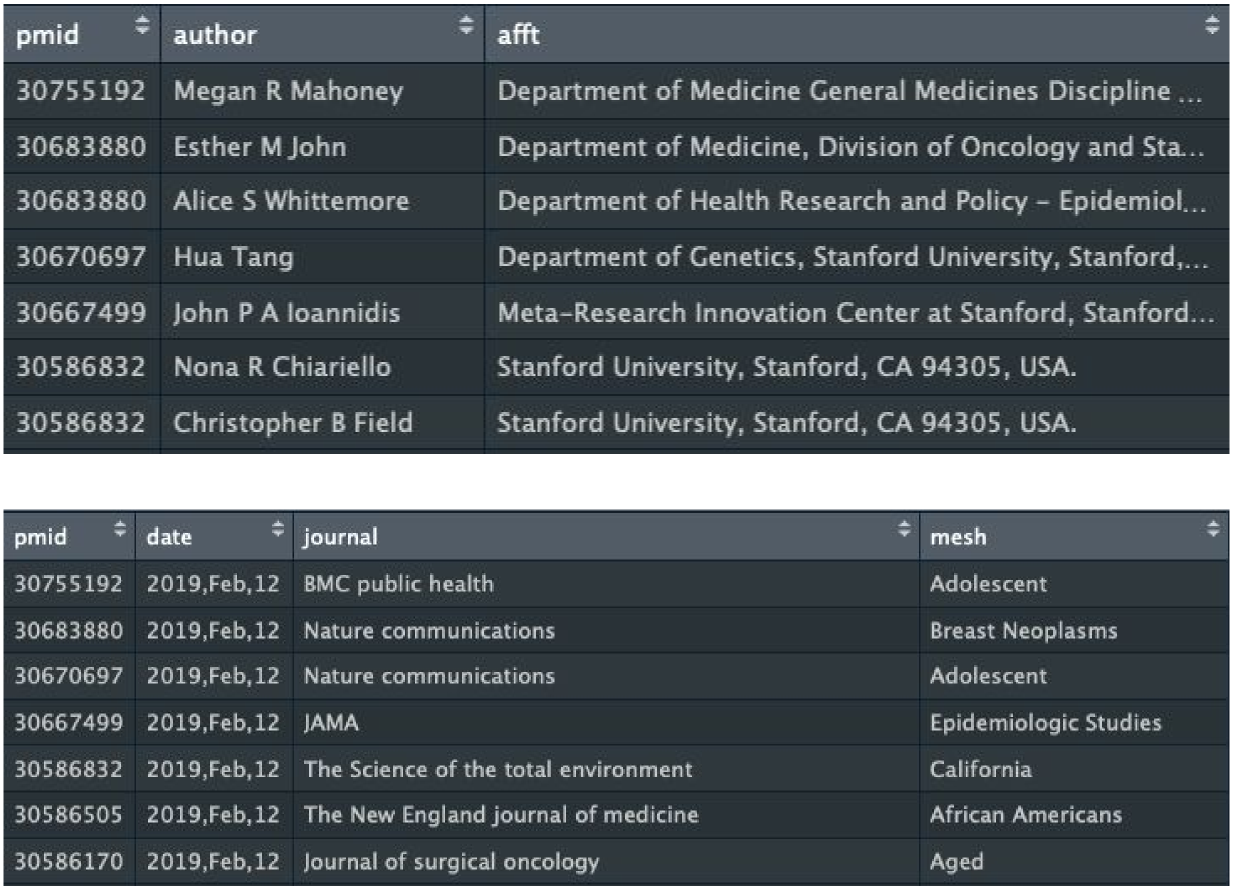
Data property assertions
foaf:firstName "Musen"^^xsd:string
foaf:lastName "Mark"^^xsd:string

Negative object property assertions

Negative data property assertions

Fig. 5 Partial list of individuals, example of FacultyMember instance Mark Musen

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.



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

Discussion and Future Work

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

We collaborated on writing this milestone. We have both been working on everything, but Sehj primarily has been scraping data and Laura has primarily been working with the ontology in Protege.

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.