Semiautomated process for generating knowledge graphs for marginalized community doctoral-recipients

Neha Keshan, Kathleen Fontaine and James A. Hendler
Department of Computer Science, Tetherless World Constellation,
Rensselaer Polytechnic Institute, Troy, NY

Received 28 February 2022
Revised 1 June 2022
Accepted 5 July 2022

Abstract

Purpose – This paper aims to describe the "InDO: Institute Demographic Ontology" and demonstrates the InDO-based semiautomated process for both generating and extending a knowledge graph to provide a comprehensive resource for marginalized US graduate students. The knowledge graph currently consists of instances related to the semistructured National Science Foundation Survey of Earned Doctorates (NSF SED) 2019 analysis report data tables. These tables contain summary statistics of an institute's doctoral recipients based on a variety of demographics. Incorporating institute Wikidata links ultimately produces a table of unique, clearly readable data.

Design/methodology/approach – The authors use a customized semantic extract transform and loader (SETLr) script to ingest data from 2019 US doctoral-granting institute tables and preprocessed NSF SED Tables 1, 3, 4 and 9. The generated InDO knowledge graph is evaluated using two methods. First, the authors compare competency questions' sparql results from both the semiautomatically and manually generated graphs. Second, the authors expand the questions to provide a better picture of an institute's doctoral-recipient demographics within study fields.

Findings – With some preprocessing and restructuring of the NSF SED highly interlinked tables into a more parsable format, one can build the required knowledge graph using a semiautomated process.

Originality/value – The InDO knowledge graph allows the integration of US doctoral-granting institutes demographic data based on NSF SED data tables and presentation in machine-readable form using a new semiautomated methodology.

Keywords Semiautomation process, Knowledge graphs, Institute demographics, Graduate mobility, NSF doctoral recipients survey data

Paper type Research paper

This work is part of the vision of "Building a Social Machine for Graduate Mobility". We would like to thank Dean Stanley Dunn, Dean of Graduate Education, who provided expert insights into this issue. We would like to thank the members of the Tetherless World Constellation Lab at Rensselaer Polytechnic Institute, especially John S. Erickson and Jamie P. McCusker, who provided insights and expertise that greatly assisted this research. This work was funded in part by the RPI-IBM AI Research Collaboration, a member of the IBM AI Horizons network.

International Journal of Web Information Systems
Vol. 18 No. 5/6, 2022
pp. 413-431
Emerald Publishing Limited
1744-0084
DOI 10.1108/IJWIS-02-2022-0046

1. Introduction

Newly minted doctoral students face a long-standing problem of "what comes next?". This transition from being a student toward their chosen career path is referred to as graduate mobility (Keshan, 2021). The preparation for graduate mobility does not start when one is approaching graduation but rather much earlier, perhaps even as early as the time of program selection. To help students have a smooth transition from graduate school to their career, it is important for them to have an adequate amount of information for doctoral graduate school selection. The information should include the demographics of past doctoral recipients and the career paths they chose. Students can use this information along with the general program ranking to make an informed decision about which graduate program to join. Therefore, the question of "what comes next" is connected to the question of "where is the best for me?" during a doctoral program selection (Keshan *et al.*, 2021). In general, doctoral programs are challenging for all students but can be especially challenging for students from marginalized communities – groups of students traditionally under-represented based on ethnicity, race, language, gender identity, age, physical ability and/or immigration status (Gay, 2004; Sevelius *et al.*, 2020). It has been shown that marginalized students have to go the extra mile to prove their worth.

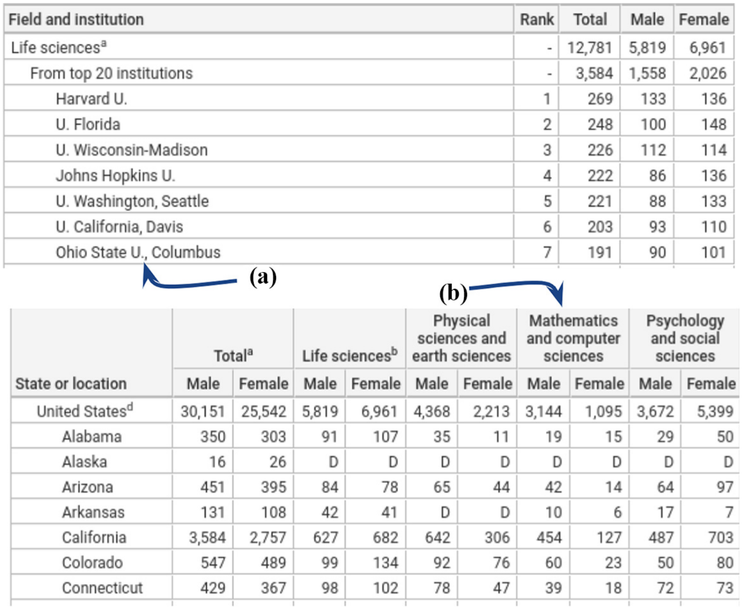
Previous work (Keshan *et al.*, 2021) proposed an Institute Demographic Ontology (InDO) designed to help with this problem. The ontology was mainly generated manually using a traditional methodology (Kendall and McGuiness, 2019). This paper builds on that work by describing a new, semiautomated process for generating an Institute Demographic knowledge graph, based on the InDO ontology, to integrate the various NSF SED survey results statistical data (Foley, 2021). Notably, National Science Foundation (NSF) recently (Dec 2021) launched the "Survey of Earned Doctorates Restricted Data Analysis System" (SED RDAS), which allows users to create their own tables for SED data from 2017 to 2020. In this restrictive model, security protocols in the NSF system do not allow the user to acquire institute-specific demographics with respect to the year. However, the institute-specific data is available through the NSF website as part of their SED analysis results across multiple tables for the years 1958 to 2020. These tables (Figure 1) can be integrated with one another using semantic techniques without compromising privacy to make the statistical data more machine-readable and, therefore, more accessible, providing a more comprehensive picture of any US doctorate-granting institute's demographics. This system integrates the available institute data from the provided results table without compromising student privacy.

In this paper, we describe a semiautomated linked-data representation of the NSF SED statistical data, knowledge representation of this statistical, demographic data and the usefulness of linking it with Wikidata [1]. Wikidata is a free and open knowledge base that can be processed by both humans and machines. The content of Wikidata, available under a free creative commons license, is interlinkable to other open data sets on the linked data Web. Our current InDO-based semiautomatically generated knowledge graph includes data points from Tables 1, 3, 4 and 9 of the published NSF SED 2019 analysis results. One hundred and ninety-four of the 448 doctoral-granting US institutes have their respective Wikidata nodes added to allow users to access our resources in conjunction with other linked data already available on the Web. Finally, as part of the evaluation, we compared blazegraph workbench results obtained from the semiautomatically generated knowledge graph and the manually generated knowledge graph. We also added new competency questions to provide a better picture of an institute's demographic based on broad study fields. The consistency of the semiautomatically generated InDO knowledge graph was checked using Pellet in Protégé – 5.5.0.

The paper is organized as follows: Section 2 discusses the related literature and background concepts used for this work. Section 3 expands on the different methods used as part of the pipeline. Section 4 describes the obtained results, and Section 5 compares the semiautomated generated knowledge graph with the manually generated one. Section 6 discusses the challenges and limitations of our work while describing how we address them, and Section 7 concludes by restating the importance of the work and how it leads to future research ventures.

2. Background work

Even when a rich profusion of research on graduate students can be found in the social science domain, there is comparatively little available on marginalized US graduate students. Most of the available research focuses on graduate students of color and females, and the major theme across research in at least the past three decades has been the role of advising in the retention and success of historically marginalized graduate students in the US (Blackwell, 1987; Abatso *et al.*, 1987; Frierson, 1990; Willie, 1991; Terenzini, 1996; Brown *et al.*, 1999; Brown, 2000; Gay, 2004; Golde and Dore, 2004; Girves *et al.*, 2005; Thomas *et al.*, 2007; Luna and Prieto, 2009; Pau, 2009; Fuhrmann *et al.*, 2011; McGee *et al.*, 2012; Sauermann and Roach, 2012; Gibbs *et al.*, 2014; McCallum, 2017; Roach and Sauermann, 2017; Sevelius *et al.*, 2020; and Ullrich *et al.*, 2021). The research is even more sparse when looking at minority groups based on other factors such as functional limitations (Jain *et al.*, 2020; Tamjeed *et al.*, 2021), citizenship (Zhou, 2010) or age (Rose, 2005). Additionally, the existing research highlights the lack of available references and reference points – someone with a similar background who has experienced or is currently experiencing the doctoral process,



Field and institution	Rank	Total	Male	Female
Life sciences[a]	-	12,781	5,819	6,961
From top 20 institutions	-	3,584	1,558	2,026
Harvard U.	1	269	133	136
U. Florida	2	248	100	148
U. Wisconsin-Madison	3	226	112	114
Johns Hopkins U.	4	222	86	136
U. Washington, Seattle	5	221	88	133
U. California, Davis	6	203	93	110
Ohio State U., Columbus	7	191	90	101

	Total[a]		Life sciences[b]		Physical sciences and earth sciences		Mathematics and computer sciences		Psychology and social sciences	
State or location	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female
United States[d]	30,151	25,542	5,819	6,961	4,368	2,213	3,144	1,095	3,672	5,399
Alabama	350	303	91	107	35	11	19	15	29	50
Alaska	16	26	D	D	D	D	D	D	D	D
Arizona	451	395	84	78	65	44	42	14	64	97
Arkansas	131	108	42	41	D	D	10	6	17	7
California	3,584	2,757	627	682	642	306	454	127	487	703
Colorado	547	489	99	134	92	76	60	23	50	80
Connecticut	429	367	98	102	78	47	39	18	72	73

Figure 1.
(a) The top screenshot depicts a portion of the NSF SED 2019 Table 4. This table consists of the “Top 20 doctorate-granting institutions ranked by total number of doctorate recipients, by broad field of study and sex”; (b) the bottom screenshot depicts a portion of the NSF SED 2019 Table 6. This table consists of the “Doctorates recipients, by state or location of doctorate institutions, broad field of study, and sex”

whether that be a student or a faculty member – for marginalized graduate students in US doctoral-granting institutes (Keshan *et al.*, 2021).

Though not referred to specifically as a lack of a reference point, social scientists have been addressing similar issues through their research on the role of mentorship in marginalized graduate students' development for at least the past 30 years. For example, Brown *et al.* (1999) state that one needs to emphasize the process of becoming a profession rather than the act of being that profession. They discuss the importance of students building the required knowledge base, skills and behaviors to become a successful professional under the guidance of an advisor. This process of learning and mentorship is backed by historical examples of individuals entering professions through apprenticeship. Terenzini (1996) brings to our attention the way students "learn – and in many cases internalize – their mentors' intellectual orientations, value systems, criteria, and standards about what constitutes appropriate topics and good research". Therefore, due to their ideological and experiential influence, advisers become one of the students' most important reference points in understanding and adapting to the scientific research system. It has been shown that effective mentoring relationships can be used to encourage underrepresented students to go to graduate schools (Luna and Prieto, 2009). Gay (2004) continues to bring to our attention how the concept of "giving back to the community" plays an influential role for these students. We can refer to such "giving back" as becoming the reference point for the next generation of graduate students. Even when researchers have been discussing the lack of reference points in different forms, the authors Thomas *et al.* (2007) remind us of the unique challenges marginalized students face – the "societal pressures and frequent negative stereotypes as well as usually being a racial token in their department, program or even college". Hence, it becomes more important to create a system to offer reference points to marginalized graduate US students.

While the literature shows a lot of inter-domain research being conducted, the work of Pau (2009) is the closest work to ours in terms of looking at the social science problem of marginalized students from a computer science perspective. The author focuses on factors influencing female students' perception of computing and computing careers (Pau, 2009). Even when the author does tackle a social science problem from a computer science perspective, the research methods followed were parallel to the ones dominantly used in the social science domain: interviews and surveys. Because "reference points" could be either an individual or information available on the Web in a structured, semistructured and unstructured manner, connecting and combining the information using semantics becomes an obvious choice. Literature shows how semantics has helped elevate required information by combining such resources in various domains (Singhal, 2012; Madhavan *et al.*, 2008; Dong, 2019; Krishnan, 2018; High, 2012).

As part of a semantic Web benchmark, Lehigh University demonstrated a unifying system of university information through the creation of the University Ontology [2]. This ontology allows access to university concepts, including information, affiliation and department for faculty, students and staff, but it does not define the demographic of an individual, nor does it include a mechanism to store the number of doctoral recipients of an institute in a given year. To build in/include this important demographic information, we looked at two prominent ontologies: NCI Thesaurus (NCIT) open biological and biomedical ontologies Edition (Balhoff *et al.*, 2017) and the Children's Health Exposure Analysis Resource (CHEAR) ontology (Balshaw *et al.*, 2017). Even though the NCIT ontology is centered on the cancer domain and the CHEAR ontology is centered on childrens' health with respect to environmental exposure, the two ontologies provide clear guidelines as to how to incorporate demographic concepts into knowledge graphs. Thus, there are multiple resources that could help store and present a section of the data for marginalized community students. However, we could not find one existing ontology that could either directly or, by extension, fulfill the requirement of a comprehensive resource. Hence, we created InDO, starting by reusing the already existing SemanticScience Integrated Ontology (SIO) (Dumontier *et al.*, 2014). Even though the InDO ontology provides us with a helpful taxonomy, it is important to create a knowledge graph by adding instances to the ontology for maximizing the ontologies' utility. The different applications, problems, challenges, refinements and evaluation of knowledge graphs using structured, semistructured and unstructured data can be found in recent knowledge graph survey papers (Paulheim, 2017; Guo *et al.*, 2020; Tiwari *et al.*, 2021; Ji *et al.*, 2021). These surveys outline and explain existing methods of creating and evaluating knowledge graphs from tables.

Some of the tools used to generate knowledge graphs are RDFLib, customized Python programming and different extract-transform-load (ETL) tools. We would like to highlight two spreadsheets/tables to rdf tools – Cellfie and Semantic Extract Transform and Load-er (SETLr) (discussed in Section 3.3). Cellfie is a Protege Desktop Plugin for importing spreadsheet data into owl ontology [3], whereas SETLr (McCusker *et al.*, 2018) is a powerful tool for creating RDF from tabular sources and provides a platform for semantic extract, transform and load workflows. As Celfie is made for a particular platform and has not been demonstrated to be used as a part of a larger build approach, its automatability is limited. On the other hand, SETLr can be incorporated into an automated build system, allowing it to be deployed directly to production knowledge graphs or part of larger build approaches. Because this work is part of a larger build approach, hence, we chose SETLr to generate our semiautomated knowledge graph based on InDO. Currently, the ingested data is from the multiple tables published by the US NSF [4] as part of their annual NSF SED report. NSF has been conducting these annual surveys since the year 1957 and publishing an aggregated summary report both in terms of data tables and an analysis report. These downloadable tables (Figure 1) are organized such that the human eye can analyze the data quickly but are highly complex and difficult for a machine to parse. Hence, we incorporated the summary statistics aggregation method (McCusker *et al.*, 2019) used by Chari *et al.* (2019) to make study cohorts visible using knowledge graphs. Using this method allowed us to incorporate the summary statistical data provided in the published NSF SED 2019 survey data tables. We also preprocessed these tables to convert them into a machine-readable format for comparatively easy ingestion.

The literature we reviewed helped us to see the current gaps in the field and use computer science semantic tools, including ontologies and knowledge graphs, to solve this giant social science problem: the lack of reference points for historically marginalized US graduate students. Now that we have created the InDO ontology, we will build upon our previous work by demonstrating a semiautomated process of generating a knowledge graph and connecting it with Wikidata, a highly recognized linked resource. We will also demonstrate how the challenges during the conversion and ingestion of the NSF SED summary statistical data were handled during the process to make the resource a valuable one.

3. Methodology

Figure 2 depicts the overview of the institute's demographic data harmonization process, from extracting them from published Web pages to flattening them and then matching them with the corresponding classes, followed by ingestion into the knowledge graph through the customized SETLr code. This is then followed by querying the semiautomatically generated knowledge graph to acquire the required knowledge. We will now discuss the overview of the InDO ontology used as the taxonomy for this work in Section 3.1, followed by the preprocessing step and data modeling of the ingested NSF tables in Section 3.2, and learn more about the SETLr tool and its usage in this study in Section 3.3.

3.1 Overview of Institute Demographic ontology

The InDO captures the demographics of an institute's doctoral recipients, broad and fine field of study and pursues career paths (Figure 3). The terminology is structured in a five-leveled hierarchy that provides room for the most abstract top level (basic components used to describe an Institute's demographics) to the most concrete bottom levels (particular graduate programs offered by the institute) with the corresponding provenance. The terminology structure is highly influenced by the NSF SED data tables and the CHEAR ontology. Both the ontology and the use case knowledge graph are created using Protégé – 5.5.0.

The ontology consistency is checked using the in-built Hermit reasoner (this reasoner helps determine if the input ontology is consistent or not while identifying subsumption relationships between classes), whereas the knowledge graph consistency is checked using the in-built Pellet reasoner (this reasoner does not only checks if the ontology contains contradictory facts but is built to check the possibilities of instances and types of individuals added).

3.2 Preprocessing of National Science Foundation tables

NSF SED 2019 (Foley, 2021) (henceforth referred to as NSF results) has 72 tables [5], and each table has a unique structure, with many being in a highly interlinked (multihierarchy levels) format. NSF results Table 2 shows an increase in the US doctoral-granting institutions from 283 institutes in 1973 to 448 institutes in 2019. We created a table (Figure 4) consisting of all the 2019 doctoral-granting US institutions, available in data tables and their respective Wikidata links. Incorporating Wikidata in our comprehensive resource system allowed us to access significant amounts of other linked-data sources connected via the Wikipedia links.

As mentioned, these tables being in a highly interlinked format, could not be easily processed by machines, so we first had to preprocess these tables and restructure them into a parsable format. We used the most logical method to convert each NSF SED complex data table to a simple table format keeping the essence intact. For example, NSF results Table 3 – "Top 50 doctorate-granting institutions ranked by a total number of doctorate recipients, by

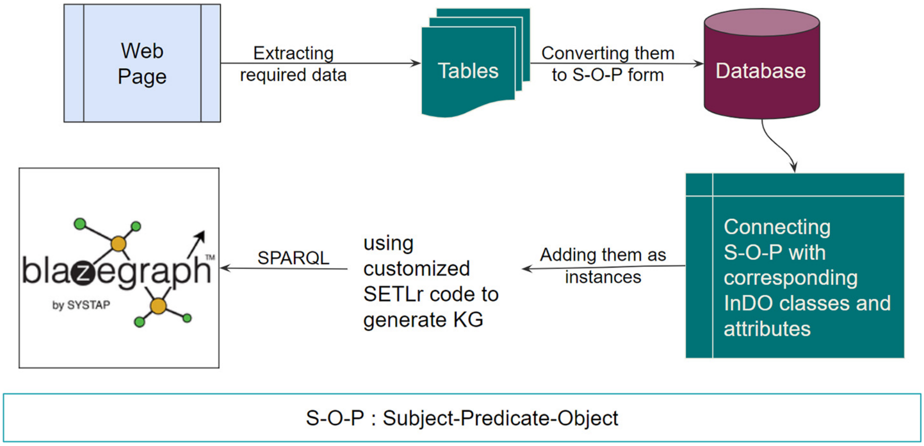
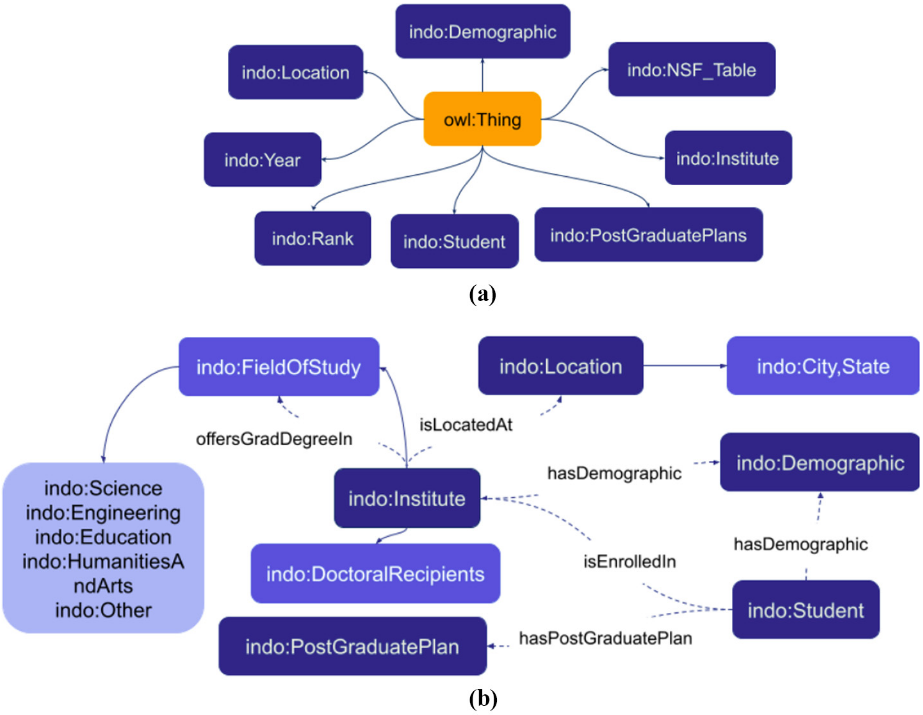
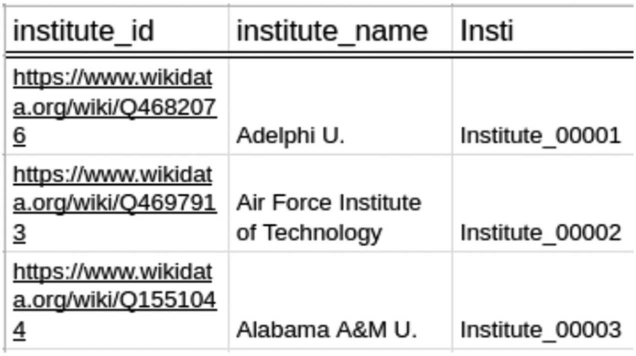

Web Page
Extracting required data
Tables
Converting them to S-O-P form
Database
Connecting S-O-P with corresponding InDO classes and attributes
Adding them as instances
using customized SETLr code to generate KG
SPARQL
blazegraph
by SYSTAP
S-O-P : Subject-Predicate-Object

Figure 2.
An overview of the semiautomated process of adding data to the knowledge graph using InDO and customized SETLr code



Notes: The arrows depict the subclass relationship between classes, whereas the dashed arrows depict the property association between two classes. It also demonstrates the broad field of studies that we intend to include in our comprehensive resource for graduate students

Figure 3. (a) A conceptual diagram depicting the top-level classes of our Institute Demographic Ontology; (b) a conceptual overview of our Institute Demographic Ontology's institute class



institute_id	institute_name	Insti
https://www.wikidata.org/wiki/Q4682076	Adelphi U.	Institute_00001
https://www.wikidata.org/wiki/Q4697913	Air Force Institute of Technology	Institute_00002
https://www.wikidata.org/wiki/Q1551044	Alabama A&M U.	Institute_00003

Figure 4. A snapshot of the newly created table of all the 2019 doctoral granting US institutes

sex: 2019" has a column for the total number of doctorate recipients and one column for each male and female doctorate recipients for the top 50 US doctoral-granting institutions for 2019. We merged these three columns into a single column that we called "Sex." This required us to add the class "All" under "indo:Sex" stating the total number of doctoral recipients. The provenance of this summary statistical data was added through a column stating the NSF results it belonged to (Figure 5). NSF results Table 4 – “Top 20 doctorate-granting institutions ranked by the total number of doctorate recipients, by broad field of study and sex: 2019” was restructured in a similar format. In addition to creating a single “Sex” column for all institutions as well as adding the NSF provenance table column, we introduced a separate study field column and replicated the information for each of the Top 20 institutes in the NSF results as depicted in Figure 6. These small tweaks allowed us to convert complex, hierarchical and highly interlinked tables into a simple table format with all the required information intact. This also demonstrates the method we can use to include other gender identities into our ontology which currently consists of just the Male and Female tags used in the NSF SED data tables.

To demonstrate the value of semantic tools for this work, we ingested Table 9 – “Top 20 doctorate-granting institutions ranked by number of minority U.S. citizen and permanent resident doctorate recipients, by ethnicity and race: 5-year total, 2015–19” in our knowledge

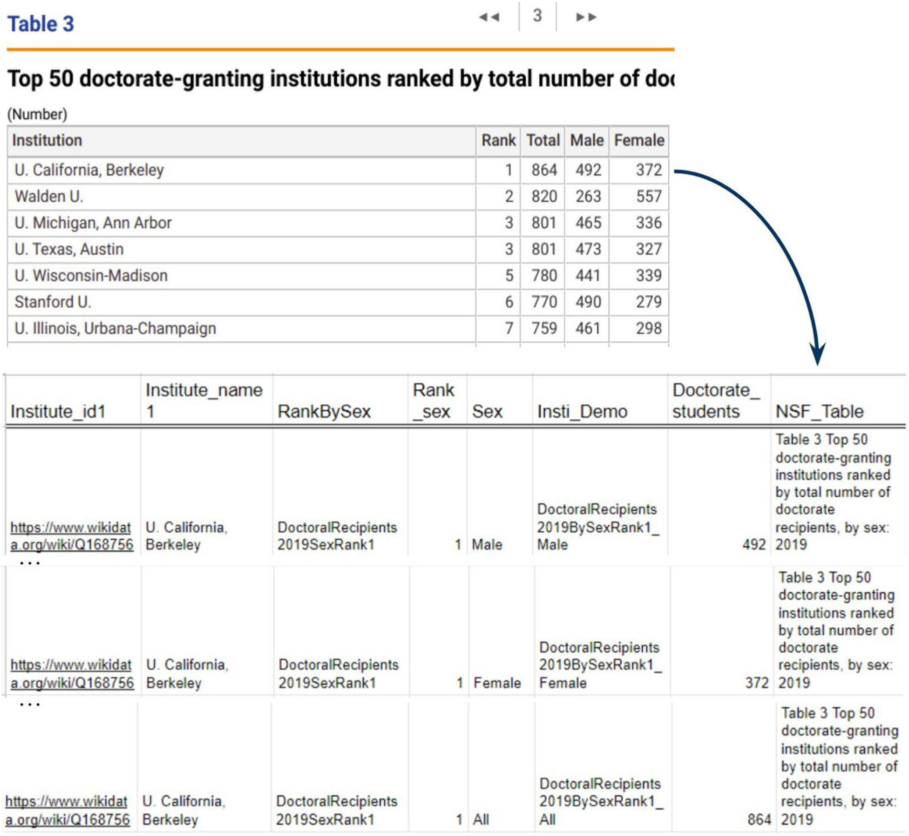


Table 3

Top 50 doctorate-granting institutions ranked by total number of do

(Number)

Institution	Rank	Total	Male	Female
U. California, Berkeley	1	864	492	372
Walden U.	2	820	263	557
U. Michigan, Ann Arbor	3	801	465	336
U. Texas, Austin	3	801	473	327
U. Wisconsin-Madison	5	780	441	339
Stanford U.	6	770	490	279
U. Illinois, Urbana-Champaign	7	759	461	298

Institute_id1	Institute_name 1	RankBySex	Rank _sex	Sex	Insti_Demo	Doctorate_ students	NSF_Table
https://www.wikidata.org/wiki/Q168756	U. California, Berkeley	DoctoralRecipients 2019SexRank1	1	Male	DoctoralRecipients 2019BySexRank1_ Male	492	Table 3 Top 50 doctorate-granting institutions ranked by total number of doctorate recipients, by sex: 2019
...							
https://www.wikidata.org/wiki/Q168756	U. California, Berkeley	DoctoralRecipients 2019SexRank1	1	Female	DoctoralRecipients 2019BySexRank1_ Female	372	Table 3 Top 50 doctorate-granting institutions ranked by total number of doctorate recipients, by sex: 2019
...							
https://www.wikidata.org/wiki/Q168756	U. California, Berkeley	DoctoralRecipients 2019SexRank1	1	All	DoctoralRecipients 2019BySexRank1_ All	864	Table 3 Top 50 doctorate-granting institutions ranked by total number of doctorate recipients, by sex: 2019

Notes: The converted U. California, Berkeley information as seen in the bottom screenshot was used to ingest the data along with the provenance in InDO knowledge graph using the customized SETLr script

Figure 5.
The top screenshot depicts the view of the NSF SED Survey Table 3 as seen on the NSF website using the view option and the bottom screenshot depicts the new processed table

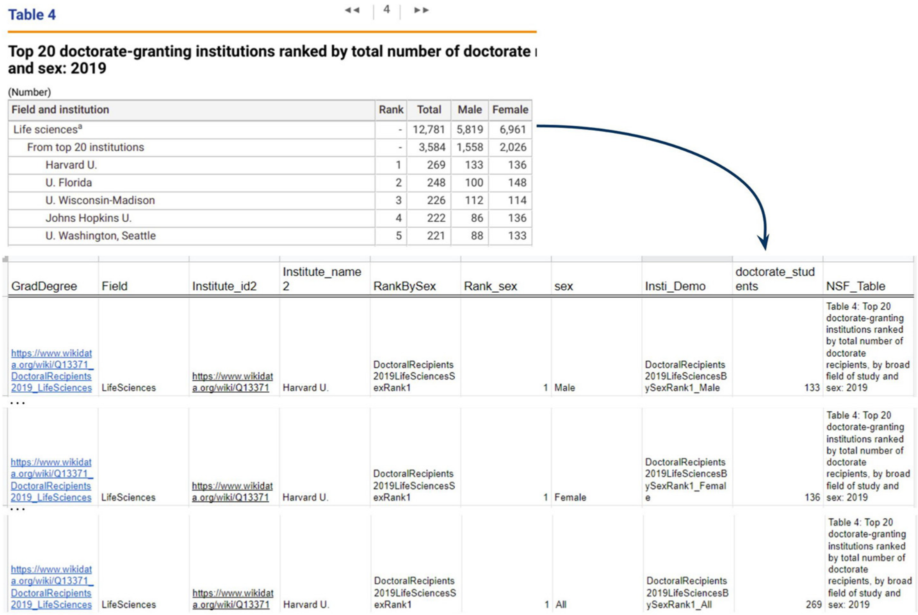


Table 4

Top 20 doctorate-granting institutions ranked by total number of doctorate r
and sex: 2019

(Number)

Field and institution	Rank	Total	Male	Female
Life sciences[a]	-	12,781	5,819	6,961
From top 20 institutions	-	3,584	1,558	2,026
Harvard U.	1	269	133	136
U. Florida	2	248	100	148
U. Wisconsin-Madison	3	226	112	114
Johns Hopkins U.	4	222	86	136
U. Washington, Seattle	5	221	88	133

GradDegree	Field	Institute_id2	Institute_name2	RankBySex	Rank_sex	sex	Insti_Demo	doctorate_students	NSF_Table
https://www.wikidata.org/wiki/Q13371_DoctoralRecipients2019_LifeSciences	LifeSciences	https://www.wikidata.org/wiki/Q13371	Harvard U.	DoctoralRecipients2019LifeSciencesSexRank1	1	Male	DoctoralRecipients2019LifeSciencesBySexRank1_Male	133	Table 4: Top 20 doctorate-granting institutions ranked by total number of doctorate recipients, by broad field of study and sex: 2019
…									
https://www.wikidata.org/wiki/Q13371_DoctoralRecipients2019_LifeSciences	LifeSciences	https://www.wikidata.org/wiki/Q13371	Harvard U.	DoctoralRecipients2019LifeSciencesSexRank1	1	Female	DoctoralRecipients2019LifeSciencesBySexRank1_Female	136	Table 4: Top 20 doctorate-granting institutions ranked by total number of doctorate recipients, by broad field of study and sex: 2019
…									
https://www.wikidata.org/wiki/Q13371_DoctoralRecipients2019_LifeSciences	LifeSciences	https://www.wikidata.org/wiki/Q13371	Harvard U.	DoctoralRecipients2019LifeSciencesSexRank1	1	All	DoctoralRecipients2019LifeSciencesBySexRank1_All	269	Table 4: Top 20 doctorate-granting institutions ranked by total number of doctorate recipients, by broad field of study and sex: 2019

Notes: The converted Life Science doctoral students demographic data for Harvard University as seen in the bottom screenshot was used to ingest the data along with the provenance in InDO knowledge graph using the customized SETLr script

Figure 6. The top screenshot depicts the view of the NSF SED Survey Table 4 as seen on the NSF website using the view option, while the bottom screenshot depicts the new processed table

graph. Our knowledge graph now consists of demographics “Sex” and “Ethnicity and Race”. Unlike for Sex statistics which is for each year, the Ethnicity and Race data is for a five-year period. It also has an addition of a total number of institutes with doctorates from a specific ethnicity and race (added within braces next to the ethnicity and race in first column before getting to the institutes names), and the Top 20–22 institutes among them. This adds new complexities to the data access and flattening process. Multiple new columns were added to properly represent all the data available in the provided table. Figure 7 provides a pictorial view of the before and after of Table 9 view which incorporates the data and the provenance of the ingested data.

3.3 Semantic extract transform and loader

We created a customized SETLr [6] script to populate our InDO knowledge graph with each of the processed NSF results tables. One can ingest and generate this knowledge graph in multiple ways. We decided to have one graph that ingests data from each of the processed NSF results tables as a whole. The connections between the data across these processed NSF results tables were inferred using the doctoral-granting US institute’s Wikidata links as their unique identifiers. This approach of ingesting data from each table separately and allowing the connection to be made between these different data points using linked data provides a mechanism to process the NSF SED data in a mutually exclusive manner. One will not be required to look at the entire graph code to modify the code for a particular processed NSF results table. Moreover, the

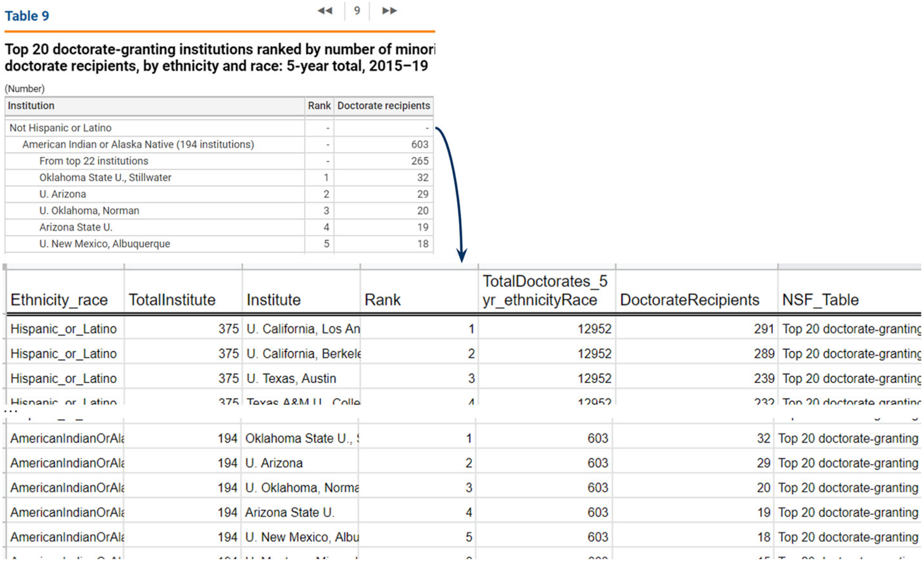


Table 9

Top 20 doctorate-granting institutions ranked by number of minori doctorate recipients, by ethnicity and race: 5-year total, 2015–19

(Number)

Institution	Rank	Doctorate recipients
Not Hispanic or Latino	-	-
American Indian or Alaska Native (194 institutions)	-	603
From top 22 institutions	-	265
Oklahoma State U., Stillwater	1	32
U. Arizona	2	29
U. Oklahoma, Norman	3	20
Arizona State U.	4	19
U. New Mexico, Albuquerque	5	18

Ethnicity_race	TotalInstitute	Institute	Rank	TotalDoctorates_5yr_ethnicityRace	DoctorateRecipients	NSF_Table
Hispanic_or_Latino	375	U. California, Los An	1	12952	291	Top 20 doctorate-grantinc
Hispanic_or_Latino	375	U. California, Berkele	2	12952	289	Top 20 doctorate-grantinc
Hispanic_or_Latino	375	U. Texas, Austin	3	12952	239	Top 20 doctorate-grantinc
Hispanic_or_Latino	375	Texas A&M U., Colle	4	12952	232	Top 20 doctorate-grantinc
…						
AmericanIndianOrAla	194	Oklahoma State U., S	1	603	32	Top 20 doctorate-granting
AmericanIndianOrAla	194	U. Arizona	2	603	29	Top 20 doctorate-granting
AmericanIndianOrAla	194	U. Oklahoma, Norma	3	603	20	Top 20 doctorate-granting
AmericanIndianOrAla	194	Arizona State U.	4	603	19	Top 20 doctorate-granting
AmericanIndianOrAla	194	U. New Mexico, Albu	5	603	18	Top 20 doctorate-granting

Figure 7.
The top screenshot depicts the view of NSF SED Survey Table 9 as seen on the NSF website using the view option and the bottom screenshot depicts the new processed table used to ingest the data

Notes: The data was ingested along with the provenance in InDO knowledge graph using the customized SETLr script. The data modeling for this table was different from the earlier tables due to its unique strudture. Here, we see how the four institutes accessed at Level 4 of the hierarchical table view are flattened to provide individual institutes ranking among the total institutes with graduates from a particular ethnicity and race for the five-year total

same code could be used to ingest the processed tables across multiple years of NSF SED survey data with similar structures. This script is also customized to follow the aggregate grouping criteria method as discussed earlier and provides room for categorizing each student as part of a demographic, field and institution. Figure 8 shows a pictorial view of this portion of the pipeline.

4. Results

Our current pipeline allows us to generate the InDO knowledge graph based on the InDO ontology in a semiautomated method and link the data set with Wikidata links as US doctoral-granting institute's unique identifier. Linking Wikidata with our data set allows us to broaden our reach of information available over the Web regarding these institutions. The consistency of the generated knowledge graph is checked through the Pellet reasoner.

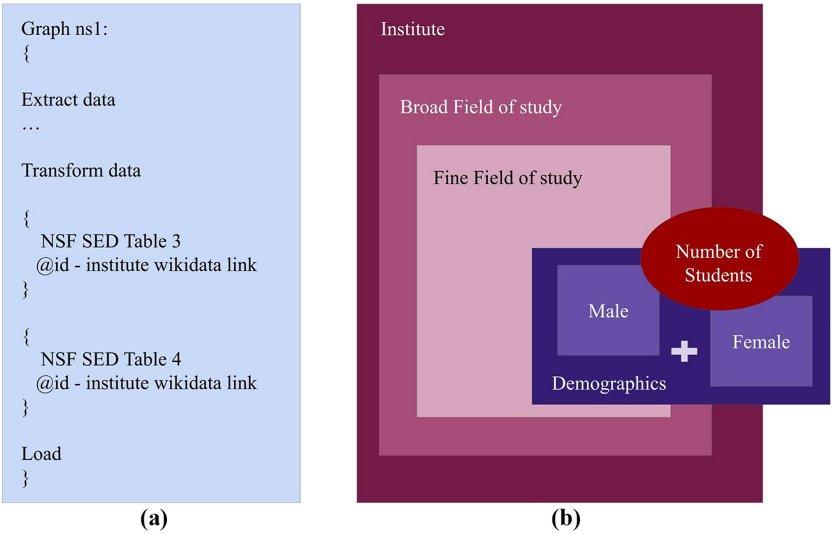
The semiautomatically generated InDO Knowledge Graph is created from 6,098 data points available across five different tables. The code connects instances of redundant instances across tables, helping us get a better overview of an institute's demographic, including its ranking and number of male and female students in one of the six broad fields of studies – Life Sciences, Physical Sciences and Earth Sciences, Mathematics and Computer Sciences, Psychology and Social Sciences, Engineering and Education. It also allows us to understand the doctoral demographics of an institute based on the five ethnicities and race groups – Hispanic_or_Latino, AmericanIndianOrAlaskanNative, Asian, BlackOrAfricanAmerican and MoreThanOneRace – over five years. Our semiautomatically generated knowledge graph consists of 109 classes and 1,138 instances. Table 1 depicts the composition of the 6,098 data points that lead to 1,138 instances in our InDO knowledge graph.

5. Evaluation

The evaluation was done in the following two ways:

(1) manually comparing the results of the competency questions obtained from the semiauto generated knowledge graph using a customized SETLr script to the manually generated InDO knowledge graph [7]; and

(2) expanding the competency questions to check for consistency and get a better overview of doctoral-granting institute's demographic information as more data points are added.



```
Graph ns1:
{

Extract data
…

Transform data

{
  NSF SED Table 3
 @id - institute wikidata link
}

{
  NSF SED Table 4
 @id - institute wikidata link
}

Load
}
```

Figure 8. (a) Demonstrates the overall structure of ingesting and connecting summary statistic data from NSF SED 2019 tables; (b) demonstrates the aggregation method used for storing the summary statistics provided in the NSF SED data tables

Table 1. Depicting the total number of data points ingested from four different table sources - three preprocessed direct tables from NSF SED 2019 survey and one additional created table consisting of doctoral-granting US institutes in 2019 with their respective Wikidata links

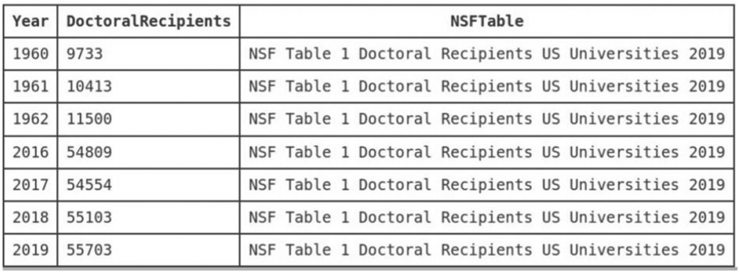
Sources	#data points
Generated table for US doctoral-granting institutes	194 × 2 = 388
NSF results Table 1: Doctorate Recipients from US Universities: 2019	61 × 2 = 122
NSF results Table 3: Top 50 doctorate-granting institutions ranked by total number of doctorate recipients, by sex: 2019	150 × 8 = 1,200
NSF results Table 4: Top 20 doctorate-granting institutions ranked by total number of doctorate recipients, by broad field of study and sex: 2019	366 × 10 = 3,660
NSF results Table 9: Top 20 doctorate-granting institutions ranked by number of minority US citizen and permanent resident doctorate recipients, by ethnicity and race: five-year total, 2015–19	104 × 7 = 728

Notes: These multiple data points have overlapping data in them, such as the institute name, their Wikidata links, and the NSF Table name as the provenance of the ingested data. These multiple instances of ingested data are identified as one instance in the knowledge graph avoiding duplication of data and helping in linking data from multiple data sources with accurate provenance information. This leads to the generation of 1,138 data instances in our semiautomated knowledge graph from the 6,098 data points across five tables

The semiautomatically generated knowledge graph simplifies the data ingestion process and removes extra instances used during the manual process. Hence, by updating the SPARQL queries for the newly semiautomatically created knowledge graph (Figure 15), we were able to obtain the same results for the following three competency questions as from its manually generated knowledge graph. These queries were run on Blazegraph Workbench:

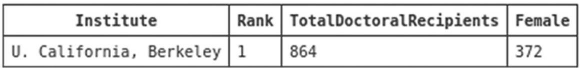
- (cq1) A given year's total doctoral recipients from US institutes between 1958 and 2019. For example, how many total doctoral recipients were there from 1960-62 and 2016-19 from US institutes (Figure 9)?
- (cq2) US institute that graduated the most doctorates in a given year and the binary gender representation of those students. For example, what is the US Institute with the maximum doctoral recipients in 2019, and how many of them were females (Figure 10)?
- (cq3) Institute with the most doctoral students in a given field of study. How many of those were from their marginalized community? For example, how many female doctoral recipients in 2019 were from the University of California Berkeley in the Mathematics and Computer Science graduate program (Figure 11)?

Now that we have more information, we have expanded the above competency questions to include other broad fields of studies and institutes. For example, we queried the (cq4) broad field of studies offered by Harvard University and the number of Male and Female students in each field. Because Harvard University is one of the 2019 top 20 institutes based on the number of doctoral recipients, we were able to obtain information about sex demographics in Life Sciences, Physical Sciences, Earth Sciences, Psychology and Social Sciences, as shown in Figure 12.



Year	DoctoralRecipients	NSFTable
1960	9733	NSF Table 1 Doctoral Recipients US Universities 2019
1961	10413	NSF Table 1 Doctoral Recipients US Universities 2019
1962	11500	NSF Table 1 Doctoral Recipients US Universities 2019
2016	54809	NSF Table 1 Doctoral Recipients US Universities 2019
2017	54554	NSF Table 1 Doctoral Recipients US Universities 2019
2018	55103	NSF Table 1 Doctoral Recipients US Universities 2019
2019	55703	NSF Table 1 Doctoral Recipients US Universities 2019

Figure 9. Result of sparql query as obtained on Blazegraph workbench for cq1-total US doctoral recipients for given years



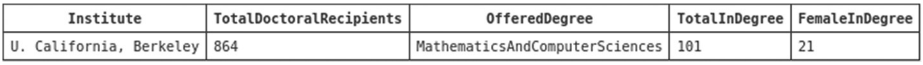
Institute	Rank	TotalDoctoralRecipients	Female
U. California, Berkeley	1	864	372

Figure 10. Result of sparql query as obtained on Blazegraph workbench for cq2-US institute with maximum 2019 doctoral recipients with female demographic data

We also queried the (cq5a) different ethnicities and races data available in the system and the (cq5b) institute that was ranked fifth in graduating doctorates from Black or African American race. Figures 13 and 14 demonstrate the respective results obtained from Blazegraph Workbench.

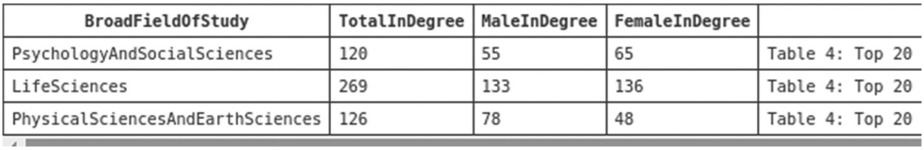
6. Discussion

This paper demonstrates a semiautomated process of generating the InDO knowledge graph based on the Institute Demographic Ontology. This process provides a scalable knowledge graph model aimed toward a comprehensible resource incorporating various demographics (sex, functional limitations, citizenship status, ethnicity and age.) The semiautomation process had its own challenges, from accessing the highly interlinked data provided by NSF SED survey 2019 results tables to finding a feasible and scalable process of creating the required knowledge using InDO, from ingesting summary statistics data to the process of connecting our database with other established Web sources, while keeping the knowledge graph clean and compact. We restructured the NSF SED 2019 survey result data tables from a highly interlinked table format to a simple table format. Restructuring the tables made processing and accessing the summary statistics data easier. This solution required the addition of a new class, "indo:All" as a subclass of "indo:Sex" under the demographic class, hence, resulting in InDO extension. The new induced class stored the total number of doctoral recipients, whether for an institute or for a broad field of study offered by an institute. The semiautomation process included the creation of a new table consisting of the 2019 doctoral granting US institutes and their respective Wikidata links. These links were



Institute	TotalDoctoralRecipients	OfferedDegree	TotalInDegree	FemaleInDegree
U. California, Berkeley	864	MathematicsAndComputerSciences	101	21

Figure 11. Result of sparql query as obtained on Blazegraph workbench for cq3-US institute with maximum 2019 doctoral recipients in the field of Mathematics and Computer Science with female demographic data



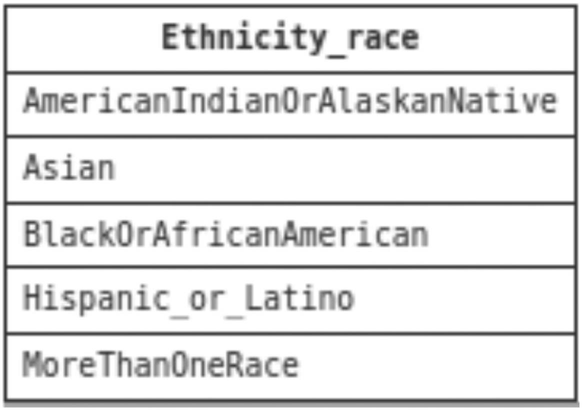
BroadFieldOfStudy	TotalInDegree	MaleInDegree	FemaleInDegree	
PsychologyAndSocialSciences	120	55	65	Table 4: Top 20
LifeSciences	269	133	136	Table 4: Top 20
PhysicalSciencesAndEarthSciences	126	78	48	Table 4: Top 20

Figure 12. Result of sparql query as obtained on Blazegraph workbench for Harvard Universities 2019 demographics based on broad field of study and sex

used as the institute's identifier, resulting in data connection across tables, maintaining consistency and preventing instance duplication. We also used the aggregation method to empower cohort visibility and keep track of students who are part of multiple groups. This method was implemented in the customized SETLr script for ingesting our current summary statistics data based on the number of doctoral recipients. Figure 6 illustrates the use of this method to create collections/groups/sets.

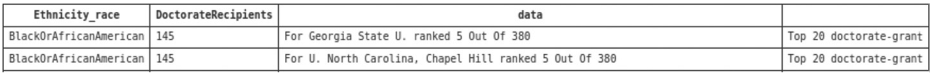
The customized SETLr script allowed quick data ingestion from our created institute's table and the preprocessed 2019 NSF SED 2019 survey's Tables 1, 3, 4 and 9. Even though these tables contain the top 50 and top 20 institutes based on total doctoral recipients and provide the total number of male/female doctoral recipients information, and the number of students for five ethnicity and race groups for a five-year period, it allows us to understand the general approach to expand and incorporate the required data available across 72 tables as part of NSF SED 2019 survey analysis report. Consistency of the generated knowledge graph was tested using the Protégé – 5.5.0 built-in Pellet reasoner. The current customized SETLr code used to create the semiautomatically generated InDO knowledge graph provides a mechanism to generate a machine-readable format of the NSF Survey of Earned Doctorates summary statistical data, originally provided in a highly interlinked and nonmachine processable format on the NSF website, without compromising the anonymity of the historically US underrepresented graduate student community.

During this automation process, we saw that the queries used on the manually created knowledge graph needed to be modified. Figure 15 shows the sparql query used



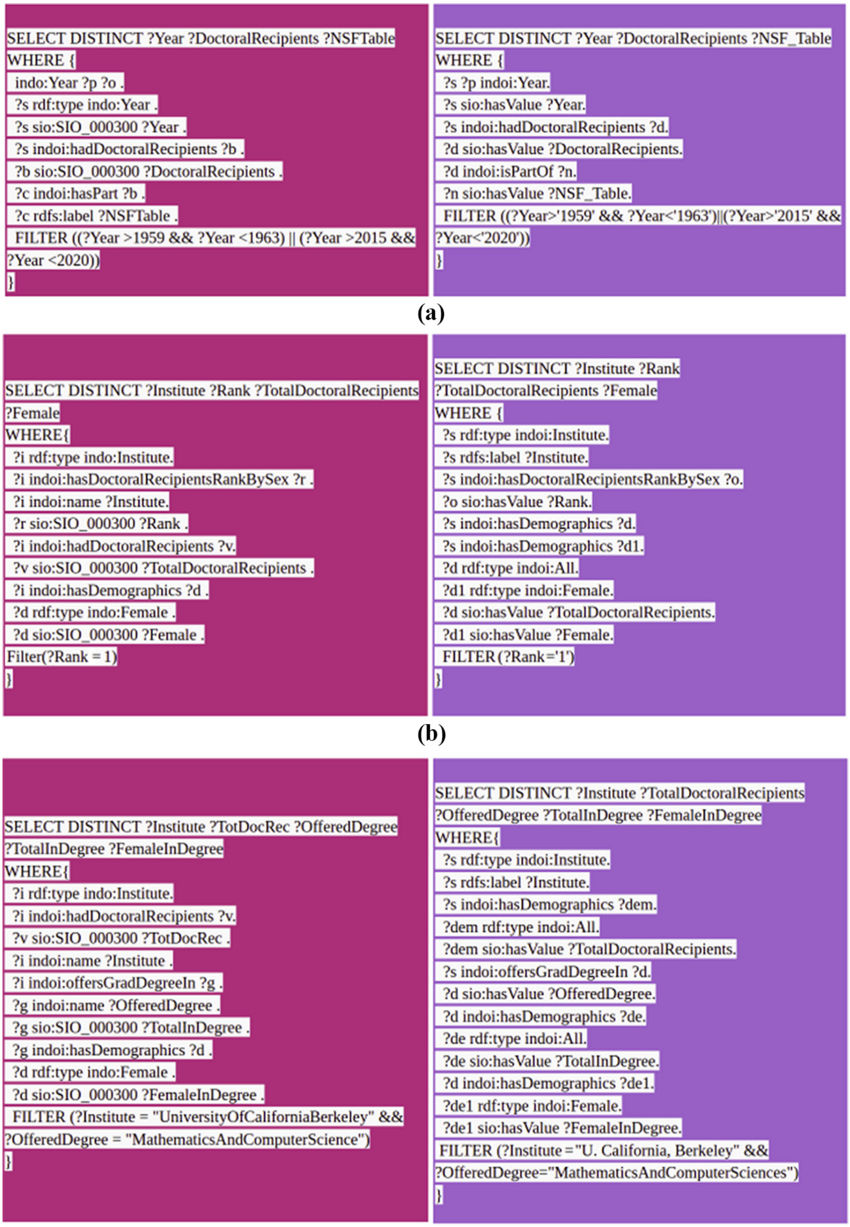
Ethnicity_race
AmericanIndianOrAlaskanNative
Asian
BlackOrAfricanAmerican
Hispanic_or_Latino
MoreThanOneRace

Figure 13.
Result of sparql query as obtained on Blazegraph workbench for ethnicities and race data available in the system



Ethnicity_race	DoctorateRecipients	data	
BlackOrAfricanAmerican	145	For Georgia State U. ranked 5 Out Of 380	Top 20 doctorate-grant
BlackOrAfricanAmerican	145	For U. North Carolina, Chapel Hill ranked 5 Out Of 380	Top 20 doctorate-grant

Figure 14.
Results of sparql query as obtained on Blazegraph workbench for institute ranked 5th with doctorates from Black or African American community



```
SELECT DISTINCT ?Year ?DoctoralRecipients ?NSFTable
WHERE {
 indo:Year ?p ?o .
 ?s rdf:type indo:Year .
 ?s sio:SIO_000300 ?Year .
 ?s indoi:hadDoctoralRecipients ?b .
 ?b sio:SIO_000300 ?DoctoralRecipients .
 ?c indoi:hasPart ?b .
 ?c rdfs:label ?NSFTable .
 FILTER ((?Year >1959 && ?Year <1963) || (?Year >2015 &&
?Year <2020))
}
```

```
SELECT DISTINCT ?Year ?DoctoralRecipients ?NSF_Table
WHERE {
 ?s ?p indoi:Year.
 ?s sio:hasValue ?Year.
 ?s indoi:hadDoctoralRecipients ?d.
 ?d sio:hasValue ?DoctoralRecipients.
 ?d indoi:isPartOf ?n.
 ?n sio:hasValue ?NSF_Table.
 FILTER ((?Year>'1959' && ?Year<'1963')||(?Year>'2015' &&
?Year<'2020'))
}
```

(a)

```
SELECT DISTINCT ?Institute ?Rank ?TotalDoctoralRecipients
?Female
WHERE{
 ?i rdf:type indo:Institute.
 ?i indoi:hasDoctoralRecipientsRankBySex ?r .
 ?i indoi:name ?Institute.
 ?r sio:SIO_000300 ?Rank .
 ?i indoi:hadDoctoralRecipients ?v.
 ?v sio:SIO_000300 ?TotalDoctoralRecipients .
 ?i indoi:hasDemographics ?d .
 ?d rdf:type indo:Female .
 ?d sio:SIO_000300 ?Female .
Filter(?Rank = 1)
}
```

```
SELECT DISTINCT ?Institute ?Rank
?TotalDoctoralRecipients ?Female
WHERE {
 ?s rdf:type indoi:Institute.
 ?s rdfs:label ?Institute.
 ?s indoi:hasDoctoralRecipientsRankBySex ?o.
 ?o sio:hasValue ?Rank.
 ?s indoi:hasDemographics ?d.
 ?s indoi:hasDemographics ?d1.
 ?d rdf:type indoi:All.
 ?d1 rdf:type indoi:Female.
 ?d sio:hasValue ?TotalDoctoralRecipients.
 ?d1 sio:hasValue ?Female.
 FILTER(?Rank='1')
}
```

(b)

```
SELECT DISTINCT ?Institute ?TotDocRec ?OfferedDegree
?TotalInDegree ?FemaleInDegree
WHERE{
 ?i rdf:type indo:Institute.
 ?i indoi:hadDoctoralRecipients ?v.
 ?v sio:SIO_000300 ?TotDocRec .
 ?i indoi:name ?Institute .
 ?i indoi:offersGradDegreeIn ?g .
 ?g indoi:name ?OfferedDegree .
 ?g sio:SIO_000300 ?TotalInDegree .
 ?g indoi:hasDemographics ?d .
 ?d rdf:type indo:Female .
 ?d sio:SIO_000300 ?FemaleInDegree .
 FILTER (?Institute = "UniversityOfCaliforniaBerkeley" &&
?OfferedDegree = "MathematicsAndComputerScience")
}
```

```
SELECT DISTINCT ?Institute ?TotalDoctoralRecipients
?OfferedDegree ?TotalInDegree ?FemaleInDegree
WHERE{
 ?s rdf:type indoi:Institute.
 ?s rdfs:label ?Institute.
 ?s indoi:hasDemographics ?dem.
 ?dem rdf:type indoi:All.
 ?dem sio:hasValue ?TotalDoctoralRecipients.
 ?s indoi:offersGradDegreeIn ?d.
 ?d sio:hasValue ?OfferedDegree.
 ?d indoi:hasDemographics ?de.
 ?de rdf:type indoi:All.
 ?de sio:hasValue ?TotalInDegree.
 ?d indoi:hasDemographics ?de1.
 ?de1 rdf:type indoi:Female.
 ?de1 sio:hasValue ?FemaleInDegree.
 FILTER (?Institute ="U. California, Berkeley" &&
?OfferedDegree="MathematicsAndComputerSciences")
}
```

(c)

Notes: We can see that restructuring made the total students as part of the demographic class, which requires an additional pattern matching with the demographic class

Figure 15. SPARQL queries on the left were used to get competency question responses from manually generated InDO knowledge graph, and the SPARQL queries on the right side were used to get the same responses from semiautomatically generated knowledge graphs

on the manually created knowledge graph [8] vs the semiautomatically created knowledge graph to obtain responses for the competency questions discussed in the evaluation section. The major change was moving the total doctorates under the demographic class "indo:All." This required an additional pattern matching in the SPARQL query to access the total number of graduate recipients in a year or for a particular institute and field. This change does add an extra pattern matching step in our sparql queries but helps keep things clean and simple by having all the aggregated values as part of the demographics class, facilitating the cohort visibility. It also allows extensibility that can be used as we extend the work to a larger class of marginalized students. The current work also helps us understand how data harmonization could take place between different demographic groups, leading to a full picture of an institute's demographics based on the provided data.

7. Conclusion

In this paper, we discussed the tools required and challenges faced while semiautomating the process of InDO knowledge graph generation and connecting it with established Web sources such as Wikidata. We also show how we found a way to tackle a social science problem from a computer science perspective. Finally, we demonstrate how our InDO ontology could be extended and made inclusive to fulfill the requirement for a comprehensive resource for marginalized graduate students.

In particular, we were able to semiautomatically generate the InDO knowledge graph, using customized SETLr code, to contain 1,138 instances across five different data sources and use the Wikidata links as the institute's identifiers. We also provide the details of how the SETLr script is structured and the use of aggregation methods for summary statistics to empower cohort visibility. Evaluation of the generated knowledge graph was illustrated in two ways: first, by comparing the competency question sparql query results obtained from the semiautomatically generated InDO knowledge graph to the manually generated one, and second, by extending those competency questions to get a better picture of an institute's demographic based on the broad field of study and sex. The queries were run on Blazegraph workbench while the consistency of the knowledge graph was checked using Pellet.

Future work includes extending the current semiautomatic process to include required data from all 72 tables provided as part of the NSF SED 2019 survey analysis report. This will allow us to include other demographic factors such as functional limitations, age and citizenship status, along with sex, ethnicity and race, as provided in these tables. This would then be extended to include data from all the available NSF SED survey analysis report data tables over the years to understand the evolution of an institute's graduate students' demographics. Even though the initial focus of our work is with institutes within the USA, we hope to demonstrate in the future how users in other countries can use InDO. Along with this, we hope to more fully address the maintainability of both InDO and InDO knowledge graphs as well as the dynamic nature of the survey data, perhaps through a controlled versioning process. In addition, we will explore increasing the representation of other marginalized communities through more sources that extend the ontology to handle the needs of students dealing with gender bias, physical disabilities and other issues.

Notes

1. www.wikidata.org/wiki/Wikidata:Main_Page
2. http://swat.cse.lehigh.edu/onto/univ-bench.owl

3. https://github.com/protegeproject/cellfie-plugin
4. Survey of Earned Doctorates | NCSES | NSF, available at: www.nsf.gov/statistics/srvydoctorates/
5. Doctorate Recipients from U.S. Universities: 2019 | NSF – National Science Foundation, available at: https://ncses.nsf.gov/pubs/nsf21308/data-tables
6. https://github.com/tetherless-world/setlr
7. https://tetherless-world.github.io/institute-demographic-ontology/competencyquestions/
8. https://tetherless-world.github.io/institute-demographic-ontology/competencyquestions/

References

Abatso, Y.R., Allen, W.R., Blackwell, J.E., Braddock, I.I.J.H., Brazziel, M.E., Brazziel, W.F., Davis, J.S., Epps, E.G., Fincher, C., Gosman, E.J. and Grigg, C.M. (1987), *Pursuit of Equality in Higher Education*, Rowman and Littlefield.

Balhoff, J.P., Brush, M.H., Christopherson, L., de Coronado, S., Fragoso, G., Haendel, M.A., Mungall, C.J., Robasky, K., Vasilevsky, N.A. and Wright, L.W. (2017), "Tailoring the NCI thesaurus for use in The OBO Library", ICBO.

Blackwell, J.E. (1987), *Mainstreaming Outsiders: The Production of Black Professionals*, Rowman and Littlefield.

Balshaw, D.M., Collman, G.W., Gray, K.A. and Thompson, C.L. (2017), "The children's health exposure analysis resource (CHEAR): enabling research into the environmental influences on children's health outcomes", *Current Opinion in Pediatrics*, Vol. 29 No. 3, p. 385.

Brown, M.C. (2000), "Involvement with students: how much can I give", *Succeeding in an Academic Career: A Guide for Faculty of Color*, pp. 71-88.

Brown, M.C., II, Davis, G.L. and McClendon, S.A. (1999), "Mentoring graduate students of color: myths, models, and modes", *Peabody Journal of Education*, Vol. 74 No. 2, pp. 105-118.

Chari, S., Qi, M., Agu, N.N., Seneviratne, O., McCusker, J.P., Bennett, K.P., Das, A.K. and McGuinness, D. L. (2019), "Making study populations visible through knowledge graphs", *International Semantic Web Conference*, Springer, Cham, pp. 53-68.

Dong, X.L. (2019), "Building a broad knowledge graph for products", *2019 IEEE 35th International Conference on Data Engineering (ICDE), IEEE*, pp. 25-25.

Dumontier, M., Baker, C.J., Baran, J., Callahan, A., Chepelev, L., Cruz-Toledo, J., Del Rio, N.R., Duck, G., Furlong, L.I., Keath, N. and Klassen, D. (2014), "The semanticscience integrated ontology (SIO) for biomedical research and knowledge discovery", *Journal of Biomedical Semantics*, Vol. 5 No. 1, pp. 1-11.

Foley, D. (2021), "Survey of doctorate recipients, 2019", NSF 21-230. National Center for Science and Engineering Statistics (NCSES), National Science Foundation, Alexandria, VA, available at: https://ncses.nsf.gov/pubs/nsf21320/

Frierson, H.T., Jr, (1990), "The situation of black educational researchers: continuation of a crisis", *Educational Researcher*, Vol. 19 No. 2, pp. 12-17.

Fuhrmann, C.N., Halme, D.G., O'sullivan, P.S. and Lindstaedt, B. (2011), "Improving graduate education to support a branching career pipeline: recommendations based on a survey of doctoral students in the basic biomedical sciences", *CBE – Life Sciences Education*, Vol. 10 No. 3, pp. 239-249.

Gay, G. (2004), "Navigating marginality en route to the professoriate: graduate students of color learning and living in academia", *International Journal of Qualitative Studies in Education*, Vol. 17 No. 2, pp. 265-288.

Gibbs, K.D., Jr, John McGready, J.C. and Bennett, G.K. (2014), “Biomedical science Ph. D. career interest patterns by race/ethnicity and gender”, *PloS One*, Vol. 9 No. 12, p. e114736.

Girves, J.E., Zepeda, Y. and Gwathmey, J.K. (2005), “Mentoring in a post-affirmative action world”, *Journal of Social Issues*, Vol. 61 No. 3, pp. 449-479.

Golde, C.M. and Dore, T.M. (2004), “The survey of doctoral education and career preparation: the importance of disciplinary contexts”, *Paths to the Professoriate: Strategies for Enriching the Preparation of Future Faculty*, pp. 19-45.

Guo, Q., Zhuang, F., Qin, C., Zhu, H., Xie, X., Xiong, H. and He, Q. (2020), “A survey on knowledge graph-based recommender systems”, *IEEE Transactions on Knowledge and Data Engineering*.

High, R. (2012), *The Era of Cognitive Systems: An Inside Look At Ibm Watson And How It Works*, Vol. 1, IBM Corporation, Redbooks, p. 16.

Jain, D., Potluri, V. and Sharif, A. (2020), “Navigating graduate school with a disability”, *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, pp. 1-11.

Ji, S., Pan, S., Cambria, E., Marttinen, P. and Philip, S.Y. (2021), “A survey on knowledge graphs: representation, acquisition, and applications”, *IEEE Transactions on Neural Networks and Learning Systems*.

Kendall, E.F. and McGuinness, D.L. (2019), “Ontology engineering”, *Synthesis Lectures on the Semantic Web: Theory and Technology*, Vol. 9 No. 1, pp. 1-102.

Keshan, N. (2021), “Building a social machine for graduate mobility”, *13th ACM Web Science Conference 2021*, pp. 156-157.

Keshan, N., Fontaine, K. and Hendler, J.A. (2021), “InDO: the institute demographic ontology”, *Iberoamerican Knowledge Graphs and Semantic Web Conference*, Springer, Cham, pp. 1-15.

Krishnan, A. (2018), *Making Search Easier: How Amazon’s Product Graph is Helping Customers Find Products More Easily*, Amazon Blog.

Luna, V. and Prieto, L. (2009), “Mentoring affirmations and interventions: a bridge to graduate school for Latina/o students”, *Journal of Hispanic Higher Education*, Vol. 8 No. 2, pp. 213-224.

McCallum, C. (2017), “Giving back to the community: how African Americans envision utilizing their PhD”, *The Journal of Negro Education*, Vol. 86 No. 2, pp. 138-153.

McCusker, J.P., Chastain, K., Rashid, S., Norris, S. and McGuinness, D.L. (2018), “SETLR: the semantic extract, transform, and load-r”, *PeerJ Preprints*, Vol. 6, p. e26476v1.

McCusker, J.P., Dumontier, M., Chari, S., Luciano, J.S. and McGuinness, D.L. (2019), “A linked data representation for summary statistics and grouping criteria”, *SA WSemStats@ ISWC*.

McGee, R., Jr, Saran, S. and Krulwich, T.A. (2012), “Diversity in the biomedical research workforce: developing talent”, *Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine*, Vol. 79 No. 3, pp. 397-411.

Madhavan, J., Ko, D., Kot, Ł., Ganapathy, V., Rasmussen, A. and Halevy, A. (2008), “Google’s deep web crawl”, *Proceedings of the VLDB Endowment*, Vol. 1 No. 2, pp. 1241-1252.

Pau, R. (2009), “Experiential factors which influence how female students perceive computing and computing careers at different stages in their education”, Doctoral dissertation, University of Southampton.

Paulheim, H. (2017), “Knowledge graph refinement: a survey of approaches and evaluation methods”, *Semantic Web*, Vol. 8 No. 3, pp. 489-508.

Roach, M. and Sauermann, H. (2017), “The declining interest in an academic career”, *Plos One*, Vol. 12 No. 9, p. e0184130.

Rose, G.L. (2005), “Group differences in graduate students’ cconcepts of the ideal mentor”, *Research in Higher Education*, Vol. 46 No. 1, pp. 53-80.

Sauermann, H. and Roach, M. (2012), "Science PhD career preferences: levels, changes, and advisor encouragement", *PLoS ONE*, Vol. 7 No. 5, p. e36307.

Sevelius, J.M., Gutierrez-Mock, L., Zamudio-Haas, S., McCree, B., Ngo, A., Jackson, A., Clynes, C., Venegas, L., Salinas, A., Herrera, C. and Stein, E. (2020), "Research with marginalized communities: challenges to continuity during the COVID-19 pandemic", *AIDS and Behavior*, Vol. 24 No. 7, pp. 2009-2012.

Singhal, A. (2012), "Introducing the knowledge graph: things, not strings", *Official Google Blog*, Vol. 5, p. 16.

Tamjeed, M., Tibdewal, V., Russell, M., McQuaid, M., Oh, T. and Shinohara, K. (2021), "Understanding disability services toward improving graduate student support", *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, pp. 1-14.

Terenzini, P.T. (1996), "Presidential address: rediscovering roots: public policy and higher education research", *The Review of Higher Education*, Vol. 20 No. 1, pp. 5-13.

Thomas, K.M., Willis, L.A. and Davis, J. (2007), "Mentoring minority graduate students: issues and strategies for institutions, faculty, and students", *Equal Opportunities International*, Vol. 26 No. 3.

Tiwari, S., Al-Aswadi, F.N. and Gaurav, D. (2021), "Recent trends in knowledge graphs: theory and practice", *Soft Computing*, Vol. 25 No. 13, pp. 8337-8355.

Ullrich, L.E. Ogawa, J.R. and Jones-London, M.D. (2021), "Factors that influence career choice among different populations of neuroscience trainees", bioRxiv.

Willie, C.V. (1991), *African-Americans and the Doctoral Experience: Implications for Policy*, Teachers College Press, *Columbia University*, New York, NY.

Zhou, Y. (2010), "Understanding of international graduate students' academic adaptation to a US graduate school", Doctoral dissertation, *Bowling Green State University*.

Further reading

Ahmed, N., Khan, S. and Latif, K. (2016), "Job description ontology", *2016 International Conference on Frontiers of Information Technology (FIT)*, *IEEE*, pp. 217-222.

Corresponding author

Neha Keshan can be contacted at: keshan@rpi.edu