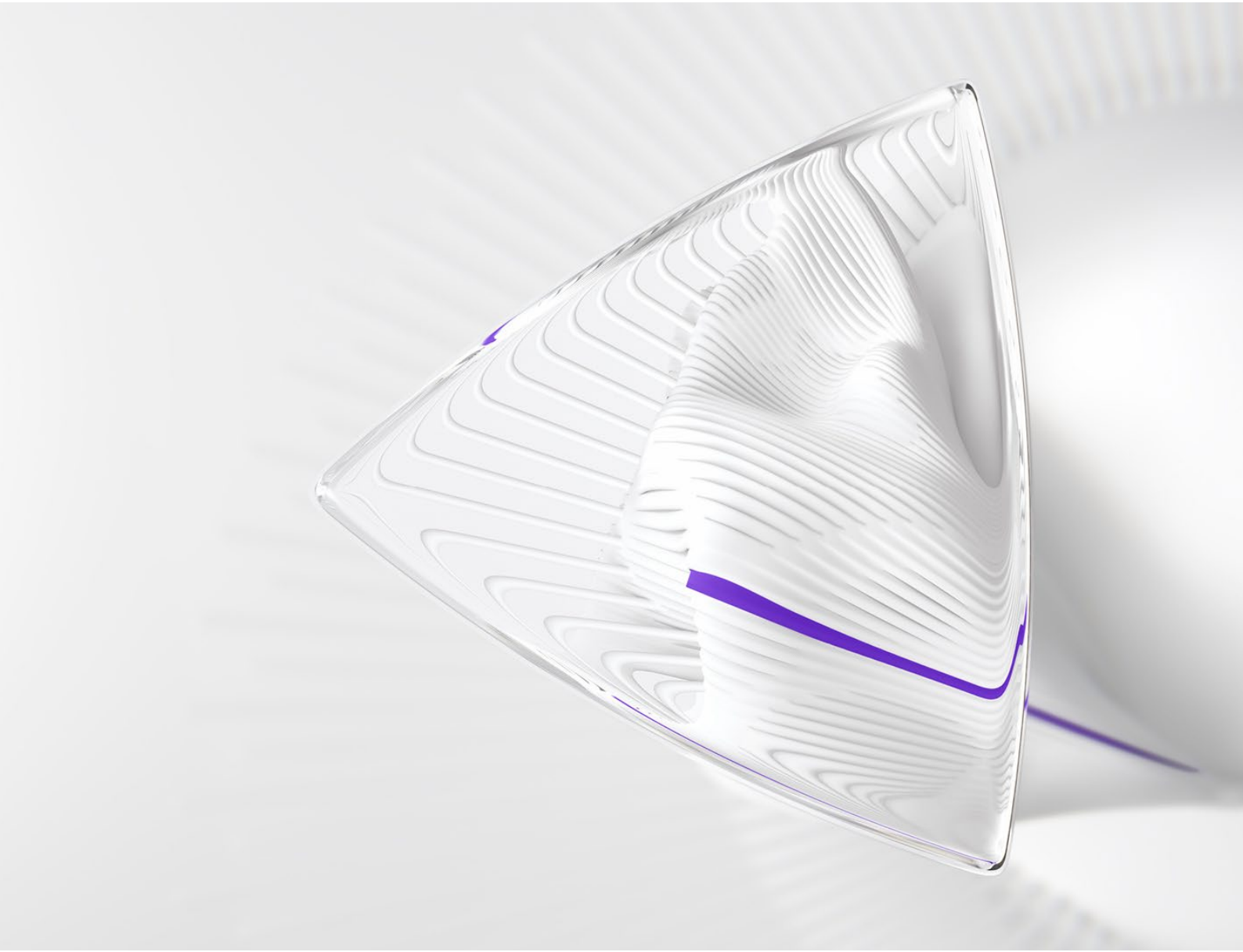
Clarivate™

Steppingblocks

Clarivate and Steppingblocks: Foreign-Born Scientists and Engineers in the US Workforce FBSE 22

Towards Profiles of Foreign-Born Scientists and Engineers

Task 5: Data analysis and quality assessment



The America's DataHub Consortium (ADC), a public-private partnership is being utilized to implement research opportunities that support the strategic objectives of the National Center for Science and Engineering Statistics (NCSES) within the U.S. National Science Foundation (NSF). This report documents research funded through the ADC and is being shared to inform interested parties of ongoing activities and to encourage further discussion. Any opinions, findings, conclusions, or recommendations expressed in this report do not necessarily reflect the views of NCSES or NSF. Please send questions to **ncsesweb@nsf.gov**. This product has been reviewed for unauthorized disclosure of confidential information under NCSES-DRN24-082.

Contents

1. Introduction

Task 5 is the culmination of a series of processes that began with the linkage of Steppingblocks, Clarivate and government data to create a data set that sheds light on the education and careers of Foreign–born Scientists and Engineers (FBSE). In Task 5 we turn toward a preliminary assessment of the FBSE data set's representativeness and, in so doing, offer an evaluation of its utility for research and policy decisions. This report is in three parts:

1. A narrative description of the work undertaken in Tasks 1-4, the challenges encountered, how they were addressed, and a procedural explanation of our confidence in the validity of the FBSE data set.
2. An examination of our data for the purposes of identifying their limitations and assessing whether the data are representative of select substrata of the FBSE population.
3. A series of data summaries, analyses, and visualizations that illustrate the types of research that can be conducted with these data and a discussion of how such data and their analyses, including simulations, can yield evidence to support policy decisions.

With the completion of Task 5, we can confirm that we have satisfied the first requirement for success of America's DataHub—the creation of a database of FBSEs in which the identified individuals are, in fact, who we claim them to be. In other words, our data are valid. At the same time, we conclude that our final validated data set is not representative of the population of FBSE in the US.

2. Creation of an FBSE data set

2.1. Available data sources

To understand this project's final data set, it is helpful to revisit the underlying data and the processes used to build the data set.

The starting point was Steppingblocks' existing 135M+ workforce profiles and one billion career milestones. Steppingblocks aggregates and distills data from many publicly available data sources, including social media platforms, online job sites, government sources, and resumes. Parallel computing was used to map hundreds of variables from the FBSE's first job title to their most recent, and everything in between, including variables indicating salary, skills, and educational qualifications.

We targeted citizenship and visa information data from the Department of Labor's Office of Foreign Labor Certification (OFLC) and Department of Homeland Security's Student and Exchange Visitor Information System (SEVIS) with the intent to validate the profiles identified in the initial FBSE data set. However, these government data proved to be inaccessible within the project timeline. The alternative approach considered publicly available government OFLC data, where the most relevant data sets were the permanent labor certifications (PERM) and H-1B visa applications (H-1B). These data sets focus on specialty occupations and were thought most likely to match FBSE candidates generated from the initial Steppingblocks' data. Of these options, the period 2016 through 2019 in the public PERM data set was selected as the richer source for linkage.

To learn more about the research productivity and research contributions of the FBSEs, we next merged the Steppingblocks FBSE candidate population with Clarivate's Web of Science (WoS). Among professional bibliometricians, the Web of Science is considered an unparalleled resource for research into the past, present and future of scholarly research.

2.2. Methodology

To build the initial 35M FBSE candidate data set (outermost circle in Figure 1), Steppingblocks identified potential foreign markers in their 135M+ career profile database. Institutions were identified from this output but had to be manually reviewed and assessed to build the tagged records for an initial data set (7.49M in the second circle in Figure 1). Steppingblocks drew from their education and job category taxonomies to match the *Science and engineering (S&E) fields* and *Science and engineering (S&E)-related occupations*, as defined in NCSES Glossary[1]. This filtering process winnowed the data set to 4.06M individuals with science and engineering degrees of whom 1.34M had identifiable links to organizations (Innermost circles in Figure 1).

Validation

Validation of the FBSE profile and the data set was the next milestone. As noted earlier, PERM records were selected as the richer data set for linkage. We chose "Certified," and "Certified-Expired" PERM records for the years 2016-2019 as selection parameters based on data completeness, linkage potential, and expected output volume.

The normalized and filtered PERM data set was then linked to the FBSE data set. The first step was an organizational linkage that provided approximately sixty-two percent linking rate and yielded 249,182 applications which could be positively linked with Steppingblocks' data.

Next, dates of employment in Steppingblocks' database were linked to the decision dates found in the 249,182 PERM records. There was a return of 90.5

[1] NCSES Glossary: **The State of U.S. Science and Engineering 2022 | NSF - National Science Foundation**

percent, reducing the size of the data set to 225,065 FBSE profiles. Next, similarity computations were employed using these variables:

- Graduation Year (blocking was performed on that variable)
- School Name (We used Jaro Winkler text similarity[2] to allow for different spellings and variations in how the name was reported.)
- Degree Level (Exact Match)
- Education Major (Cosine Similarity of Sentence Embeddings, BERT)
- Job Title (Cosine Similarity on Sentence Embeddings, BERT)

A disadvantage of this approach was the potential difficulty in distinguishing among a large number of candidates with the same education domain working at the same company and performing the same job function (e.g., Computer Science major, hired at Google as a Software Engineer). To overcome that risk, an additional filtering layer used a machine learning model to predict ethnicities from Steppingblocks' potential candidate's full names (when possible) to determine the best candidate for each OFLC record. From that effort, 81,162 Steppingblocks profiles were linked to OFLC PERM data sets between 2016 and 2019. Using the NCSES criteria for foreign-born scientists or engineers as a final filter yielded 77,823 linked FBSE profiles.

In light of these data validation measures, we have high confidence in the resulting data set, referred to henceforth as PERM-SB. PERM-SB is, in itself, a high-quality data output. More broadly, it demonstrates the feasibility of using Steppingblocks' profiles to obtain verified profiles of FBSEs using only publicly available data.

Linking to Web of Science

Following the creation and validation of the PERM-SB data set, we turned to the task of linking it to the Web of Science (WoS) data on publications and citations. The first filter initially linked records with a shared last name and first initial. FBSE profiles of individuals with common names yielded multiple matches among the WoS profiles. Of the original data set of 77,823 FBSE profiles, 11,862 FBSEs did not share a common last name or first initial with any WoS profile. Additionally, 3,174 FBSEs lacked an identified last name altogether. The linkage was performed on the remaining 62,787 FBSEs.

PERM-SB was linked to WoS profiles using five administrative data variables common to both data sets: institutional email domain name, first name, institutional affiliations, and name commonality. This filtering process yielded 10,550 (16.8%) PERM-SB profiles that matched a unique WOS profile. Another 809 PERM-SB profiles (1.3%) were matched solely on personal email address. This matched data set consisting of 11,359 FBSE profiles is referred to as PERM-SB-WoS and is depicted in the center of the Venn diagram in Figure 2.

[2] Jaro Winkler text similarity definition: https://en.wikipedia.org/wiki/Jaro%E2%80%93Winkler_distance#Jaro%E2%80%93Winkler_similarity

Figure 1. Progression of FBSE database development

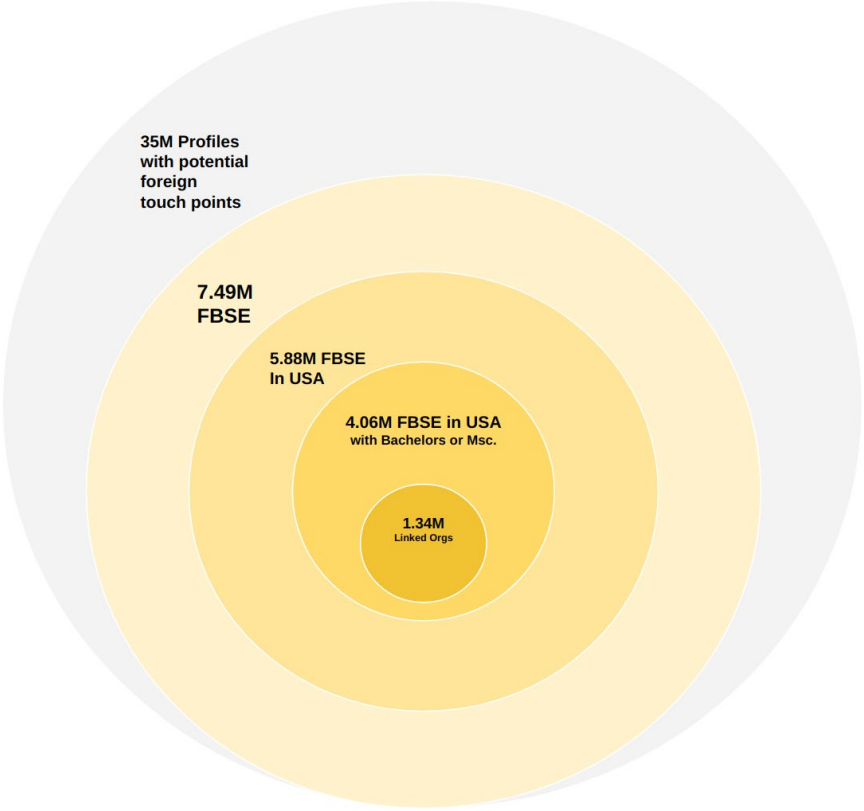
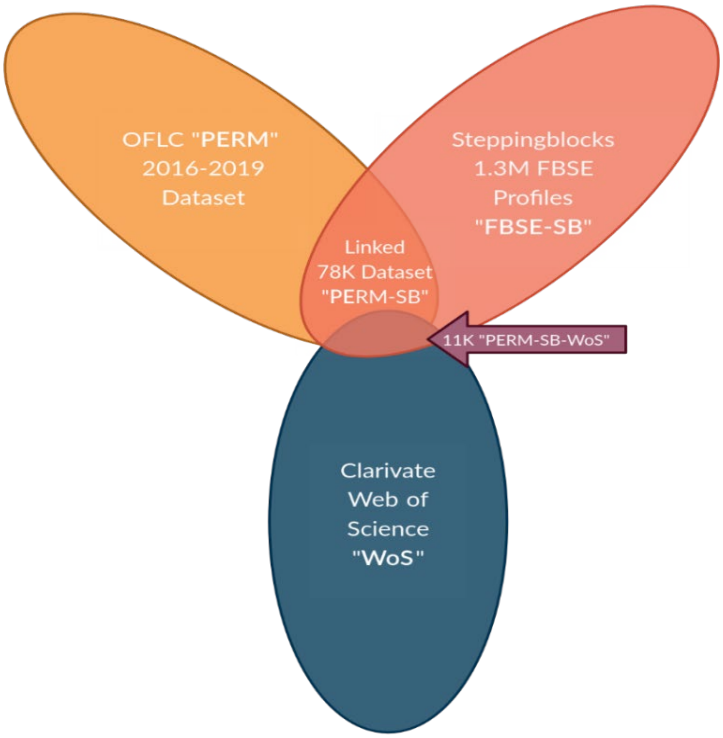


Figure 2: Venn diagram depicting relationships among the OFLC PERM data set, Steppingblocks FBSE profiles and the Web of Science



2.3. Challenges

At several junctures, the work of developing the final PERM-SB-WoS data set presented challenges that had to be analyzed and addressed before a high-quality PERM-SB-WoS FBSE data set could be constructed and deployed with confidence. Fields within each data source had to be carefully considered for gaps, accuracy, and level of completeness as well as compatibility with the next data source. The data sources each required varying degrees of disambiguation and normalization. At the onset of the project, Steppingblocks examined the data collected in their processes to select those observations for their Extraction, Transformation, and Loading (ETL) process, using the following criteria for inclusion:

- Any location markers outside the United States, e.g., job, training, and education locations.
- Languages e.g., Korean
- Group membership, awards, or projects linked to foreign populations, e.g., "Association of Chinese Students and Scholars at Yale" or "ALPFA.org | Latino Professionals for America"

Foreign institutions and degree programs were identified in the first FBSE profiles from Steppingblocks database. Often this information followed native language and norms and was not easily mapped to US degrees and naming conventions.

Cases of US students who were studying in a foreign country or US entities with foreign campuses, for example, also posed an unexpected impediment. For instance, the use of foreign markers as an initial screen identified:

- US students studying abroad
- Military personnel
- US expatriates
- Enrollees in foreign campuses of US universities, e.g., the degree granting foreign campuses[3] of New York University in Abu Dhabi or Shanghai.

We had to carefully review our data, often manually, to confirm that such individuals did meet the NCSES definition[4] for an FBSE before including them in our data set for further processing.

There are several government databases that have reliable information for validating our selection process by verifying that the individuals in our data set were indeed FBSE. However, we did not have access to such data. Hence our risk mitigation strategy, in lieu of the government data, was to find publicly available information that could be used to verify that we were correctly identifying FBSEs. The publicly available government data such as the OFLC H-1B visa applications data, have ninety-six distinct variables for each applicant. Of these, Steppingblocks identified thirteen variables for linking. The variables could potentially allow Steppingblocks to distinguish between different individuals who started working for the same company around the same date. The PERM data sets have even more data than the H-1B set, with up to 154 variables for each case, including sponsoring employers, intended employee positions, previous education, and country of birth/citizenship. After careful review, nine PERM variables were identified for linkage.

Although better results were possible with the PERM data versus the H-1B data, certain challenges persisted, including self-reporting on fields such as job title, and employer. In linking the FBSE data set with the Web of Science, challenges were noted in self-reported variable spellings or omissions of FBSE first names (e.g.,

[3] NYU foreign campuses: **https://www.nyu.edu/faculty/faculty-in-the-global-network.html**
[4] NCSES definitions: **The State of U.S. Science and Engineering 2022 | NSF - National Science Foundation**

name with diacritics transliterated), inconsistent use of a middle initial, and variations in self-reported institutional affiliation name.

3. Representativeness of the data

As each data source was assessed and incorporated, assumptions and options were reviewed and biases inherent in the data were identified. Identifying FBSEs using the methods devised thus far resulted in a preponderance of FBSEs who stayed and worked in the United States.

Despite excellent classification accuracy and an F1-score[5,6] of 0.86, our model is heavily biased towards accurately identifying only the most common ethnic groups represented in the US.

3.1. Implications of linking to the PERM data

While the PERM data provided the greatest potential for linkage, it has a built-in bias towards H-1B profiles[7], that is individuals with advanced degrees and in specialized areas. Steppingblocks used 2016-2019 public PERM records as they offered the most consistent and recent data to maximize our chances of a good link to Steppingblocks' database.

Each PERM record had a case status with four possible outcomes: Certified, Certified-Expired, Denied, and Withdrawn. Applications with the status of Denied or Withdrawn could have been denied or withdrawn for various reasons, including data inaccuracy. Therefore, Steppingblocks decided to filter these out. The resulting data set numbered 406,921 individual applications after this initial filtering.

Furthermore, to increase chances of generating high-quality matches, only records that could be linked directly to Steppingblocks' list of organizations were selected, which generated a slight bias towards larger organizations.

Identifying FBSEs using the methods devised thus far will result in a bias towards FBSEs who stayed and worked in the United States. A simple volumetric analysis shows that in 2019, for every certified PERM application, roughly twice as many H-1Bs visas and four times as many student[8] visas (F or M) were granted.

Clarivate successfully matched 11,359 individuals from the PERM-SB data set to their publications to generate a new PERM-SB-WoS data set. Biographical information from the two data sources enabled Clarivate to establish the linkage between PERM-SB and WoS and to provide information about FBSE publication productivity and impact, and their research topic area. An analysis of the combined PERM-SB-WoS data set (see section IV) provides a robust picture of FBSE research activities and their contribution to the US scientific community.

There is at least 97% coverage for each field provided by the PERM-SB-WoS data set. That is, at least 97% of records have a valid value for each WoS variable.

---

[5] Cai, L. and Zhu, Y., 2015. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Science Journal, 14, p.2. DOI: http://doi.org/10.5334/dsj-2015-002, Taleb, I., Serhani, M.A., Bouhaddioui, C. et al. 2021. Big data quality framework: a holistic approach to continuous quality management. J Big Data 8, 76.

6 F1-score is the harmonic mean of precision and recall of a classification. Its value ranges between 0 and 1 and measures how good a method is at classification. The precision of a classification, (in our case, of the match) is the number of true positive results divided by the number of all positive results, (both true and false positives), and the recall is the number of true positive matches divided by the number of all matches that should have been identified as positive. https://en.wikipedia.org/wiki/F-score

[7] H-1B visa: **https://www.immi-usa.com/h1b-visa/h-1b-visa-requirements/**

[8] Student visa categories: **https://travel.state.gov/content/travel/en/us-visas/study/student-visa.html**

Table 1. Percentage coverage of the bibliometric fields in PERM-SB-WoS

Field	Coverage
Number of Publications	100%
Impact Percentile	99.95%
Research Topic	97.65%
Collaborating Country	99.89%

The linkage between PERM-SB and WoS is predicated on the individual's having published. The PERM-SB-WoS data set contains FBSEs who have published research in peer-reviewed journals.

3.2. Assessment of the representativeness of the PERM-SB-WoS data set

Having established the feasibility of creating a high quality and procedurally valid FBSE data set, we turn to the question of data representativeness. Standards of data representativeness are defined in relation to the data's intended uses. Data intended for use by decision-makers should be both error free and representative of the population about which the decisions are to be made. On the other hand, researchers may focus on data quality itself, examining data sets for systematic or random errors including problems attributable to data collection methods, null values, data redundancy and outliers. Determining use cases for the data set produced in this project necessitates an assessment of its quality.

While the individual data sets used in this project are highly reliable, the process of merging them necessarily introduces certain questions *vis a vis* the objective of creating FBSE profiles. Namely, is the resulting data set representative of the PERM data population (2016-2019) as well as the larger 1.34M FBSE population from Steppingblocks? We address various facets of this question below.

Do the data adequately represent the FBSEs among the various classes of visa holders? The table below, detailing the distribution of the FBSE's visa types, confirms that our linked data set (78K) is biased towards H-1B visa holders.

Table 2. Evaluation of representation against PERM 2016-2019 data set

Visa type (Top 5)	2016-2019 PERM (N=~446,000)	PERM-SB data set (n=~78,000)
H-1B	0.701	0.813
L-1	0.069	0.078
F-1	0.059	0.062
TN	0.018	0.017
Outside US	0.034	0.006
Other	0.057	0.016

Do the data adequately represent FBSEs from various undergraduate majors? Given the preponderance of H-1B visas in the PERM data set, which have been historically issued to people in the high-tech industries, we note that the PERM-SB data set over-represents those with Computer Science and Engineering majors.

Table 3. Comparison of individuals represented in SB and PERM-SB

Majors (Top 6)	FBSE (%) (N=1.34 million)	PERM-SB (%) (n=78,000)	Representation Ratio
Computer Engineering	1.8	5.2	2.88
Electronics Engineering	1.9	5.0	2.61
Electrical Engineering	4.4	10.9	2.50
Computer Science	9.8	21.2	2.15
Engineering	5.9	8.0	1.37
Business Administration	2.4	2.2	0.93
Other	73.8	47.5	0.64

Hence, we conclude that our PERM-SB data set is not representative of the full FBSE population.

Is there a bias in the distribution of highest degrees attained in PERM-SB-WoS? Because the WoS data set has only published researchers, who are more likely to have doctorate degrees, the FBSEs in the PERM-SB-WoS data set are likely to over-represent doctorate degrees. The data displayed in Table 4 of the overall proportion of highest degree levels among all degree holders for the PERM-SB and PERM-SB-WoS data set confirms that observation.

Table 4. Comparing the Highest Degree Level between the PERM and PERM-SB-WoS data sets

Highest Degree Level	Percent of PERM-SB data set	Percent of PERM-SB-WoS data set	Representation Ratio
Bachelor's or Master's	89.0	57.1	0.64
Doctorate	11.0	42.9	3.88

Table 5 demonstrates the high percentage of Computer Science and Engineering degrees in PERM-SB. The concentration in Computer and Information Sciences is somewhat attenuated in PERM-SB-WoS. However, these concentrations are reflective of the preponderance of H-1B visas in PERM.

Table 5. Comparing the Top Eight Highest Degree Categories between the PERM-SB and PERM-SB-WoS data set

Highest Degree Category	Percent of PERM-SB data set	Percent of PERM-SB-WoS data set	Representation Ratio
Computer And Information Sciences	42.8	36.3	0.85
Engineering	29.8	31.9	1.07
Business	6.3	3.3	0.53
Mathematics & Statistics	2.0	3.4	1.56
Physical Sciences	2.0	4.7	2.32
Finance	1.6	1.2	0.73
Social Sciences & Liberal Arts	1.3	2.5	1.90
Marketing	1.2	0.8	0.69

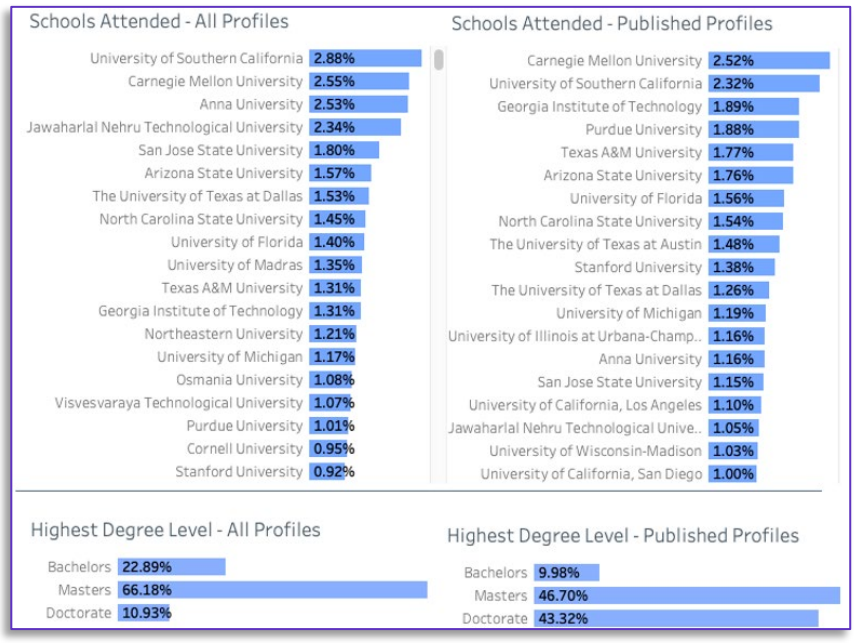
Other	4.0	5.3	1.32

4. Analysis of the PERM-SB-WoS data

We begin our discussion of the analysis of the data with an important caveat and a warning regarding the interpretation of the numbers. As we have mentioned, the PERM-SB-WoS data set is not a representative sample of the FBSE population in the US. However, we want to illustrate the utility and power of the data to address research and policy questions by providing examples of the types of analyses that could be conducted with a representative data set. Hence, the analysis results are illustrative. Any patterns or trends observed in the results of the analysis are purely coincidental and should not be used to draw inferences about the FBSE population in the US. The PERM data set we used to validate our data has a substantial number of FBSEs who entered the US on H-1B visas, however, we do not know whether this substratum of the FBSE population in our data set is representative of the stratum of all H-1B visa holders.

To familiarize readers with the data we begin with select descriptions of the contents of these data. An important criterion for uniquely identifying foreign-born individuals is to find links to their educational history in the country in which they were born and similarly to track them by their educational histories in the US. For instance, Figure 3 illustrates the ability of the PERM-SB-WoS data to identify the schools attended by FBSEs in the US and the highest degree earned. In addition to the information on the degree, our data also include information on their undergraduate major and disciplinary training should they go on to earn more advanced degrees.

Figure 3. Most frequently attended schools and highest degrees attained.



Such data when coupled with information on wages (Table 6) disaggregated by the type of degree can provide researchers, university administrators, and policy makers with basic information about educational institutions, degrees they strive for and majors that are most frequently represented among FBSEs. We can also use such data to examine variations in these choices based on the country of origin, the type of entry visa and other demographic characteristics included in the hundreds of variables in the PERM-SB-WoS data set.

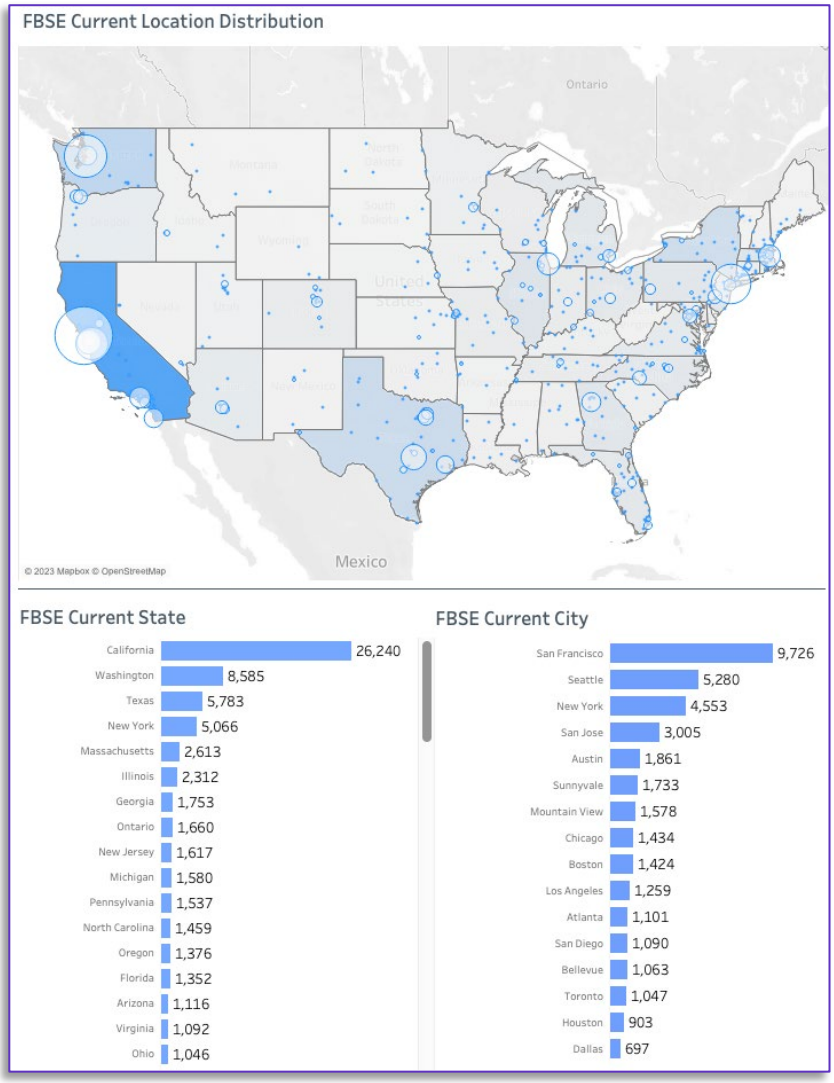
Table 6. Wages by degree attained

Entry Degree Level	Population (n)	Avg. Entry Salary	Avg. Current Salary
Bachelor's	25,917	100,591	103,401
Master's	45,844	99,588	103,628
Doctorate	5,707	83,334	103,342

These data are invaluable for demographers and students of migration patterns and policy analysts attempting to understand trends in foreign born arrivals and the opportunities they seek in the US.

Our data not only indicate the age and other demographic characteristics of the individuals, where they first settled or the university they came to attend, but also where they moved to as they entered the workforce (Figure 4). Face validity can be established by noting the preponderance of H-1B visa holders in our data set and their concentrations in California and Washington in the west and other locations with large corporations with a strong demand for a workforce with specialized computer science-oriented skills.

Figure 4. Distribution of FBSEs across the US



Other policy relevant information such as wages (Tables 6 and 7), trends in wages (Table 6), and locations (Figure 4) can be combined and used to inform policies

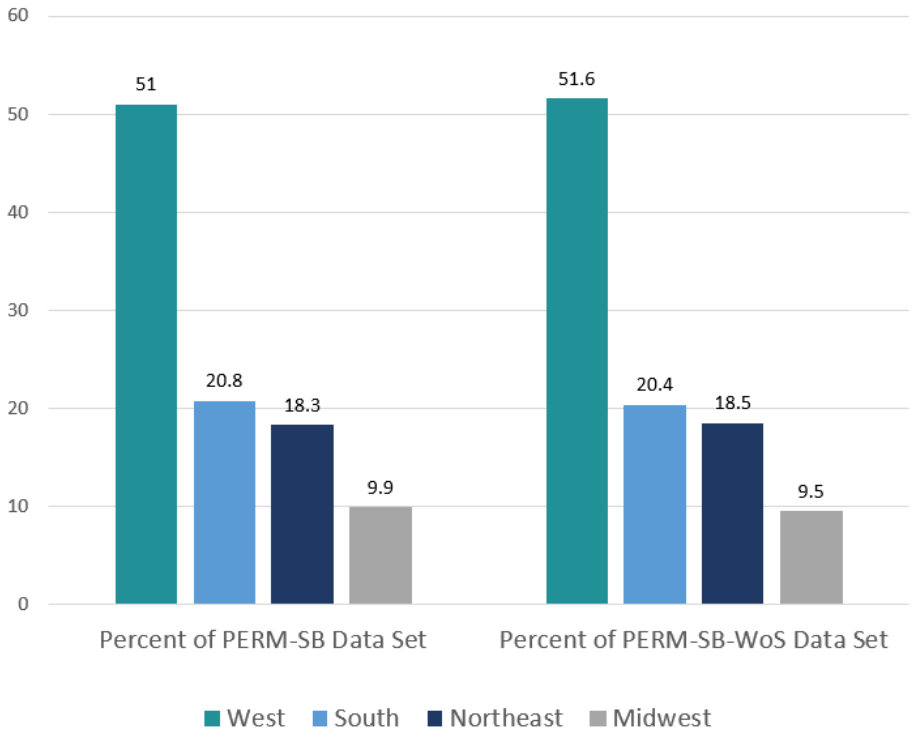
regarding subsidies and incentives to encourage FBSEs to settle in targeted (EPSCoR)[9] locations (Table 7).

Table 7. Wages of FBSEs in and outside of EPSCoR jurisdictions

	EPSCoR Jurisdictions	Non-EPSCoR Jurisdictions
Population (n)	2,228	75,230
Avg. Entry Salary	$72,448	$99,510
Avg. Current Salary	$103,884	$103,521

Examining the augmented PERM-SB-WoS data set, which includes their publication characteristics, provides another example of how these data can be used to compare different strata within our data set. For instance, the comparison between the PERM-SB data and its subset PERM-SB-WoS data illustrates (at a different level of granularity) the feasibility of examining within and between group differences in the FBSE population (Figure 5).

Figure 5. Comparing regional employment observed in PERM-SB and PERM-SB-WoS



Comparative analyses between the FBSE population and its substratum (FBSE with publications) shown in Figure 6 (employers) and Figure 7 (industries in which they are employed) further illustrates the richness of the data set and its potential value for researchers and relevance for policy making.

[9] A number of jurisdictions (state, federal territory, or commonwealth) have been targeted for funding by the Established Program for Stimulating Competitive Research (EPSCoR) Program by NSF and other federal agencies to enhance their research competitiveness. **https://beta.nsf.gov/funding/initiatives/epscor**.

Figure 6. Top Employers

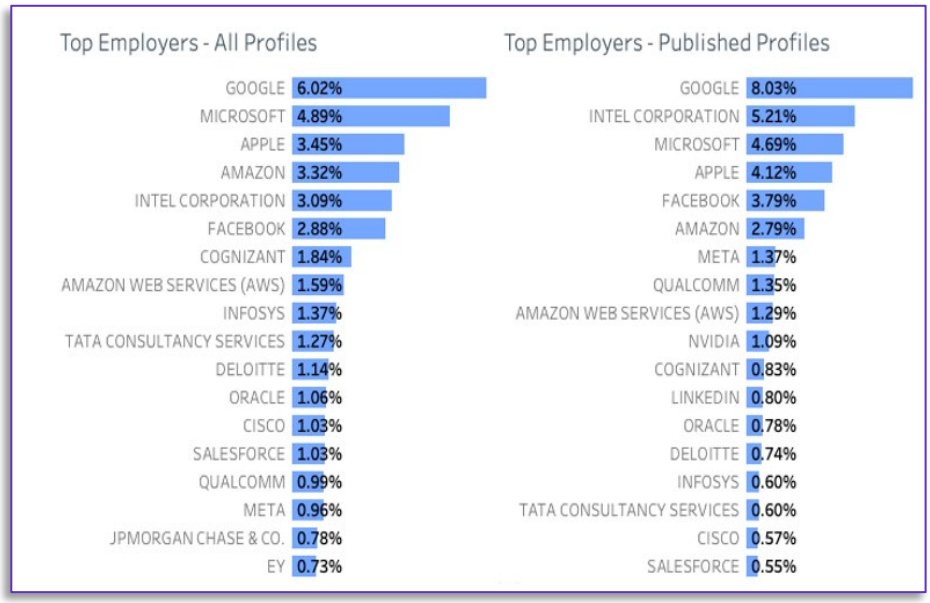
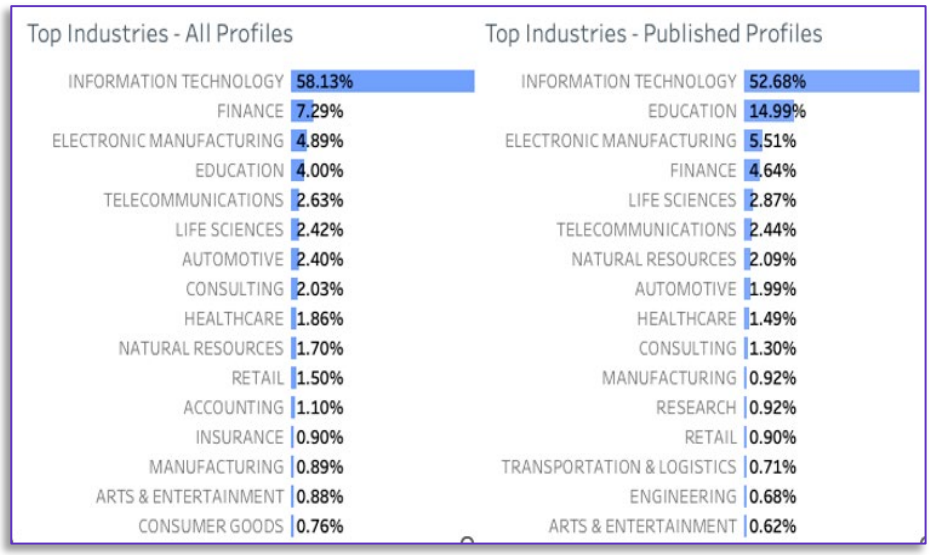


Figure 7. Top Industries



We turn next to analyses that illustrate the potential for research and relevance for research policy of the substratum of our dataset that consists of FBSEs who have published in scholarly journals indexed in WoS. We return to the policy themes, of the location of these individuals, their research topics and the influence or impact of their research. These illustrative examples, when coupled with the potential analyses shown above, using the full PERM-SB data set, demonstrate the potential and relevance of such data to support sophisticated policy analyses.

Table 8 complements Figure 8 in illustrating the distribution of FBSEs with doctorates across the country. It also illustrates the loss of information in aggregating the data into four regions rather than showing the distribution of the FBSEs across the fifty states. A potential analysis based on metropolitan areas would show yet another clustering of FBSEs that would provide a perspective that differs from the state and the regional level distribution of FBSE.

Table 8. Comparing Doctorate Percentage by Worker Region

Worker Region	Percent Doctorates
West	11.4
South	10.8
Northeast	10.8
Midwest	10.3

The seemingly even distribution of the FBSEs across the four regions (Table 8) is in sharp contrast to that shown in Figure 4 and highlights the relationship between data granularity and the types of research the data can support. Doctorate holders produce the highest impact research, and it is expected that the composition of doctorates would be strongly related to research impact. Indeed, the even distribution of doctorate holding FBSEs across the four regions (Table 8) mirrors the evenness of the median impact percentile by region (Figure 8). On the other hand, our approach also highlights the richness of our individual level data with the large number of variables for each FBSE that can support analyses along a broad range of socio-economic-geographic and demographic characteristics along with education and career characteristics.

Using the same regional classification of the FBSEs, Figure 8 shows the median value of the impact indicator of the individual's publications over the course of the individual's career. Such analyses are useful for determining which variables are associated with researchers who publish the most impactful work. Hence, the topic in which the FBSEs in our data set have published the most impactful research is Electrical Engineering, Electronics & Computer Science regardless of FBSE geographical region.

Figure 8. Comparing Median Impact Percentile by Worker Region and Top 5 Research Topics

Research Topic	West	South	Northeast	Midwest
Chemistry	47	45	51	51
Clinical & Life Sciences	42	41	41	32
Electrical Engineering, Electronics & Computer Science	55	52	54	54
Physics	46	43	49	46
Social Sciences	46	49	45	42

Worker Region

Figure 9 displays information on the most frequent job categories in which the FBSEs are working. How individuals report job categories is not standardized, which makes the linking of research topics in which the FBSEs are publishing to job categories such as "science" and "education" imprecise. This lack of compatibility in taxonomies is also not a unique feature of our data set. For instance, changes in the nature of work have led to multiple classification schemes making the crosswalk from one categorization to another difficult[10].

Figure 9. Comparing Median Impact Percentile by Worker Region and Top Five Current Job Categories

Job Category	Worker Region: West	South	Northeast	Midwest
Education	54	50	53	53
Engineering	45	45	45	45
Information Technology	45	46	46	44
Management	44	42	39	29
Science	59	57	62	50

These illustrative analyses of the FBSE profiles in the PERM-SB-WoS data set suggests a broad range of options for more detailed examinations. The potential of these data for policy analysis is perhaps more exciting than the methodological opportunities in that these analyses point to a broad range of policy options within the control of policymakers at the state and federal levels. They also point to potential triple helix collaborations among government, universities, and industry to enhance both the foreign born and native STEM workforce.

[10] National Research Council. 1999.*The Changing Nature of Work: Implications for Occupational Analysis*. Washington, DC: The National Academies Press.

5. Lessons learned

This project was a true learning experience along multiple dimensions. Linking data sets has a multiplier effect on their utility that should not be underestimated. At the same time, linking data sets is simultaneously difficult, time consuming and yet, a rewarding experience. Our processes for ensuring high quality linked data, while predicated on the unique characteristics of the Steppingblocks profiles and the Clarivate publications data, have yielded transferable insights that apply to all such endeavors. While we are still processing the lessons we have learned, we mention, below, some that are specific to this project and discuss others that have broader relevance in the Conclusion.

- Given our reliance on the publicly available PERM data set and the lack of access to other government data for validation, our resulting data set reflects the biases inherent in the PERM data set, that is, there is over-representation of engineering and technology fields. Obtaining a data set that is representative of each of the scientific fields will require access to other, more representative, data sources. This lack of good data sources cannot be overcome by devising better modeling and methodological solutions. By good sources we mean those that represent the diversity of the FBSE population which can be used for validating the data solutions from data providers such as Steppingblocks.
- The overwhelming dominance by citizens from China or India in our data set requires that we conduct analysis with and without profiles from these FBSEs originating in these two countries. Such selective analyses will be necessary to prevent eclipsing meaningful patterns in the data on citizens from other countries.
- Our data set does not contain information on people whose applications for entry to the US were denied. To gain insight for policy analysis it is useful to examine the characteristics of the full range of individuals seeking to enter the US. Inclusion of individuals whose applications were rejected or withdrawn is necessary to conduct a proper policy analysis of the consequences of changing the eligibility criteria for entry and potential residence in the United States.
- The absence of the full range of data on individuals seeking entry to the US led us to explore methodological solutions to compensate for the lack of representative data for validation. We see great potential in developing predictive models using innovative career vectorization methods such as career2vec that could lead to more robust induction processes as alternatives to the one we used in the early filtering of the Steppingblocks career profiles.
- The linkage between PERM and WoS successfully identified a data set of FBSEs who have published peer-reviewed research. This data set can be used to characterize differences in the regional distribution of FBSEs within the US, the research productivity and impact of FBSEs, as well as FBSE contributions to strategic research areas of interest. Further expansion of the source data set will enable a deeper analysis of the FBSE population and its contribution to the US science and technology enterprise.

6. Conclusion

In identifying a FBSE data subset from Steppingblocks' data, linking it to publicly available OFLC PERM data, and successfully enriching it with publication details via Clarivate's Web of Science, the team demonstrated its ability to build a robust process to create a rich data set consisting of information on FBSEs' education background, career trajectories, and research contributions. An assessment of the resulting data set confirmed the high-quality of the data, despite some representation challenges, which reflect, in part, the existing biases of the 2016-2019 OFLC PERM data set selected for our proof of concept.

We have demonstrated that it is possible to create a high-quality data set suitable for analysis, albeit of a limited group of FBSEs who entered the US through visa categories that were dominated by H-1B (highly specialized temporary workers) and L-1 (intracompany transferees) visas. This experience highlights the value of access to the full range of entry visas to create data sets that are representative of the FBSE population in the US.

This problem of lack of full representation of the FBSE population is not unique to our data set. As noted in NSF's Science and Engineering Indicators report[11] the analysis of FBSE is often conducted by visa type. For instance, the analysis of foreign-born undergraduates is based on individuals who entered the US on F-1 (student) visas whereas the analysis of individuals with more advanced degrees includes individuals who entered the US on J-1 (exchange visitor) visas. Analysis for policy purposes, say, eligibility criteria, is probably best done on individuals in the STEM workforce who entered the US on H-1B (highly specialized temporary workers) or L-1 (intracompany transferees) visas. Another consideration in the analysis of FBSE career trajectories is that the taxonomies used by the immigration services and NCSES surveys and other statistical data sources to classify foreign born individuals or labor categories are not always compatible[12].

Although our final PERM-SB-WoS data set is relatively small, it captures the richness of and variety of the FBSE profiles. With access to the relevant government data which would enhance the representativeness of our data, it is possible to conduct multiple analyses and simulations to examine the potential outcomes of policies that the government can implement. One obvious example of such analysis is the use of simulation to model different eligibility criteria for different visa categories to explore the potential consequences of the changes in these criteria on the inflow of individuals with different professional qualifications and skill sets.

In a similar vein, it is possible to use these data to inform simulations of the consequences of forces beyond the control of US policy makers such as changes in the flow of potential FBSEs from specific parts of the world due to political or other considerations (as experienced during the COVID lockdown).

Further downstream, it is conceivable that a more machine learning-oriented approach to representing individuals' sequential education and career milestones using tools such as word2vec[13] could help develop a novel career model (career2vec)

---

[11] **The STEM Labor Force of Today: Scientists, Engineers, and Skilled Technical Workers,** https://ncses.nsf.gov/pubs/nsb20212/participation-of-demographic-groups-in-stem

[12]The STEM Labor Force of Today: Scientists, Engineers, and Skilled Technical Workers **https://ncses.nsf.gov/pubs/nsb20212/u-s-stem-workforce-definition-size-and-growth**

[13] **https://towardsdatascience.com/word2vec-explained-49c52b4ccb71**

trained on the Steppingblocks' corpus of career profiles to identify quickly and reliably FBSEs while eliminating a large part of the manual curation and data checks that were needed in this project. This approach could also help develop more scalable and reliable models of career trajectories for examining different professional lifecycles and their relationship to education, training, and careers.

In light of our experience in developing our proof of concept and the findings and the lessons learned from the multiple projects involved in this pilot for America's DataHub, we believe that there is sufficient information to focus future efforts on important research questions and research priorities so that both the data set generation and analysis can yield actionable insights. For instance, our FBSE data set contains information that can be refined to model and estimate FBSEs' societal and economic contributions.

Looking to the future and the eventual establishment of the National Secure Data Service (NSDS)[14], this pilot effort reinforces the need to address the challenges identified throughout this project to ensure data accessibility, interoperability, and linkages. On a more optimistic note, this project demonstrates that while there are obstacles, progress can be made, and more opportunities lie ahead to bring together data sets on which to conduct analysis to examine the role and contributions of FBSEs in the US workforce.

[14] CHIPS act of 2022, PUBLIC LAW 117–167—AUG. 9, 2022.
https://www.congress.gov/117/plaws/publ167/PLAW-117publ167.pdf