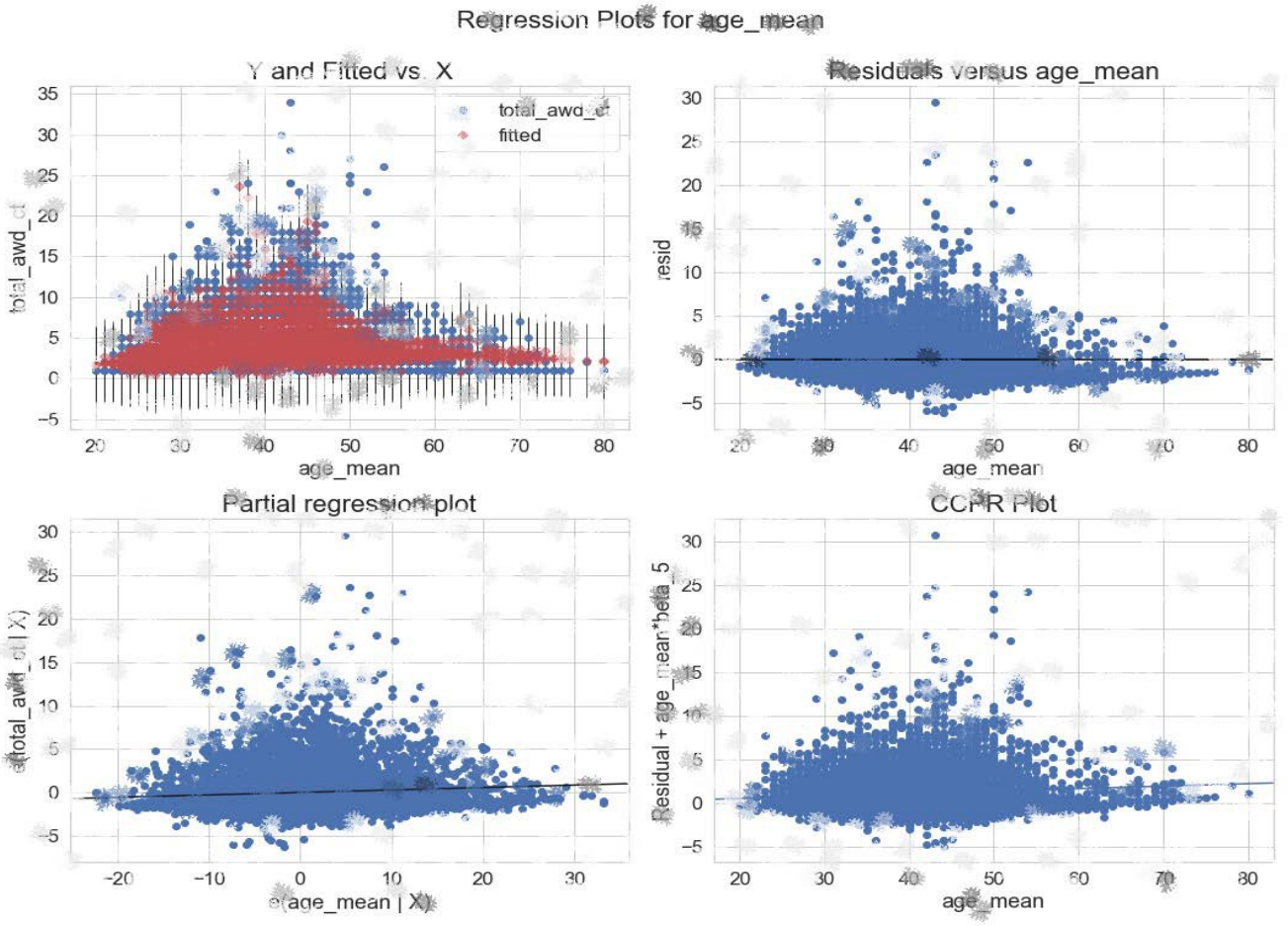
Summary Documentation for the IRIS UMETRICS 2020 Data Release



INSTITUTE FOR RESEARCH ON INNOVATION & SCIENCE

June 2020

Data Access Statement

This is the publicly available summary documentation for the 2020 IRIS UMETRICS dataset. Access to the full documentation is restricted to authorized IRIS data users.

Citation

Nicholls, N., Ku, R. L., VanEseltine, M., & Owen-Smith, J. (2020). Summary Documentation for the IRIS UMETRICS 2020 Data Release. Institute for Research on Innovation and Science (IRIS). https://doi.org/10.21987/9WYN-8W21

Table of Contents

About the IRIS Data Release

IRIS has produced a de-identified research data release for researchers in the IRIS VDE each spring since an initial release in April 2017, and we are pleased to announce the fourth release in June 2020. As in the past, IRIS UMETRICS data release files are available in two environments, the IRIS Virtual Data Enclave (VDE) and the U.S. Census Bureau FSRDC system. With the IRIS membership growth, the annual release continues to grow (see Table 1).

Table 1: Data Growth from Past Releases

	2017 Release	2018 Release	2019 Release	2020 Release
Number of Universities	19	26	31	33
Number of Files	14	15	18	18
Number of All Awards	176,971	296,253	392,125	442,945
Number of Federal Awards	90,827 (51%)	151,816 (51%)	215,628 (55%)	251,922 (57%)
Number of Employees	333,944	478,815	643,463	720,679
Number of Vendors	237,690	582,797	821,420	902,351
Number of Subawards	9,139	13,262	21,888	23,552
Number of Awards (as funding source of subawards)	12,282	22,212	30,691	35,805
Award Total Direct Expenditures (indirect cost excluded)	$ 36.4 billion	$ 61.6 billion	$ 83.5 billion	$ 98.9 billion
Vendor payment total	$ 18.1 billion	$ 18.7 billion	$48.5 billion	$ 25.9 billion
Subaward payment total	$ 6.0 billion	$ 8.5 billion	$12.6 billion	$14.9 billion

2020 Dataset

Release Highlights

IRIS is pleased to announce new features to this year's annual release. Highlights include:

1. Adding an indicator for federally funded sponsored projects;
2. Adding a new data field for standardized (official) names of nonprofit foundations as part of a continuous effort toward cleaning and disambiguating funding source names;
3. Adding a new field for standardized sub-organization unit names by applying 18 categories (e.g., Arts & Sciences, Health Sciences, Administration, Graduate School, etc.) which we hope enables a cross-sectional analysis with a focus on units through which grants are administered and research activities are performed;
4. Expanding demographic variables (imputed gender and ethnicity, as well as, age) that help to better characterize employees who are paid on sponsored projects;
5. Advancing record linkage work by linking federal award (NIH and NSF) and UMETRICS data at both award and individual Principal Investigator (PI) levels; and,
6. Providing new linkage files focusing on research teams in which team leaders are identified and verified as NSF and/or NIH PIs and are connected to their team members who are paid on the same sponsored project.

We would like to make a special note that many of these changes are a product of our communication and discussion with IRIS researchers who made significant contributions to this annual release through their data inquiry, exploration, experimentation, and implementation since the last release. We continue to advance release data quality and quantity in order to further meet researchers' needs through future supplementary and annual releases.

Dataset Overview

The 2020 dataset for research is based on the fourth quarter 2019 Census data transfer (UMETRICS 2019Q4) and contains data from 33 IRIS universities (see Appendix A) including coverage between 2001 and 2019 (FY2001-2018). This coverage varies by institution (see Data Summary section for more details). Data profile information (file name, record count, and file size if exported in csv) is shown in Tables 2, 3, and 4.

Core Collection

The core collection includes data submitted to IRIS by IRIS member universities. These files include university financial and personnel administrative data pertaining to sponsored project expenditures at each university during a given year, drawn directly from sponsored projects, procurement, and human resources data systems on each IRIS university's campus. Individual campus files are de-identified, cleaned and aggregated by IRIS to produce the core collection files. The 2020 release includes transactions from about 440,000 unique federal and non-federal awards including wage payments to about 700,000 individuals as well as transactions to about 900,000 unique vendors (both organizations and individuals). In addition, about 23,500 unique organizations / institutions received subawards from IRIS universities transferring their prime awards. About 36,000 unique awards were used by IRIS universities as the funding source to transfer subawards to subrecipients.
Vendor and subaward payment by awards total $100 billion. (See Appendix B UMETRICS Core File Relationship in Monetary Flow).

Table 2: Release 2020 Core Files

File Name	SQL Table Name	Record Count	File Size (csv)
award	release2020.core_award	10,506,992	2,855,655 KB
employee	release2020.core_employee	26,376,022	4,968,165 KB
vendor	release2020.core_vendor	23,982,459	5,451,092 KB
subaward	release2020.core_subaward	785,522	212,242 KB
For file & field descriptions, see the next section or the 2020 Data Dictionary			

Auxiliary Collection

The auxiliary collection provides researchers with contextual information on institutions and demographic information about UMETRICS employees who are paid on sponsored projects. Files can also help to retrieve more details about the purpose of each employee / vendor / subaward transaction and sub-organization units (including standardized sub-organization unit names) where grants are administered and research is conducted.

Table 3: Release 2020 Auxiliary Files

File Name	SQL Table Name	Record Count	File Size (csv)
institution_fastfacts	release2020.aux_institution_fastfacts	594	69 KB
comprehensive_award_list	release2020.aux_comprehensive_award_list	443,074	21,032 KB

cfda	release2020.aux_cfda	6,359	1,015 KB
suborganization	release2020.aux_suborganization	2,569	155 KB
emp_demographics	release2020.aux_emp_demographics	548,483	42,989 KB
object_code	release2020.aux_	6,373	283 KB
For file & field descriptions, see the next section or the 2020 Data Dictionary			

Linkage Collection

Finally, the linkage collection includes crosswalks between IRIS data and external datasets (e.g., federal award and publication data) at the award level. The improved award match rate (in both NSF and NIH award linkage) reflects our continuous effort to minimize false positives and false negatives. As indicated above, we expanded award linkage from award to individual level this year by linking UMETRICS employee name and NSF / NIH Principal Investigator (PI) names. Based on individual name matching, we developed two team-focused data files that can be useful in addressing research questions concerning team characteristics (such as diversity) and their impact on research performance and productivity if team files are linked to Core files.

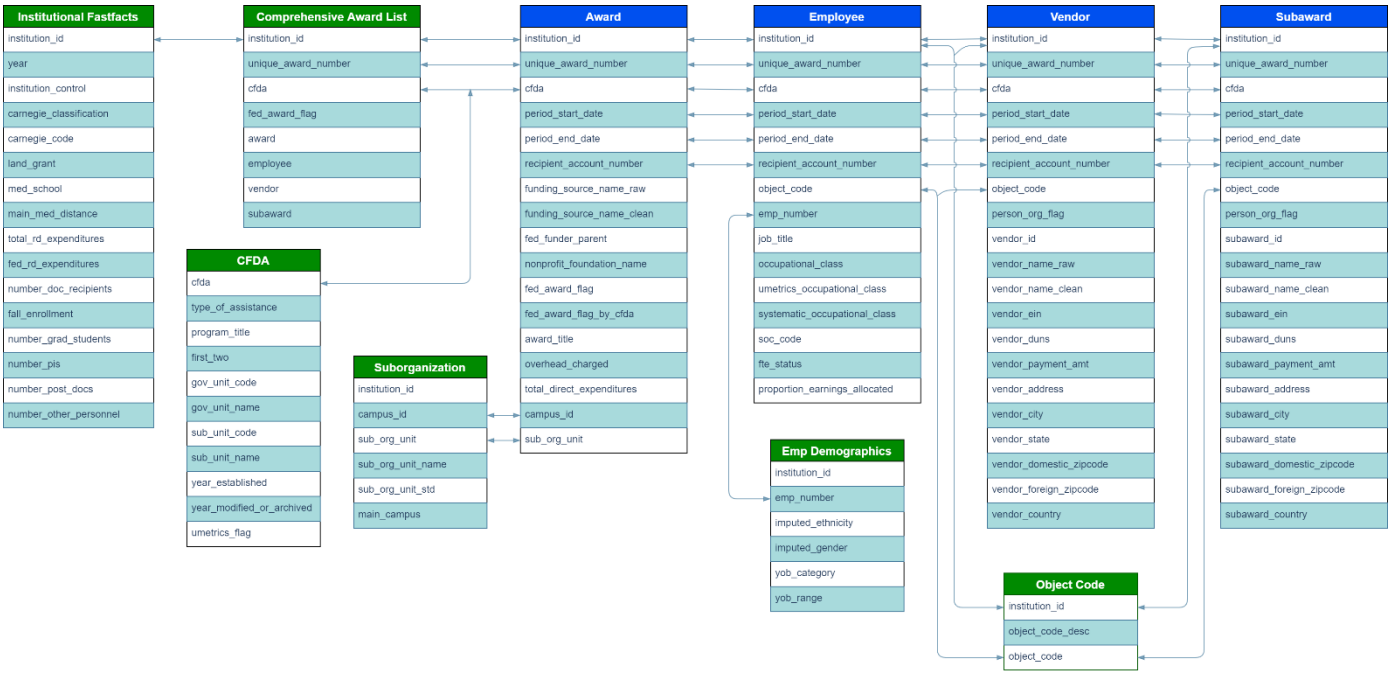
Table 4: Release 2020 Linkage Files

File Name	SQL Table Name	Record Count	File Size (csv)
nih_xwalk	release2020.link_nih_xwalk	86,957	3,990 KB
nsf_xwalk	release2020.link_nsf_xwalk	35,244	1,285 KB
nih_pub_xwalk	release2020.link_nih_pub_xwalk	3,066,310	169,995 KB
nih	release2020.link_nih	1,706,524	5,745,798 KB
nsf	release2020.link_nsf	246,941	606,513 KB
nih_pub	release2020.link_nih_pub	2,422,831	541,693 KB
team_leader_award_year	release2020.link_team_leader_award_year	340,754	40,170 KB
team_leader_member_year	release2020.link_team_leader_member_year	2,423,223	356,877 KB
For file & field descriptions, see the next section or the 2020 Data Dictionary			

Relationships between Collections

In the figures below, we provide a visual representation of what data field elements link between files. The data diagram (Figure 1) shows Core and Auxiliary files with linking elements highlighted.

Figure 1: Data Diagram for Core and Auxiliary Files



The following data diagrams (Figures 2 & 3) show relationships between the Core Award file and Linkage collection files (award and team linkage files respectively).

Figure 2: Data Diagram for Core and Linkage Files (Award Linkage Files)

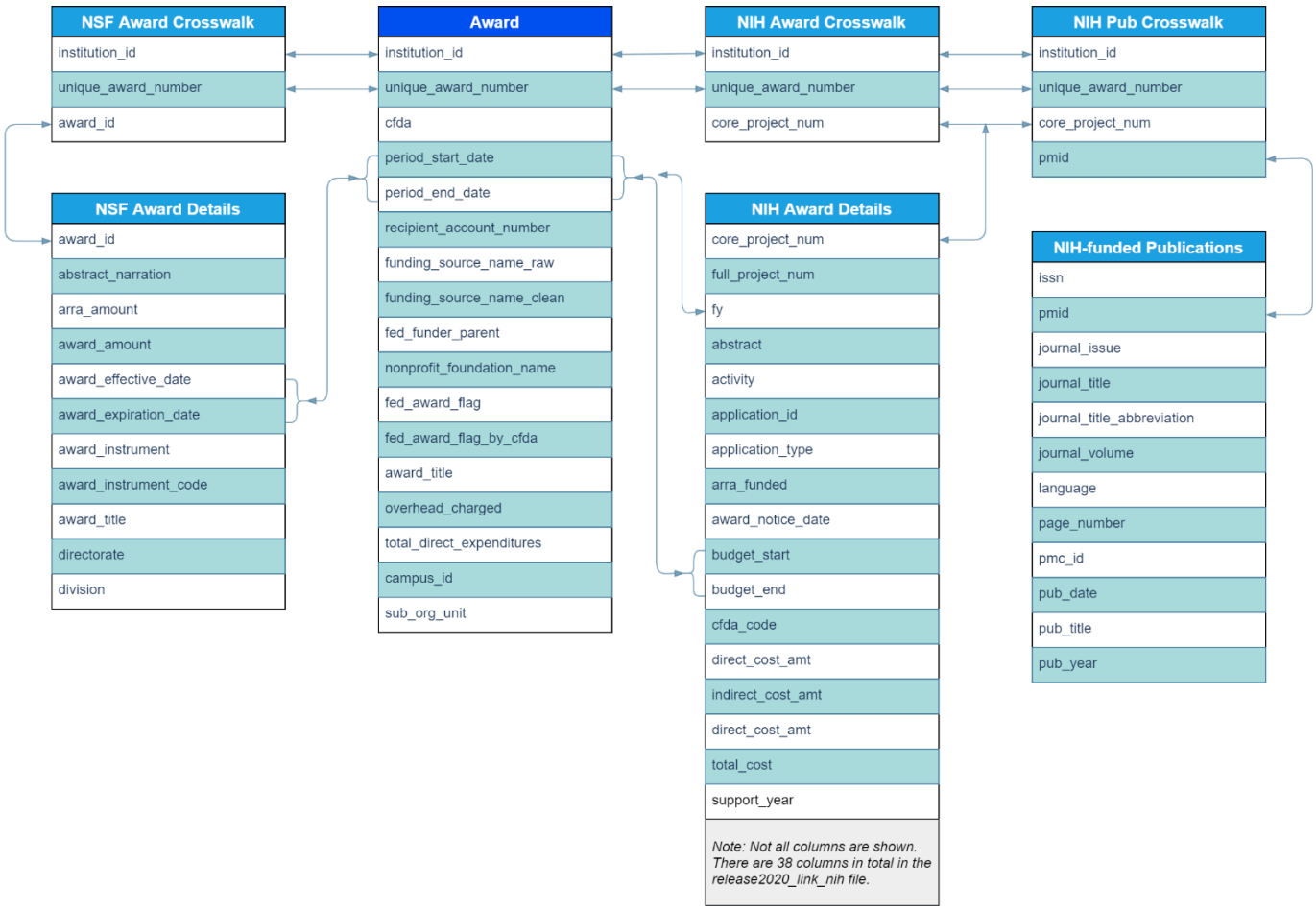
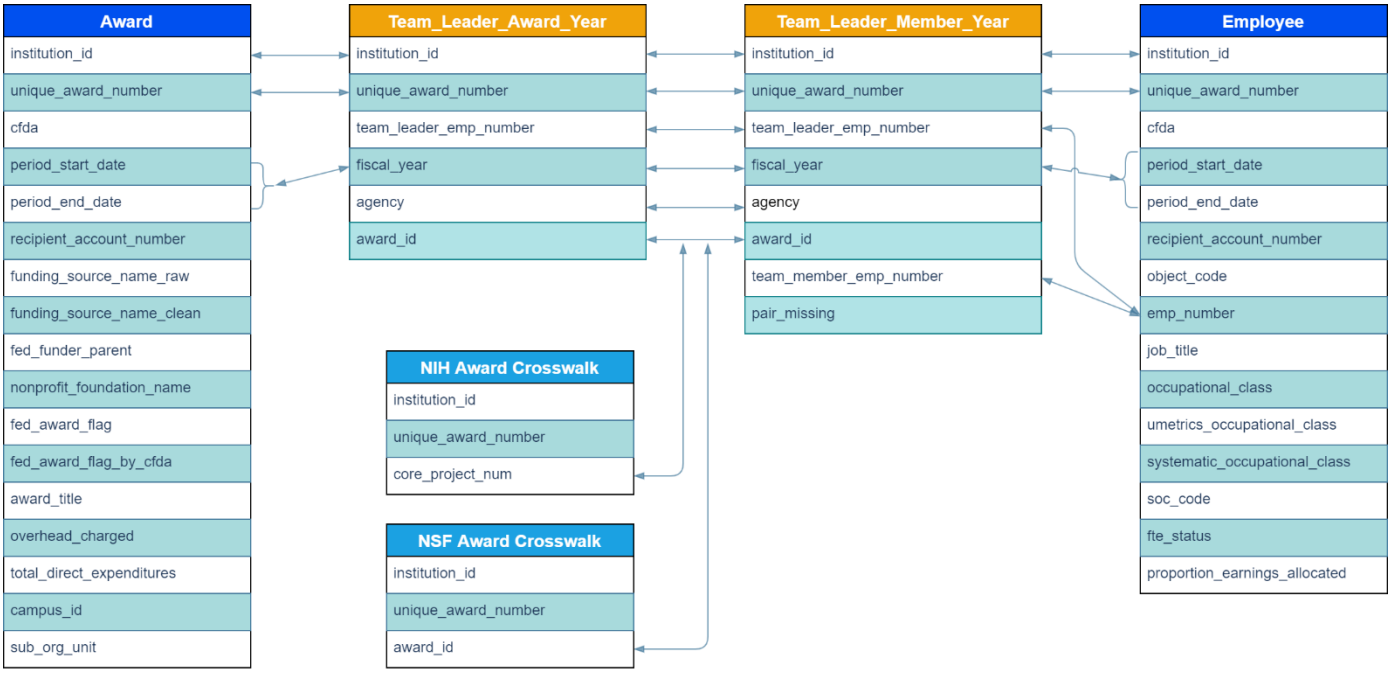


Figure 3: Data Diagram for Core and Linkage Files (Award and Team Files)



File & Field Descriptions

Note: In selected files, we included data questions addressed by researchers and our answers to them. In addition, we included our notes regarding data fields that are new to this current release. Linkage-related questions are incorporated in the Methodology

Award

File Details

File Name: core_award
Record Counts: 10,506,992
Field/Column Counts: 17

File Summary

The Award file is the centerpiece of the release dataset and directly connects to other files, thus a fundamental source for record linkages. This file contains transaction data on every sponsored project that has direct or overhead (also commonly referred to as indirect) expenditures during the period of time covered in the file This file can be considered as an award profile in that types of sponsors can be identified through a federal grant indicator, CFDA number, funding source names, and grant administering sub-organization units on campus. The file includes all funded awards that IRIS universities received during a given year. Awards include (but are not limited to):

(1) Research-related,
    i) Federal and
    ii) Nonfederal awards, and;
(2) Non-research related activities such as work-study programs.

Data Fields

Fields appearing across files	Fields unique to this file
institution_id	funding_source_name_clean
unique_award_number	funding_source_name_raw
cfda	fed_funder_parent
period_start_date	award_title
period_end_date	overhead_charged
recipient_account_number	total_direct_expenditures
campus_id	fed_award_flag_by_cfda
sub_org_unit	nonprofit_foundation_name
fed_award_flag	

Table 5: Core_Award Data Fields

Field Name	Column Name	Data Type	Set Length	Max Length	Field Definition
Award Title	award_title	varchar	500	496	Title of award
Campus ID	campus_id	varchar	50	8	IRIS-generated identifier assigned to each campus for de-identification purposes. Each ID is a combination of institution_id and a serial number helpful to identify the campus to which a given award is made and by which the award is being administered
CFDA	cfda	varchar	10	6	A five-digit Catalog of Federal Domestic Assistance (CFDA) number retrieved from the unique_award_number that universities submit. This number is assigned to awards to represent the source of funding. If the first two digits are between 10 and 98, awards are federally funded. If the first two digits range between 00 and 09, or 99, sources of awards are most likely non-federal
Federal Award Flag by CFDA	fed_award_flag_by_cfda	int	4	4	A binary code to identify types of awards based on the first two-digit number of CFDA; coded 1 if the award is federal; coded 0 if non-federal
Federal Award Flag	fed_award_flag	int	4	4	A binary code to identify types of awards based on the first two-digit number of CFDA and funding source name; coded 1 if the award is federal; coded 0 if non-federal
Funding Source Name Clean	funding_source_name_clean	varchar	200	92	Cleaned name of funding source assigned to each project; if university is a recipient of a prime award, the funding source name is the direct sponsor; if university is a subrecipient of an award, this is the name of the pass-through entity (exceptions noted in documentation)

Federal Funding Agency (Parent)	fed_funder_parent	vachar	200	50	Name of federal sponsoring department or parent agency, e.g., Department of Health and Human Services for NIH awards
Funding Source Name Raw	funding_source_name_raw	varchar	200	104	Raw name of funding source assigned to each project; if university is a recipient of a prime award, the direct sponsor is the name of funding source; if university is a subrecipient of an award, this is the name of the pass-through entity (exceptions are noted in documentation)
Institution ID	institution_id	int	4	4	IRIS-generated unique identifier assigned to each IRIS member university for de-identification purposes. Values are four or five digit numbers
Nonprofit Foundation Name	nonprofit_foundation_name	varchar	200	88	Disambiguated name of a nonprofit foundation closely corresponding to the funding source name in the field of 'funding_source_name_clean'; this field is null if a funder is not a nonprofit foundation
Overhead Charged	overhead_charged	numeric	9	9	Overhead charged to the award in the specified period in actual dollars; Overhead charge is also commonly referred to as indirect cost
Period End Date	period_end_date	date	3	3	End of period in which a monthly expense transaction took place; each period end day is the last day of a month: e.g., 3/30/2008 or 12/31/2014
Period Start Date	period_start_date	date	3	3	Beginning of period in which a monthly expense transaction took place; each period start day is the first day of a month: e.g., 4/1/2009 or 10/1/2015
Recipient Account Number	recipient_account_number	varchar	50	18	A university's internal account number to uniquely identify each project; typically an accounting code used to allocate funds received from an award

Sub-organization Unit	sub_org_unit	varchar	100	12	IRIS-generated identifier assigned to sub-organization units to which each funded project is assigned, such as a particular college within a given IRIS member university. (This is not at the level of individual departments.) Each ID is a combination of campus_id described above and a serial number assigned to each sub-organization unit within each campus. This ID helps to identify the college or unit to which a given award is made and by which the award is being administered at a lower-level
Total Direct Expenditures	total_direct_expenditures	numeric	9	9	Total direct expenditures charged to the award in the specified period
Unique Award Number	unique_award_number	varchar	500	100	University-generated unique identifier specifying an award and its funding source, made up of the 5-digit funding source code (e.g., CFDA number) and an award identifier. Award identifier may include the awarding agency's federal award ID (e.g., federal grant number, contract number, or loan number) or an internal award ID for non-federal awards. Values may include a space or dash in between them: e.g., “10.310 2010-12345-54321” (USDA example), “47.050 1234567” (NSF example), “93.865 2-R01-DK-012345-15-S1” (NIH example), “00.000 1234567” and “00.200 State Award 1” (Non-federal grant examples)

Q&A re. Award File

1. Why is the Award file a centerpiece in the IRIS annual release dataset?

This is because UMETRICS data are centered around university financial and personnel administrative data pertaining to sponsored project expenditures submitted by IRIS member universities. Each IRIS member university contributes records from its sponsored projects, procurement and human resources systems. While other sources, such as NSF HERD (Higher Education Research and Development) survey data or USAspending.gov, provide some information on research spending, our Award file sits at the center of employment, vendor purchases, and subaward transfer records, which is unique to IRIS UMETRICS data.

2. What kind of expenses does "total direct expenditure" include?

This includes salary expenditures, research training, equipment / service purchased from R&D project accounts (the payment amount from Vendor and Subaward files, respectively) paid by a given award. This, however, does not include indirect costs which are separately reported and recorded in the 'overhead charged' field in the Award file. Note that total and federally funded R&D expenditure data in the Institutional Fastfacts file (in the Auxiliary collection) were collected from NSF HERD survey and this dollar amount includes both direct and indirect costs. What makes our UMETRICS data unique is that direct and indirect spending are separately reported and made available to researchers at the institutional level.

3. What is the difference between the total amount of spending in the Award file and the award amount reported in NSF or NIH Award Detail files?

The difference between them is ***actual*** spending in UMETRICS (the total amount of spending as a sum of 'total direct expenditures' and 'overhead charged') and approved award (obligation) amount reported in NIH / NSF award data. What you see in our Award file are the details based on monthly spending. So, if a given NSF award duration runs between 2018 and 2021, you may see two years of spending on a 3-year award that is scheduled to run through 2021. This is why these two cannot easily be directly compared since lots can happen to spending in the course of a project that was not registered in obligations reported when the project was funded.

4. How can I tell which awards are federally funded?

In this release, we included an indicator field 'fed_award_flag' with a binary code (1 for federal and 0 for non-federal awards). The 'fed_funder_parent' field is also useful for researchers to know a federal funding department or parent agency as the origin of a given award. For non-federal awards, the field 'nonprofit_foundation_name' provides disambiguated names of nonprofit organizations as a funder.

5. Why are there university names in the field of 'funding_source_name' and what are they?

These universities are pass-through entities from which IRIS member universities receive subawards. The named university indicates a prime award recipient that provides a subaward to a subrecipient (other organizations including IRIS universities) to carry out part of a federal program. If researchers are interested in tracking pass-through funding that IRIS universities receive, funding source names are useful. Pass-through entity names could include not only academic institutions, but also state agencies, nonprofit organizations, etc.

Note that, as for pass-through funds, IRIS asks universities to directly map CFDA numbers to originating prime awards—for instance, if a pass-through funding is originated in NSF, the CFDA should be 47.### even though the funding source name indicates academic institutions.

IRIS has been working on cleaning and disambiguating funding source names. While the data are cleaner every release, the quality of disambiguation has still much room to be improved.

6. Why does the field 'cfda' include the five-digit number that does not look like official CFDA code, for instance, '00.000' or '00.600?

This is because IRIS asks universities to submit the contents in the 'unique award number' field in the same format for both federal and non-federal awards—either the CFDA code for federal awards or a STAR Other Funding Source (OFS) code (see Appendix C) for non-federal. As described in the data dictionary, a CFDA code indicates awarding federal agencies and is defined in the format of a 5-digit number (##.###) where the first 2-digits represent the funding agency and the 3-digits represent the federal domestic assistance program. If the funding source is a non-federal organization, universities follow the OFS codes like '00.600' indicating nonprofit foundations, with some exceptions like using '00.000'.

7. Are there any unique award numbers in the Award file that lack the uniqueness needed to accurately link records to other files?

Unfortunately, yes. There are cases in which the unique award number field has no particular award number assigned by funding organizations— examples include '00.000 agreement", '00.000 subrecipient', '00.000 addendum to agreement', etc. Many of them are non-federal awards. When no unique award identifier is given (thus no uniqueness of awards), this prevents researchers from finding corresponding unique award numbers between files. Filtering these cases may be recommended for more efficient and accurate data linkage. We provide a list of such cases in the public folder in VDE.

Employee

File Details

File Name: core_employee
Record Counts: 2,6376,022
Field/Column Counts: 15

File Summary

The Employee file provides information about the individuals working on awards at IRIS member universities. It connects directly with the Award and Object Code file, as well as, indirectly with the crosswalk generated by linkages at the employee-level. The file should contain a record for every employee that received any type of compensation from an award or spent time working on an award. While all individuals who charge time to federal or nonfederal grants are included in the data, the unit of record is a payment to an individual on an award in a pay-period. Thus, individuals routinely appear in multiple periods, on multiple awards. Although each employee should have only one entry per award/account number pairing per month, an employee could work on several awards concurrently or sequentially.

Data Fields

Fields appearing across files	Fields unique to this file
institution_id	job_title
emp_number	occupational_class
unique_award_number	umetrics_occupational_class
cfda	soc_code
period_start_date	fte_status
period_end_date	proportion_earnings_allocated
recipient_account_number	systematic_occupational_class
object_code	

Table 6: Core_Employee Data Fields

Field Name	Column Name	Data Type	Set Length	Max Length	Field Definition
CFDA	cfda	varchar	10	6	A five-digit Catalog of Federal Domestic Assistance (CFDA) number retrieved from the unique_award_number that universities submit. This number is assigned to awards to represent the source of funding. If the first two digits are between 10 and 98, awards are federally funded. If the first two digits range between 00 and 09, or 99, sources of awards are most likely non-federal
Employee Number	emp_number	varchar	200	32	IRIS-generated unique identifier assigned to all personnel being paid by awards
FTE Status	fte_status	numeric	9	5	Designation of the status of the funded personnel (e.g., full time = 1.0, half time = .5); FTE is a university specific, not an award specific field; the value ranges between 0 and 1
Systematic Occupational Class	systematic_occupational_class	varchar	50	22	Systematic occupational classification generated by IRIS in 2019; this includes six groups, Undergraduate, Graduate Student, Other Student, Post Graduate Researcher, Faculty, and Staff; this new classification is a result of analyzing a combination of university supplied job titles, occupational classes, CDFA numbers, and object code descriptions in order to provide a more accurate picture of student involvement in sponsored research
Institution ID	institution_id	int	4	4	IRIS-generated unique identifier assigned to each IRIS member university for

					de-identification purposes. Values are four or five digit numbers
Job Title	job_title	varchar	200	142	Job or occupation title assigned to the funded personnel by IRIS member universities
Object Code	object_code	varchar	50	39	Internal object code or other expense type category assigned to a transaction to identify payment purposes or resources
Occupational Class	occupational_class	varchar	50	50	Job classification provided by IRIS member universities
Period End Date	period_end_date	date	3	3	End of period in which a monthly expense transaction took place; each period end day is the last day of a month: e.g., 3/30/2008 or 12/31/2014
Period Start Date	period_start_date	date	3	3	Beginning of period in which a monthly expense transaction took place; each period start day is the first day of a month: e.g., 4/1/2009 or 10/1/2015
Proportion of Earnings	proportion_earnings_allocated	numeric	9	9	Calculated portion of earnings charged by funded personnel to the award in the specified period. This is not actual salary or dollar amounts, the value ranges between 0 and 1 depending on how much of the salary is derived from an award
Recipient Account Number	recipient_account_number	varchar	50	15	A university's internal account number to uniquely identify each project; typically an accounting code used to allocate funds received from an award
SOC Code	soc_code	varchar	50	30	Standard Occupational Classification codes that are required for federal agency reporting (http://www.bls.gov/soc/); each occupation in the SOC is placed within one of 23 major groups

UMETRICS Occupational Class	umetrics_occupational_class	varchar	50	22	Job classification generated by IRIS; jobs are categorized into 6 major aggregate groups (Faculty, Staff, Post Graduate Research, Graduate Student, Undergraduate, and Other). The Staff group is further classified into 6 categories (Clinical, Research, Research Facilitation, Technical Support, Instructional, Other Staff)
Unique Award Number	unique_award_number	varchar	500	51	University-generated unique identifier specifying an award and its funding source, made up of the 5-digit funding source code (e.g., CFDA number) and an award identifier. Award identifier may include the awarding agency's federal award ID (e.g., federal grant number, contract number, or loan number) or an internal award ID for non-federal awards. Values may include a space or dash in between them: e.g., “10.310 2010-12345-54321” (USDA example), “47.050 1234567” (NSF example), “93.865 2-R01-DK-012345-15-S1” (NIH example), “00.000 1234567” and “00.200 State Award 1” (Non-federal grant examples)

Q&A re. Employee File

1. **There are three different but similar occupational classification fields for employees. What is the difference among them and which one should I use for what purposes?**

   If you have used IRIS UMETRICS data before, you should be familiar with the two fields ('occupational_class' and 'umetrics_occupational_class') that remain the same from previous releases—the former is university-submitted occupational data and the latter is based on manual classification built on processes from STARMETRICS and coded by Dr. Bruce Weinberg (an IRIS Co-PI)'s team at OSU.

   The existing UMETRICS occupational classification coding aims to categorize the production function of an employee's work at the job title level and include 12 categories. The new classification ('systematic_occulational_class') is built on a set of variables (Occupational Classification, CFDA, and Object Code Description) towards improving the way we had grouped students before, which include only 6 broad categories.

   The new field is built to complement the UMETRICS coding, not replace the existing process; however, we recommend using the new classification if employees' student status is central to one's analysis. For more details, see the method section in this documentation.

2. **As a data user who was given access to the previous release, I am interested in using the previous linkage crosswalk that includes employee numbers and applying it to the new Employee file from this release. Is it possible to track the same individuals through their de-identified Employee Numbers?**

   IRIS continues to randomly generate employee numbers as a unique ID, using the HashBytes function built in the SQL database. Although the method is similar, previously a unique number was generated by reading particular data elements (including employee name and YOB/MOB). These data elements used for hashing have been changed and the same individuals are not assigned the same employee number in this current release. Until these linkages are updated in the summer / fall of 2020 with current employee numbers, should you be interested in using the linkage work at the individual level such as ProQuest-UMETRICS or Patent-UMETRICS crosswalk from previous releases, please reach out to the IRIS Research Support team for a crosswalk that links current IDs to those from the past release.

3. **Can you explain how I can use the information on Full Time Equivalent (FTE) Status and Proportion of Earnings?**

   This describes whether an employee is a full time or part time employee of the institution. Note that the value in the FTE status field is not specific to a given award; instead this is specific to an employee. The FTE status is the employee's status at the institution and it is the designation of the status (percent) of the funded personnel (full time = 1.0, half time = 0.5) during the given reporting period.

The value in the field of fte_status should be larger than 0 and equal to or smaller than 1 (indicating an employee working 40+ hours per week is considered full time with an FTE of 1, 100% Full Time Equivalent). However, some universities do not consistently report employees' percentage of FTEs. They often fill in FTE with a zero or an empty field.

The Proportion of Earnings is the calculated portion of earnings charged by funded personnel to the award in the specified period. This field is a simple calculation:

The 'Total Earnings for Employee' is the total gross amount that an employee earned in a given transactional period. The 'Earnings from Specified Award' are those earnings that came exclusively from work or association with the given award for the given employee in the given time period. Note that an employee can work on more than one award, so it is possible that an employee could have multiple entries in the Employee file during a given time period. In that case, each one would be for a specific award/account number, and have its own separate Proportion of Earnings Allocated.

Vendor

File Details

File Name: core_vendor
Record Counts: 2,3982,459
Field/Column Counts: 20

File Summary

The Vendor file includes transaction information on all vendor purchases (both goods and services) made on the awards contained in the Award file. It connects directly with the Award and Object Code files. Vendor transactions are rolled up to the month at the vendor establishment level. In other words, the vendor payment amount field contains the sum of all expenditures with the given vendor establishment in the given month off of the given unique award number / recipient account number. Vendor establishment in this instance refers to a single address of a given vendor. The file does not include spending to collaborating institutions through subawards or subcontracts, as this type of expense is included in the Subaward file.

Data Fields

Fields appearing across files	Fields unique to this file
institution_id	vendor_id
unique_award_number	vendor_name_clean
cfda	vendor_name_raw
period_start_date	vendor_ein
period_end_date	vendor_duns
recipient_account_number	vendor_payment_amt
object_code	vendor_address
person_org_flag	vendor_city
	vendor_state
	vendor_domestic_zipcode
	vendor_foreign_zipcode
	vendor_country

Table 7: Core_Vendor Data Fields

Field Name	Column Name	Data Type	Set Length	Max Length	Field Definition
CFDA	cfda	varchar	10	6	A five-digit Catalog of Federal Domestic Assistance (CFDA) number retrieved from the unique_award_number that universities submit. This number is assigned to awards to represent the source of funding. If the first two digits are between 10 and 98, awards are federally funded. If the first two digits range between 00 and 09, or 99, sources of awards are most likely non-federal
Institution ID	institution_id	int	4	4	IRIS-generated unique identifier assigned to each IRIS member university for de-identification purposes. Values are four or five digit numbers
Object Code	object_code	cfda	50	6	Internal object code or other expense type category assigned to a transaction to identify payment purposes or resources
Period End Date	period_end_date	date	3	3	End of period in which a monthly expense transaction took place; each period end day is the last day of a month: e.g., 3/30/2008 or 12/31/2014
Period Start Date	period_start_date	date	3	3	Beginning of period in which a monthly expense transaction took place; each period start day is the first day of a month: e.g., 4/1/2009 or 10/1/2015
Person Organization Flag	person_org_flag	varchar	1	1	A binary code ('P' for person or 'O' for organization) to differentiate type of vendors. This dichotomous category was utilized to mask vendors' personally identifiable information
Recipient Account Number	recipient_account_number	varchar	50	16	A university's internal account number to uniquely identify each project; typically an accounting code used to allocate funds received from an award

Unique Award Number	unique_award_number	varchar	500	51	University-generated unique identifier specifying an award and its funding source, made up of the 5-digit funding source code (e.g., CFDA number) and an award identifier. Award identifier may include the awarding agency's federal award ID (e.g., federal grant number, contract number, or loan number) or an internal award ID for non-federal awards. Values may include a space or dash in between them: e.g., “10.310 2010-12345-54321” (USDA example), “47.050 1234567” (NSF example), “93.865 2-R01-DK-012345-15-S1” (NIH example), “00.000 1234567” and “00.200 State Award 1” (Non-federal grant examples)
Vendor Address	vendor_address	varchar	200	148	Address of the vendor. IRIS has replaced vendor address with the string of 'masked' if vendor addresses are provided by universities (thus not null) and vendors are individuals
Vendor City	vendor_city	varchar	50	37	City of the vendor associated with the vendor address
Vendor Country	vendor_country	varchar	50	16	Country of the vendor associated with the vendor address
Vendor Domestic Zipcode	vendor_domestic_zipcode	varchar	50	15	US ZIP code of vendor associated with the vendor address
Vendor DUNS	vendor_duns	varchar	50	15	A vendor's nine-digit the Data Universal Numbering System (DUNS) number to identify business entities on a location-specific basis is copyrighted and provided by Dun & Bradstreet (D&B). If a DUNS number was provided (i.e., not null) by universities for vendors that are individuals, IRIS has replaced DUNS number with hashed DUNS number for de-identification purposes

Vendor EIN	vendor_ein	varchar	50	30	A vendor's nine-digit Employer Identification Number (EIN). If an EIN was provided (i.e., not null) by universities for vendors that are individuals, IRIS has replaced EIN with hashed EIN for de-identification purposes
Vendor Foreign Zip Code	vendor_foreign_zipcode	varchar	50	17	Foreign ZIP/postal code of vendor associated with the vendor address
Vendor ID	vendor_id	varchar	200	32	IRIS-generated unique identifier assigned to the vendor (an organization or individual) that provides goods or services paid by an IRIS member university's award. IRIS cleans vendor name records from the data submitted by universities and generates this identifier based on the cleaned names
Vendor Name	vendor_name	varchar	500	170	Cleaned name of the vendor. IRIS has replaced vendor names with the string of 'masked' if names are provided by universities (thus not null) and vendors are individuals
Vendor Name Raw	vendor_name_raw	varchar	500	142	Raw name of the vendor. IRIS has replaced vendor names with the string of 'masked' if names are provided by universities (thus not null) and vendors are individuals
Vendor Payment Amount	vendor_payment_amt	numeric	9	9	Funds charged to the award by the vendor in the specified period; vendor payment amounts can be negative
Vendor State	vendor_state	varchar	50	23	State of the vendor associated with the vendor address

Subaward

File Details

File Name: core_subaward
Record Counts: 785,522
Field/Column Counts: 20

File Summary

The Subaward file includes transaction information on all subawards or subcontracts made off the awards contained in the Award file, allowing to track funding transferred to subawardees (also commonly referred to as subrecipients) off of the awards reported in the Award file—IRIS member universities serve as a 'path-through entity' if a given prime award was spent for subawards sent to other institutions or organizations. The file connects directly to the Award and Object Code files. University expenditure transactions of subawards are rolled up to the month at a subawardee (subrecipient) level. In other words, the file contains the sum of all expenditures with the given subawardee in the given month off of the given unique award number / recipient account number.

Data Fields

Fields appearing across files	Fields unique to this file
institution_id	subaward_id
period_start_date	subaward_ein
period_end_date	subaward_duns
unique_award_number	subaward_payment_amt
cfda	subaward_name_clean
recipient_account_number	subaward_name_raw
object_code	subaward_address
person_org_flag	subaward_city
	subaward_state
	subaward_domestic_zipcode
	subaward_foreign_zipcode
	subaward_country

Table 8: Core_Subaward Data Fields

Field Name	Column Name	Data Type	Set Length	Max Length	Field Definition
CFDA	cfda	varchar	10	6	A five-digit Catalog of Federal Domestic Assistance (CFDA) number retrieved from the unique_award_number that universities submit. This number is assigned to awards to represent the source of funding. If the first two digits are between 10 and 98, awards are federally funded. If the first two digits range between 00 and 09, or 99, sources of awards are most likely non-federal
Institution ID	institution_id	int	4	4	IRIS-generated unique identifier assigned to each IRIS member university for de-identification purposes. Values are four or five digit numbers
Object Code	object_code	varchar	50	6	Internal object code or other expense type category assigned to a transaction to identify payment purposes or resources
Period End Date	period_end_date	date	3	3	End of period in which a monthly expense transaction took place; each period end day is the last day of a month: e.g., 3/30/2008 or 12/31/2014
Period Start Date	period_start_date	date	3	3	Beginning of period in which a monthly expense transaction took place; each period start day is the first day of a month: e.g., 4/1/2009 or 10/1/2015
Person Organization Flag	person_org_flag	varchar	1	1	A binary code ('P' for person or 'O' for organization) to differentiate type of subawardees. This dichotomous category was utilized to mask vendors' personally identifiable information
Recipient Account Number	recipient_account_number	varchar	50	15	A university's internal account number to uniquely identify each project; typically an accounting code used to

					allocate funds received from an award
Subaward Address	subaward_address	varchar	200	121	Address of the subawardee. IRIS has replaced subawardee address with the string of 'masked' if subawardee address are provided by universities (thus not null) and subawardees are individuals
Subaward City	subaward_city	varchar	50	33	City of the subawardee associated with the subawardee address
Subaward Country	subaward_country	varchar	50	16	Country of the subawardee associated with the subawardee address
Subaward Domestic Zipcode	subaward_domestic_zipcode	varchar	50	12	US ZIP code of subawardee associated with the subawardee address
Subaward DUNS	subaward_duns	varchar	50	13	A subawardee's nine-digit the Data Universal Numbering System (DUNS) number to identify business entities on a location-specific basis is copyrighted and provided by Dun & Bradstreet (D&B). If a DUNS number was provided (i.e., not null) by universities for subawardees that are individuals, IRIS has replaced DUNS number with hashed DUNS number for de-identification purposes
Subaward EIN	subaward_ein	varchar	50	30	A subawardee's nine-digit Employer Identification Number (EIN). If an EIN was provided (i.e., not null) by universities for subawardees that are individuals, IRIS has replaced EIN with the string 'masked'
Subaward Foreign Zip Code	subaward_foreign_zipcode	varchar	50	19	Foreign ZIP/postal code of subawardee associated with the subawardee address
Subaward ID	subaward_id	varchar	200	32	IRIS-generated unique identifier assigned to the subaward recipient organization to which an IRIS member university provides program awards / subgrants /

					subcontracts. IRIS cleans subawardee name records from the data submitted by universities and generates this identifier based on the cleaned names
Subaward Name Clean	subaward_name_clean	varchar	500	199	Cleaned name of the subawardee. IRIS has replaced subawardee names with the string of 'masked' if names are provided by universities (thus not null) and subawardees are individuals
Subaward Name Raw	subaward_name_raw	varchar	500	200	Raw name of the subawardee. IRIS has replaced subawardee names with the string of 'masked' if names are provided by universities (thus not null) and subawardees are individuals
Subaward Payment Amount	subaward_payment_amt	numeric	9	9	Funds charged to the award by the subawardee in the specified period; subaward payment amounts can be negative
Subaward State	subaward_state	varchar	50	11	State of the subawardee associated with the subawardee address
Unique Award Number	unique_award_number	varchar	500	66	University-generated unique identifier specifying an award and its funding source, made up of the 5-digit funding source code (e.g., CFDA number) and an award identifier. Award identifier may include the awarding agency's federal award ID (e.g., federal grant number, contract number, or loan number) or an internal award ID for non-federal awards. Values may include a space or dash in between them: e.g., “10.310 2010-12345-54321” (USDA example), “47.050 1234567” (NSF example), “93.865 2-R01-DK-012345-15-S1” (NIH example), “00.000 1234567” and “00.200 State Award 1” (Non-federal grant examples)

Institution Fastfacts

File Details

File Name: aux_institution_fastfacts
Record Counts: 594
Field/Column Counts: 16

File Summary

This file contains information on 33 universities, characterizing each institution by its R&D expenditures, student enrollment, number of PIs, and research personnel, etc. Each university in this file is associated with the de-identified university ID (institution_id) so that the file directly connects to other files. Data sources include the NSF Higher Education R&D Survey (HERD), NSF-NIH Survey of Graduate Students & Post-doctorates in Science and Engineering (GSS), the NSF Survey of Earned Doctorates (SED), and the Integrated Postsecondary Education Data System (IPEDS) Fall Enrollment Survey. The data coverage in this file (2000-2018) aligns with the one from the core collection (Award, Employee, Vendor, and Subaward) files although the FY 2019 data are not available from NSF HERD at the time of this release. Institutional Fastfacts File

Data Fields

Fields appearing across files	Fields unique to this file
institution_id	year
	institution_control
	carnegie_classification
	carnegie_code
	land_grant
	med_school
	main_med_distance
	total_rd_expenditures
	fed_rd_expenditures
	number_doc_recipients
	fall_enrollment
	number_grad_students
	number_pis
	number_post_docs
	number_other_personnel

Table 9: Aux_Institution_Fastfacts Data Fields

Field Name	Column Name	Data Type	Set Length	Max Length	Field Definition
Carnegie Classification	carnegie_classification	varchar	50	50	Derived from the 2018 Classification Update of the traditional Carnegie Classification Framework for each academic institution; data retrieved Carnegie Classification of Institutions of Higher Education website: http://carnegieclassifications.iu.edu/
Carnegie Code	carnegie_code	int	4	4	A numerical code for the Carnegie Classification. Based on the 2018 Carnegie Classification Update of the traditional Carnegie Classification Framework for each academic institution; Description of codes are available on the Carnegie Classification of Institutions of Higher Education website: http://carnegieclassifications.iu.edu/
Fall Enrollment	fall_enrollment	int	4	4	The number of students enrolled in courses that are creditable toward a degree, diploma, certificate, or other formal award, or are part of a vocational or occupational program including any students enrolled in off-campus centers; data retrieved from the Integrated Postsecondary Education Data System (IPEDS) Fall Enrollment Survey
Federally Financed R&D Expenditures in All Fields	fed_rd_expenditures	int	4	4	R&D expenditures in all fields, including direct and recovered indirect costs, funded by all agencies of the Federal government; data retrieved from NSF HERD
Institution Control	institution_control	int	4	4	Defined for academic institutions as private or public (not applicable to biomedical institutions); values include 1 (Private) and 0 (Public); data retrieved from NSF HERD
Institution ID	institution_id	int	4	4	IRIS-generated unique identifier assigned to each IRIS member university for de-identification

					purposes. Values are four or five digit numbers
Land Grant	land_grant	numeric	9	5	Indicator for whether an institution is a Land Grant institution; values include 1 (institution is a Land Grant institution) and 0 (not a Land Grant institution); data retrieved from IPEDS and verified on https://nifa.usda.gov/land-grant-colleges-and-universities-partner-website-directory
Distance between medical school and main campus	main_med_distance	int	4	4	The geographical distance (in miles) of the medical school from the main campus if an IRIS member university has a medical school
Medical School	med_school	int	4	4	Indicator for each institution having a medical school included as part of its reporting unit; Values include 1 (has medical school) and 0 (does not include medical school); data retrieved from NSF HERD
Number of Doctorate Recipients	number_doc_recipients	int	4	4	All earned doctorates granted by universities; data retrieved from the NSF Survey of Earned Doctorates (SED)
Number of Graduate Students	number_grad_students	int	4	4	The number of graduate students enrolled in GSS-eligible science, engineering, and health (SEH) units in the fall of the data collection year; data retrieved from the NSF-NIH Survey of Graduate Students & Postdoctorates in Science and Engineering (GSS)
Number of Other Personnel	number_other_personnel	int	4	4	All other personnel paid from the R&D salaries, wages and fringe benefits reported on the NSF Higher Education Research and Development Survey (HERD) who are not categorized as principal investigators; data retrieved from NSF HERD
Number of Principal Investigators	number_pis	int	4	4	Personnel paid from the R&D salaries, wages and fringe benefits reported on the survey (NSF Research and Development Expenditures at Universities and Colleges/Higher Education

					Research and Development Survey), and designated by the institution to direct the R&D project or program and be responsible for the scientific and technical direction of the project; Co-investigators (co-PIs) may be designated for this role and are also included. Missing data for this question were not imputed, therefore aggregate totals represent an undercount; data retrieved from NSF HERD
Number of Postdocs	number_post_docs	int	4	4	Personnel defined as postdocs, namely recent doctorate recipients with limited-term appointments primarily intended for training under the supervision of a senior scholar. Data retrieved from the NSF-NIH Survey of Graduate Students & Postdoctorates in Science and Engineering (GSS)
Total R&D Expenditures in All Fields	total_rd_expenditures	int	4	4	R&D expenditures from the institution's current operating funds that were separately accounted for, including expenditures for organized research as defined by 2 CFR 220 Part 200 Appendix III and expenditures from funds designated for research. Expenditures came from internal or external funding and included recovered and unrecovered indirect costs. Funds passed through to subrecipient organizations were also included. R&D was excluded if it was conducted by university faculty or staff at outside institutions and was not accounted for in the reporting institution's financial records. Data retrieved from NSF HERD
Year	year	int	4	4	Year (ranging between 2010 and 2017) is defined in two ways: 1) Academic year: doctorate recipients, fall enrollment, and other personnel-related data; 2) Fiscal year: R&D expenditures as

					defined in the data source, the NSF Higher Education R&D Survey (NSF HERD)

Comprehensive Award List

File Details

File Name: aux_comprehensive_award_list
Record Counts: 443,074
Field/Column Counts: 8

File Summary

This file contains all awards that appear in the Award, Employee, Vendor, and Subaward Files. Technically, all awards that are spent and have associated transactional records in the Employee, Vendor, or Subaward file should be in the Award file, but some awards are missing from the Award file. This file helps researchers know which award originates in what transaction file, thereby identifying either an existing or missing connection between files for each award. If a binary value indicates 0 in the Award field for a given award but 1 in the Vendor file, for this particular award the connection between Award and Vendor files is missing. Therefore, the number of unquietly counted awards using this file is about 10% larger than from the Award file.

Data Fields

Fields appearing across files	Fields unique to this file
institution_id	award
unique_award_number	employee
cfda	vendor
fed_award_flag	subaward

Table 10: Aux_Comprehensive_Awards_List Data Fields

Field Name	Column Name	Data Type	Set Length	Max Length	Field Definition
Award File	award	int	4	4	A binary code to differentiate the file from which a given award originates; coded 1 if the award is present in Award file; coded 0 otherwise
CFDA	cfda	varchar	10	6	A five-digit Catalog of Federal Domestic Assistance (CFDA) number retrieved from the unique_award_number that universities submit. This number is assigned to awards to represent the source of funding. If the first two digits are between 10 and 98, awards are federally funded. If the first two digits range between 00 and 09, or 99, sources of awards are most likely non-federal
Employee File	employee	int	4	4	A binary code to differentiate the file from which a given award originates; coded 1 if the award is present in Employee file; coded 0 otherwise
Federal award flag	fed_award_flag	int	4	4	A binary code to identify federally funded awards; coded 1 if award is federal, 0 otherwise. This flag for each unique award number corresponds to the value in the 'fed_award_flag' field in the Award file if the unique award number appears in the Award file, in other words 0 in the Award field in this file; if a given unique award number does not appear in the Award file (thus 0 in the Award field), then this indicator is based on CFDA
Institution ID	institution_id	int	4	4	IRIS-generated unique identifier assigned to each IRIS member university for de-identification purposes. Values are four or five digit numbers
Subaward File	subaward	int	4	4	A binary code to differentiate the file from which a given award originates; coded 1 if the award is present in Subaward file; coded 0 otherwise
Unique Award Number	unique_award_number	varchar	500	100	University-generated unique identifier specifying an award and its funding source, made up of the 5-digit funding source code (e.g., CFDA number) and an award identifier.

					Award identifier may include the awarding agency's federal award ID (e.g., federal grant number, contract number, or loan number) or an internal award ID for non-federal awards. Values may include a space or dash in between them: e.g., "10.310 2010-12345-54321" (USDA example), "47.050 1234567" (NSF example), "93.865 2-R01-DK-012345-15-S1" (NIH example), "00.000 1234567" and "00.200 State Award 1" (Non-federal grant examples)
Vendor File	vendor	int	4	4	A binary code to differentiate the file from which a given award originates; coded 1 if the award is present in Vendor file; coded 0 otherwise

Q&A re. Comprehensive Award List File

1. The 'unique_award_numbers' in the Employee file could not be found in the Award file. Why is that?

It is because some award information was already missing from the Award file before the data were submitted by universities to IRIS. We have been monitoring and reporting the issue, i.e., the level of file completeness or file mergeness. This Comprehensive Award List file will help researchers know which given unique award number originates in which core files. For more details on the level of file mergeness, see the Data Summary section.

In addition to the aforementioned issue that some unique award numbers are missing in the award information from the Award file, researchers will also encounter some cases in which supposedly the same unique award numbers are recorded slightly differently across files. This prevents researchers from linking files (joining SQL relational tables) on the field of 'unique award number.' Although we cleaned as much as possible to minimize such problems, researchers may want to watch out for such cases, especially from a handful of universities whose file mergeness level is low. One solution to identify and link these two unique award numbers. If awards are funded by NIH or NSF, these cases are nicely linked through matched and verified NSF award ID or NIH core project numbers. For this purpose, the award crosswalk is helpful.

CFDA Lookup

File Details

File Name: aux_cfda
Record Counts: 6,359
Field/Column Counts: 11

File Summary

The CFDA file provides a historical listing of the US government domestic assistance programs (1960-) which helps to characterize federal awards that IRIS universities received through the information on program title, sponsoring agencies (both at the highest and sub-unit level if relevant), types of domestic assistance, etc. The CFDA number is one the linking assets across most of the release files at the award level although this is limited to federal awards. Note that this file lists a lot more CFDA numbers than what we find in the core files— the umetrics_flag field in this file is of use for identifying which CFDA is directly associated with federal wards in core files.

Data Fields

Fields appearing across files	Fields unique to this file
cfda	first_two
	type_of_assistance
	gov_unit_code
	gov_unit_name
	program_title
	sub_unit_code
	sub_unit_name
	year_established
	year_modified_or_archived
	umetrics_flag

Table 11: Aux_CFDA Data Fields

Field Name	Column Name	Data Type	Set Length	Max Length	Field Definition
CFDA	cfda	varchar	10	7	A Catalog of Federal Domestic Assistance (CFDA) number used to identify and sort federal financial assistance programs. Each CFDA number contains five digits and appears in the following format: ##.### (e.g., 10.001 or 93.301)
First Two Digits of CFDA	first_two	int	4	4	The first two digits of CFDA numbers that appear in the CFDA field described above. The first two digit number can be used to map each federal financial assistance program to funding agencies. The first two digit numbers (in integer format) range between 10 and 98
Government Unit Code	gov_unit_code	varchar	50	14	The official acronym of the assisting federal agency at the highest level of a given domestic assistance program. For example, if a domestic assistance program is administered and managed by the National Institutes of Health, this field indicates its parent organization acronym, HHS
Government Unit Name	gov_unit_name	varchar	200	62	The name of the assisting federal agency at the highest level of a given domestic assistance program. For example, if a domestic assistance program is administered and managed by the National Institutes of Health, this field indicates its parent organization, the Department of Health and Health Sciences
Program Title	program_title	varchar	500	239	The program title that is available and downloaded from the website, cfda.gov (sam.gov). This field is null for CFDAs that are not found in the historical file.

Sub-unit Code	sub_unit_code	varchar	50	21	The official acronym of the federal agency that directly provides and administers a given domestic assistance program. For example, if a domestic assistance program is provided and administered by the National Institutes of Health, this field indicates NIH
Sub-unit Name	sub_unit_name	varchar	200	86	The name of the federal agency that directly provides and administers a given domestic assistance program. For example, if a domestic assistance program is provided and administered by the National Institutes of Health, this field indicates National Institutes of Health.
Type of Assistance	type_of_assistance	varchar	500	500	Each program is identified in terms of one or more of the 15 types of assistance provided. This field has values only for active CFDAs
UMETRICS File	umetrics_flag	int	4	4	This field is a binary value to differentiate the file from which a given CFDA number originates; valued 1 if the CFDA number is present in umetrics file; valued 0 otherwise. The UMETRICS file was created by IRIS staff and includes 1879 unique five-digit CFDA numbers that were retrieved from core award, employee, vendor, and subaward files in December 2018
Program Established Year	year_established	int	4	4	The year when a given domestic assistance program (associated with a unique CFDA number) was established. The field is null if CFDA numbers are not found in the historical file.
Program Modified or Archived Year	year_modified_or_archived	int	4	4	The year when a given domestic assistance program was modified or archived

Q&A re. CFDA Lookup File

1. How do you suggest using this CFDA lookup file?

This file helps researchers identify the source of federal funding through CFDA numbers that are mapped to funding agency names at parent- and sub-agency levels. The 5-digit CFDA numbers in this file is a linking element to other release files. The first 2-digits of CFDA numbers (the value of the 'gov_unit_code'), e.g. 47, can be used to map each federal finance assistance program to funding departments / agencies (in the field of 'gov_unit_name'). Types of assistance (in 15 categories) are particularly useful when differentiating research grants from non-research ones.

Note that this file includes a lot more domestic assistance programs and associated information than what we find in the 'cfda' fields from the Award and other core files. We recommend using the indicator field 'umetrics_flag' that helps you identify which CFDA numbers are relevant to awards found in UMETRICS files.

2. Why are data missing in all fields except for the CFDA field in some cases, for instance, 10.000, 12,000, 47.000, 93.000, etc.?

This is because these 5-digit numbers are not official CFDA numbers. Since we created this CFDA lookup file based on the information available from the website (https://beta.sam.gov/help/assistance-listing), you find no detail if the 5-digit numbers were not considered official CFDA numbers (found in neither historical not active CFDA data files). We included these cases for the indicator flag ('umetrics_flag') even though we know such 5-digit numbers in the CFDA field were erroneously recorded on the university end. In particular, some universities intentionally use the correct first 2-digit numbers that maps to federal funding agencies (e.g., 10 for Department of Agriculture), but the last 3-digit numbers are simply replaced with '000' (e.g. 10.000) when the program-level number and information was either unavailable or unknown.

Suborganization

File Details

File Name: aux_suborganization
Record Counts: 2,569
Field/Column Counts: 6

File Summary

This file includes a list of sub-organization unit IDs and their names (if a university provides names of sub-organization units). This lookup file helps to map sub-organization unit IDs that appear in the Award file, and in particular, enables a cross-sectional analysis through the use of 18 standardized sub-organization units including, arts & sciences, education, health sciences, administration, etc.

Data Fields

Fields appearing across files	Fields unique to this file
institution_id	main_campus
campus_id	sub_org_unit_name
sub_org_unit	sub_org_unit_std

Table 12: Aux_Suborganization Data Fields

Field Name	Column Name	Data Type	Set Length	Max Length	Field Definition
Campus ID	campus_id	varchar	50	8	IRIS-generated identifier assigned to each campus for de-identification purposes. Each ID is a combination of institution_id and a serial number helpful to identify the campus to which a given award is made and by which the award is being administered
Institution ID	institution_id	int	4	4	IRIS-generated unique identifier assigned to each IRIS member university for de-identification purposes. Values are four or five digit numbers
Main Campus Flag	main_campus	int	4	4	A binary value to indicate the main campus of each member university; valued 1 if a given campus is the main campus (flagship campus); valued 0 otherwise; if a given university has one campus in UMETRICS data, then that is the main campus
Sub-organization Unit	sub_org_unit	varchar	100	12	IRIS-generated identifier assigned to sub-organization units to which each funded project is assigned, such as a particular college within a given IRIS member university. (This is not at the level of individual departments.) Each ID is a combination of campus_id described above and a serial number assigned to each sub-organization unit within each campus. This ID helps to identify the college or unit to which a given award is made and by which the award is being administered at a lower-level
Sub-organization Unit Name	sub_org_unit_name	varchar	200	100	Sub-organizational unit name that maps to sub-organizational unit code, e.g., the college of natural sciences, the medical school, or the college of engineering. If the sub-org unit name provided by IRIS member universities contains identifiable information, IRIS replaced all information in this field with the string 'masked'
Standardized Sub-organization Unit Name	sub_org_unit_std	varchar	100	26	Standardized sub-organization unit name with 18 categories (e.g., arts & sciences, health sciences, administration, etc.); this classification

					is a result of manual mapping and grouping based on the existing sub-organization unit information provided by universities as well as additional research done by IRIS through web resources that describe field and/or discipline of a given university campus organization unit

Q&A re. Suborganization File

1. There are two different but similar fields indicating sub-organization unit names in this file. How do you suggestusing these fields in this Suborganization file in general?

Suborganization unit names can be used to map to the Award file, which includes Campus ID ('campus_id') and a sub-organization unit code ('sub_org_unit') without any identifiable information. With this lookup file, researchers can identify names of campus units where a given award is administered on each campus to potentially draw connections between research activities and collaboration networks.

For de-identification purposes IRIS has to mask some descriptions of sub-organization unit names if names happen to include the name of campus buildings or anything that is very specific to universities. To complement the information lost from masking as well as to create more general unit names that are standardized across universities for cross-sectional analysis, this year's release includes a new field called 'sub_org_unit_std." As noted in the release highlights earlier, we have consolidated various university-specific campus units into 18 generic categories which we hope enables a cross-sectional analysis with a focus on units through which grants are administered and research activities are performed. These 18 categories are:

- Administration
- Agriculture
- Arts and Sciences
- Business
- Communication and Journalism
- Earth Sciences
- Education
- Engineering
- Fine Arts
- Graduate School
- Health and Human Services
- Health Sciences
- Law
- Library and Information
- Medicine
- Public Policy
- Research Institute
- Veterinary Medicine

Note that some universities (Institution IDs 10151, 27201, 45264, and 62799) do not provide details of their sub-organization units, and the Institution ID 11479 is missing about 50% of sub-organization unit names. For these universities, unfortunately, standardized names are also missing.

Employee Demographics

File Details

File Name: aux_emp_demographics
Record Counts: 548,483
Field/Column Counts: 6

File Summary

This file provides demographic characteristics of employees who are paid by research grants. Not only does this file helps researchers identify the range in which an individual's age falls, but it also provides gender and ethnicity. Both gender and ethnicity information is the result of imputation work done by IRIS—employee names provided by IRIS universities are matched to a source file including over a million names. Note that employee names and year of birth (YOB) information from active member universities was used to generate this file, but neither name nor YOB is released for privacy and confidentiality protection.

Data Fields

Fields appearing across files	Fields unique to this file
institution_id	imputed_ethnicity
emp_number	imputed_gender
	yob_category
	yob_range

Table 13: Aux_Emp_Demographics Data Fields

Field Name	Column Name	Data Type	Set Length	Max Length	Field Definition
Employee Number	emp_number	varchar	200	32	IRIS-generated unique identifier assigned to all personnel being paid by awards
Imputed Ethnicity	imputed_ethnicity	varchar	200	19	Imputed ethnicity; in addition to 28 broad ethnicity groups (e.g., French, Chinese, English, Thai, etc.), this field has many hyphenated categories to represent multiple ethnicities (e.g., Arab-Hispanic, English-German); this field is null if an employee name was neither found nor matched to the source data that IRIS used to impute ethnicity
Imputed Gender	imputed_gender	varchar	200	1	Imputed gender (M for male and F for female); this field is null if an employee name was neither found nor matched to the source data that IRIS used to impute gender
Institution ID	institution_id	int	4	4	IRIS-generated unique identifier assigned to each IRIS member university for de-identification purposes. Values are four or five digit numbers
Year of Birth Category	yob_category	int	4	4	To simplify age bands in the field of "yob_range" above, we converted age bands into integers between 1 and 11. If yob_range value is "na" it is coded as 99, and if "masked" it is coded as 98
Year of Birth Range	yob_range	varchar	50	21	The range in which an employee's year of birth falls, e.g., if one's birth year falls between 1988 and 1992, it is coded as "between_1988_and_1992". Employee age is mapped to unique employee ID found in the Employee File, and due to disclosure risk, 11 age bands are used to aggregate years. Additional values include "na" and "masked." If no information on an employee's birth year was provided by universities, it is coded as "na". Some age bands are rolled up into one and coded as "masked" due to disclosure risk

Q&A re. Employee Demographics File

1. How do you suggest using the Employee Demographics file?

All demographic variables could be of use for team- or individual-level analysis and here are some tips when using demographic fields.

As to age, we provide employees' age in 5-year ranges, for instance, a given employee's age falls within the range of 'born between 1963-1967'. This is for de-identification purposes—although IRIS receives personally identifiable information (PII) data, these are not available to share with researchers per agreements with IRIS universities. When using the age variable, it is important to keep in mind that every employee's age would vary by the time of transaction (payment) in our data. Since the temporal coverage (2001-2019) varies by universities, the age of a graduate student at the time of 2001 transaction data is different from the time of payment two years later in 2003. Using the year or actual date of transaction ('pay_period_start/end_date') information would be essential for more accuracy. However, one caveat is that if the comparison is made between two dates that are less than 5 years apart , the age band in which an employee falls will remain the same.

In this release, in addition to age, researchers will find gender and ethnicity data as well. These two fields are ***not*** the data we received from universities but instead as a result of imputation. Since our imputation work heavily depends on the source file (names that we match for retrieval of ethnicity and gender information), note that about 13% of employees are not assigned imputed gender and about 9% are missing imputed ethnicity.

Object Code

File Details

File Name: aux_object_code
Record Counts: 6,373
Field/Column Counts: 3

File Summary

This file includes a list of different object codes assigned to all transactions. Each transaction that appears in Employee, Vendor, and Subaward files is assigned into a different object code (classification) in order to identify payment purposes or resources.

Data Fields

Fields appearing across files	Fields unique to this file
institution_id	object_code_desc
object_code	

Table 14: Aux_Object_Code Data Fields

Field Name	Column Name	Data Type	Set Length	Max Length	Field Definition
Institution ID	institution_id	int	4	4	IRIS-generated unique identifier assigned to each IRIS member university for de-identification purposes. Values are four or five digit numbers
Object Code	object_code	varchar	50	6	Internal object code or other expense type category assigned to a transaction to identify payment purposes or resources
Object Code Description	object_code_desc	varchar	500	297	Description of internal object code or other expense type category assigned to a transaction; maps to object code

NSF Crosswalk

File Details

File Name: link_nsf_xwalk
Record Counts: 35,244
Field/Column Counts: 3

File Summary

This file is a crosswalk between UMETRICS unique award number and NSF award ID. These UMETRICS unique award numbers listed in this file are successfully matched to NSF awards—therefore, they are verified as NSF awards. With this crosswalk, researchers are able to link UMETRICS data to NSF award details available from the NSF source that includes award effective and expiration dates, award titles, abstracts, etc. A linking asset is NSF award ID.

Data Fields

Fields appearing across files
institution_id
unique_award_number
award_id

Table 15: Link_NSF_Xwalk Data Fields

Field Name	Column Name	Data Type	Set Length	Max Length	Field Definition
Award ID	award_id	varchar	100	7	The NSF unique award identifier assigned to each NSF award that is matched to UMETRICS award number
Institution ID	institution_id	int	4	4	IRIS-generated unique identifier assigned to each IRIS member university for de-identification purposes. Values are four or five digit numbers
Unique Award Number	unique_award_ number	varchar	400	46	University-generated unique identifier specifying an award and its funding source, made up of the 5-digit funding source code (e.g., CFDA number) and an award identifier. Award identifier may include the awarding agency's federal award ID (e.g., federal grant number, contract number, or loan number) or an internal award ID for non-federal awards. Values may include a space or dash in between them: e.g., “10.310 2010-12345-54321” (USDA example), “47.050 1234567” (NSF example), “93.865 2-R01-DK-012345-15-S1” (NIH example), “00.000 1234567” and “00.200 State Award 1” (Non-federal grant examples)

NSF Award Details

File Details

File Name: link_nsf
Record Counts: 246,941
Field/Column Counts: 11

File Summary

This file includes publicly available NSF award data downloaded from NSF in January 2020. Of 38 original data fields available from NSF, IRIS dropped a set of fields, including PI information and funded institution's name and location, for de-identification purposes. From this release, we included a flag field to identify organizations that are IRIS universities.

Data Fields

Fields appearing across files	Fields unique to this file
award_id	award_title
	award_effective_date
	award_expiration_date
	award_amount
	award_instrument
	award_instrument_code
	directorate
	division
	abstract_narration
	arra_amount

Table 16: Link_NSF Data Fields

Field Name	Column Name	Data Type	Set Length	Max Length	Field Definition
Abstract Narration	abstract_narration	varchar	8000	8000	Abstract of the award
ARRA Amount	arra_amount	varchar	50	9	Amount of funding obligated designated as ARRA funding
Award Amount	award_amount	numeric	9	5	The amount obligated to date for the project
Award Effective Date	award_effective_date	date	3	3	Effective date of the award
Award Expiration Date	award_expiration_date	date	3	3	The date on which the award expires
Award ID	award_id	varchar	20	7	The agency assigned award number (a seven digit number)
Award Instrument	award_instrument	varchar	100	33	Type of Award
Award Instrument Code	award_instrument_code	varchar	100	0	Code associated with type of award
Award Title	award_title	varchar	500	181	Descriptive title of the project
Directorate	directorate	varchar	200	60	Department of NSF funding the award
Division	division	varchar	200	70	Division of NSF funding the award

NIH Crosswalk

File Details

File Name: link_nih_xwalk
Record Counts: 86,957
Field/Column Counts: 3

File Summary

This file is a crosswalk between UMETRICS unique award number and NIH Core Project Number. These UMETRICS unique award numbers listed in this file are successfully matched to NIH Core Project Numbers—therefore, they are verified as NIH awards. With this crosswalk, researchers are able to link UMETRICS data to NIH award details available from the NIH source file that includes full project numbers, budget start / end dates, direct / indirect costs, support year, award titles, abstracts, etc. A linking asset is NIH Core Project Number.

Data Fields

Fields appearing across files
institution_id
unique_award_number
core_project_num

Table 17: Link_NIH_Xwalk Data Fields

Field Name	Column Name	Data Type	Set Length	Max Length	Field Definition
Core Project Number	core_project_num	varchar	200	11	The NIH core project number assigned to each NIH-funded project that is matched to the core project part of UMETRICS award number
Institution ID	institution_id	int	4	4	IRIS-generated unique identifier assigned to each IRIS member university for de-identification purposes. Values are four or five digit numbers
Unique Award Number	unique_award_ number	varchar	200	61	University-generated unique identifier specifying an award and its funding source, made up of the 5-digit funding source code (e.g., CFDA number) and an award identifier. Award identifier may include the awarding agency's federal award ID (e.g., federal grant number, contract number, or loan number) or an internal award ID for non-federal awards. Values may include a space or dash in between them: e.g., “10.310 2010-12345-54321” (USDA example), “47.050 1234567” (NSF example), “93.865 2-R01-DK-012345-15-S1” (NIH example), “00.000 1234567” and “00.200 State Award 1” (Non-federal grant examples)

NIH Award Details

File Details

File Name: link_nih
Record Counts: 1,706,524
Field/Column Counts: 38

File Summary

This file includes publicly available NIH award data downloaded from NIH ExPORTER in February 2020. Of 42 original data fields available from NIH, IRIS dropped 4 fields, including PI names and funded institution name and location, for de-identification purposes. From this release, we include a flag field through which award recipient organizations are identified as IRIS universities.

Data Fields

Fields appearing across files	Fields unique to this file	
core_project_num	application_id	org_fips
	activity	phr
	administering_ic	pi_ids
	application_type	project_start
	arra_funded	project_end
	award_notice_date	project_terms
	budget_start	project_title
	budget_end	serial_number
	cfda_code	study_section
	ed_inst_type	study_section_name
	foa_number	subproject_id
	full_project_num	suffix
	funding_ics	support_year
	fy	direct_cost_amt
	ic_name	indirect_cost_amt
	nih_spending_cats	total_cost
	org_dept	total_cost_sub_project
	org_district	abstract
	program_officer_name	

Table 18: Link_NIH Data Fields

Field Name	Column Name	Data Type	Set Length	Max Length	Field Definition
Abstract	abstract	varchar	8000	8000	Abstract of the funded project
Activity	activity	varchar	50	3	A 3-character code identifying the grant, contract, or intramural activity through which a project is supported. Within each funding mechanism, NIH uses 3-character activity codes (e.g., F32, K08, P01, R01, T32, etc.) to differentiate the wide variety of research-related programs NIH supports
Administrator IC	administering_ic	varchar	50	2	Administering Institute or Center - A two-character code to designate the agency, NIH Institute, or Center administering the grant
Application ID	application_id	int	4	4	A unique identifier of the project record
Application Type	application_type	varchar	50	1	A one-digit code to identify the type of application funded: 1) New Application; 2) Competing continuation; 3) Application for additional support; 4) Competing extension for an R37 award or first non-competing year of a Fast Track SBIR/STTR award; 5) Non-competing continuation; 7) Change of grantee institution; 9) Change of NIH awarding Institute or Division
ARRA Funded	arra_funded	varchar	50	1	"Y" indicates a project supported by funds appropriated through the American Recovery and Reinvestment Act of 2009
Award Notice Date	award_notice_date	date	3	3	Award notice date or Notice of Grant Award (NGA) is a legally binding document stating the government has obligated funds and which defines the period of support and the terms and conditions of award
Budget End	budget_end	date	3	3	The date when a project's funding for a particular fiscal year ends
Budget Start	budget_start	date	3	3	The date when a project's funding for a particular fiscal year begins
CFDA Code	cfda_code	varchar	50	22	A CFDA (Catalog of Federal Domestic Assistance) number used

					to identify and sort federal financial assistance programs
Core Project Number	core_project_num	varchar	50	32	Core project number
Direct Cost	direct_cost_amt	numeric	9	5	Total direct cost funding for a project for a given fiscal year. Available only for NIH awards funded in FY 2012 onward and not available for SBIR/STTR awards
ED Inst Type	ed_inst_type	varchar	50	50	Institution type
FOA Number	foa_number	varchar	50	14	The number of the funding opportunity announcement, if any, under which the project application was solicited. Funding opportunity announcements may be categorized as program announcements, requests for applications, notices of funding availability, solicitations, or other names depending on the agency and type of program. Funding opportunity announcements can be found at Grants.gov/FIND and in the NIH Guide for Grants and Contracts
Full Project Number	full_project_num	varchar	50	35	Commonly referred to as a grant number, intramural project, or contract number. For grants, this unique identification number is composed of the type code, activity code, Institute/Center code, serial number, support year, and (optional) a suffix code to designate amended applications and supplements
Funding ICs	funding_ics	varchar	50	50	The NIH Institute or Center(s) providing funding for a project are designated by their acronyms (see Institute/Center acronyms). Each funding IC is followed by a colon (:) and the amount of funding provided for the fiscal year by that IC. Multiple ICs are separated by semicolons (;). Project funding information is available only for NIH projects awarded in FY 2008 and later fiscal years

FY	fy	int	4	4	The fiscal year appropriation from which project funds were obligated
IC Name	ic_name	varchar	100	79	Full name of the administering agency, Institute, or Center
Indirect Cost	indirect_cost_amt	numeric	9	5	Total indirect cost funding for a project for a given fiscal year. Available only for NIH awards funded in FY 2012 and onward and not available for SBIR/STTR awards
NIH Spending CATS	nih_spending_cats	varchar	500	500	Congressionally-mandated reporting categories into which NIH projects are categorized. Available for fiscal years 2008 and later. Each project's spending category designations for each fiscal year are made available the following year as part of the next President's Budget request. See the Research, Condition, and Disease Categorization System for more information on the categorization process
Org Dept	org_dept	varchar	50	30	The departmental affiliation of the contact principal investigator for a project, using a standardized categorization of departments. Names are available only for medical school departments
Org District	org_district	varchar	50	2	The congressional district in which the business office of the grantee organization or contractor is located. Note that this may be different from the research performance site
Org FIPS	org_fips	varchar	50	2	The country code of the grantee organization or contractor as defined in the Federal Information Processing Standard
PHR	phr	varchar	2000	2000	Submitted as part of a grant application, this statement articulates a project's potential to improve public health
PI IDs	pi_ids	varchar	50	50	A unique identifier for each of the project Principal Investigators. Each PI in the RePORTER database has a unique identifier that is constant from project to project and year to year, but changes may be observed for investigators that have had multiple accounts in the

					past, particularly for those associated with contracts or sub-projects
Program Officer Name	program_officer_name	varchar	50	38	Name of program officer assigned to a project
Project End	project_end	date	3	3	The current end date of the project, including any future years for which commitments have been made. For subprojects of a multi-project grant, this is the end date of the parent award. Upon competitive renewal of a grant, the project end date is extended by the length of the renewal award
Project Start	project_start	date	3	3	The start date of a project. For subprojects of a multi-project grant, this is the start date of the parent award
Project Terms	project_terms	varchar	3000	3000	Thesaurus terms assigned by NIH CRISP indexers, only applicable to projects funded prior to the fiscal year 2008
Project Title	project_title	varchar	500	200	Title of the funded grant, contract, or intramural (sub)project
Serial Number	serial_number	varchar	50	8	A six-digit number assigned in serial number order within each administering organization
Study Section	study_section	varchar	50	4	A designator of the legislatively-mandated panel of subject matter experts that reviewed the research grant application for scientific and technical merit
Study Section Name	study_section_name	varchar	100	94	The full name of a regular standing Study Section that reviewed the research grant application for scientific and technical merit. Applications reviewed by panels other than regular standing study sections are designated by “Special Emphasis Panel”
Sub Project ID	subproject_id	int	50	4	A unique numeric designation assigned to subprojects of a “parent” multi-project research grant
Suffix	suffix	varchar	50	6	A suffix to the grant application number that includes the letter "A" and a serial number to identify an amended version of an original application and/or the letter "S"

					and serial number indicating a supplement to the project
Support Year	support_year	int	4	4	The year of support for a project, as shown in the full project number. For example, a project with number 5R01GM0123456-04 is in its fourth year of support
Total Cost	total_cost	numeric	9	5	Total project funding from all NIH Institute and Centers for a given fiscal year. Costs are available only for: 1) NIH and CDC grant awards (only the parent record of multi-project grants) funded in FY 2000 and later fiscal years; 2) NIH intramural projects (activity codes beginning with "Z") in FY 2007 and later fiscal years; 3) NIH contracts (activity codes beginning with "N") in FY 2007 and later fiscal years. For multi-project grants, Total Cost includes funding for all of the constituent subprojects. This field will be blank on subproject records; the total cost of each subproject is found in Total_Cost_Sub_Project (FY 2000 and later fiscal years only)
Total Cost Sub Project	total_cost_sub_project	numeric	9	5	Applies to subproject records only. Total funding for a subproject from all NIH Institute and Centers for a given fiscal year. Costs are available only for NIH awards funded in FY 2000 and later

NIH Publication Crosswalk

File Details

File Name: link_nih_pub_xwalk
Record Counts: 3,369,708
Field/Column Counts: 4

File Summary

This file is a crosswalk between UMETRICS unique award number and NIH-funded publications. There are two linking assets from this file to connect to other files, NIH Core Project Number and PubMed D. This crosswalk is built on the aforementioned linkage file (NIH crosswalk). With this crosswalk, researchers are able to link UMETRICS data to NIH-funded publication records.

Data Fields

Fields appearing across files
institution_id
unique_award_number
core_project_num
pmid

Table 19: Link_NIH_Pub_Xwalk Data Fields

Field Name	Column Name	Data Type	Set Length	Max Length	Field Definition
Core Project Number	core_project_num	varchar	200	11	The NIH core project number assigned to each NIH-funded project that is matched to the core project part of UMETRICS award number
Institution ID	institution_id	int	4	4	IRIS-generated unique identifier assigned to each IRIS member university for de-identification purposes. Values are four or five digit numbers
PMID	pmid	int	4	4	PubMed unique identifier assigned by the NIH National Library of Medicine to papers indexed in PubMed (index of abstracts). The number is 1- to 8- digits with no leading zeros
Unique Award Number	unique_award_ number	varchar	200	61	University-generated unique identifier specifying an award and its funding source, made up of the 5-digit funding source code (e.g., CFDA number) and an award identifier. Award identifier may include the awarding agency's federal award ID (e.g., federal grant number, contract number, or loan number) or an internal award ID for non-federal awards. Values may include a space or dash in between them: e.g., “10.310 2010-12345-54321” (USDA example), “47.050 1234567” (NSF example), “93.865 2-R01-DK-012345-15-S1” (NIH example), “00.000 1234567” and “00.200 State Award 1” (Non-federal grant examples)

NIH Publication Details

File Details

File Name: link_nih_pub
Record Counts: 2,242,831
Field/Column Counts: 12

File Summary

This file includes publicly available NIH publications data downloaded from NIH ExPORTER in December 2019.

Data Fields

Fields appearing across files	Fields unique to this file
pmid	issn
	journal_issue
	journal_title
	journal_title_abbr
	journal_volume
	lang
	page_number
	pmc_id
	pub_date
	pub_title
	pub_year

Table 20: Link_NIH_Pub Data Fields

Field Name	Column Name	Data Type	Set Length	Max Length	Field Definition
ISSN	issn	varchar	20	9	The International Standard Serial Number, an eight-character value that uniquely identifies the journal
Journal Issue	journal_issue	varchar	200	132	An issue, part, or supplement of the journal in which the article was published
Journal Title	journal_title	varchar	1000	280	Full journal title, taken from the NIH National Library of Medicine's cataloging data
Journal Title Abbreviation	journal_title_abbr	varchar	500	108	Standard abbreviation for the title of the journal in which the article appeared
Journal Volume	journal_volume	varchar	200	99	Volume number of the journal in which the article was published
Language	lang	varchar	20	3	Three-letter abbreviation representing the language(s) in which an article was published. List of abbreviations is available at: https://www.nlm.nih.gov/bsd/language_table
Page Number	page_number	varchar	500	138	Pages for the article, including document numbers for electronic articles
PMCID	pmc_id	varchar	20	9	A unique identifier for the article in PubMed Central (index of full-text papers). The PMCID or PMC Identifier, is assigned to each full-text paper in PubMed Central by the National Library of Medicine
PMID	pmid	int	4	4	PubMed unique identifier assigned by the NIH National Library of Medicine to papers indexed in PubMed (index of abstracts). The number is 1- to 8- digits with no leading zeros; this is the field one should use to retrieve publication details when using the award- level NIH - publication - UMETRICS crosswalk

Publication Date	pub_date	varchar	50	23	Date on which the issue of the journal was published. The standardized format includes a 4-digit year, a 3-character abbreviated month, and a 1 or 2-digit day, but the data are taken as published in the journal issue so not every record contains all elements
Publication Title	pub_title	varchar	8000	2000	Title of the article; if originally published in a non-English language this is a translation
Publication Year	pub_year	int	4	4	Year of publication, from pub_date

Team Leadership Details

File Details

File Name: link_team_leadership
Record Counts: 239,294
Field/Column Counts: 6

File Summary

This file features one row for each year of a team leader's awards. Awards are federally funded sponsored projects for which the team leader is a PI in a given year. Each row includes agency award numbers associated with Unique Award Numbers, the unique identifier used in the Award Transaction file. The unique award number is a University-generated unique identifier specifying an award and its funding source, made up of the 5-digit funding source code (e.g., CFDA number) and an award identifier.

We selected two major federal sponsors, the National Institutes of Health (NIH) and the National Science Foundation (NSF), and linked PIs to UMETRICS employees in data from 25 IRIS member universities. Several sources were integrated to construct this file, including the UMETRICS employee transaction and award files from the core collection, the award crosswalk, and the grant and PI information maintained by the NSF and NIH from the linkage collection in this release.

This file includes about 24,000 (uniquely counted) PIs who led their teams through over 60,000 sponsored projects by NIH or NSF between 2001 to 2019 (contingent on each university's available data).

Data Fields

Fields appearing across files	Fields unique to this file
agency	
award_id	
emp_number	
institution_id	
unique_award_number	
fiscal_year	

Table 21: Link_Team_Leader_Award_Year Data Fields

Field Name	Column Name	Data Type	Set Length	Max Length	Field Definition
Agency	agency	varchar	3	3	Sponsoring agency of a given award associated with the 'unique_award_number' field; values take either 'nih' or 'nsf'
Award ID	award_id	varchar	200	11	NSF-assigned award number or NIH-assigned core project number
Employee Number	emp_number	varchar	200	32	IRIS-generated unique identifier assigned to all personnel being paid by awards
Fiscal Year	Fiscal_year	int	4	4	The 12-month period beginning 1 July and ending 30 June of the following year. This is the 12-month period when a team leader is paid on sponsored projects and the same year that falls within the duration of the NSF or NIH award that pays the team leader
Institution ID	institution_id	int	4	4	IRIS-generated unique identifier assigned to each IRIS member university for de-identification purposes. Values are four or five digit numbers
Unique Award Number	unique_award_number	varchar	200	50	University-generated unique identifier specifying an award and its funding source, made up of the 5-digit funding source code (e.g., CFDA number) and an award identifier. Award identifier may include the awarding agency's federal award ID (e.g., federal grant number, contract number, or loan number) or an internal award ID for non-federal awards. Values may include a space or dash in between them: e.g., “10.310 2010-12345-54321” (USDA example), “47.050 1234567” (NSF example), “93.865 2-R01-DK-012345-15- S1” (NIH example), “00.000 1234567” and “00.200 State Award 1” (Non-federal grant examples)

Team Leader Member Year Details

File Details

File Name: link_team_membership
Record Counts: link_team_leader_member_year
Field/Column Counts: 8

File Summary

This file is the underlying data on which the team_leader_award_year file is built. The team leader-member is connected through sponsored projects in a given fiscal year, which helps researchers build their own team-based subsets.

We selected two major federal sponsors, the National Institutes of Health (NIH) and the National Science Foundation (NSF), and linked PIs to UMETRICS employees in data from 25 IRIS member universities. Several sources were integrated to construct this file, including the UMETRICS employee transaction and award files from the core collection, the award crosswalk, and the grant and PI information maintained by the NSF and NIH from the linkage collection in this release.

This file includes about 25,000 (uniquely counted) PIs who led their teams through over 60,000 sponsored projects by NIH or NSF between 2001 to 2019 (contingent on each university's available data).

Data Fields

Fields appearing across files	Fields unique to this file
agency	team_is_empty
award_id	team_leader_emp_number
institution_id	team_member_emp_number
unique_award_number	fiscal_year

Table 22: Link_Team_Leader_Member_Year Data Fields

Field Name	Column Name	Data Type	Set Length	Max Length	Field Definition
Agency	agency	varchar	3	3	Sponsoring agency of a given award associated with the 'unique_award_number' field; values take either 'nih' or 'nsf'
Award ID	award_id	varchar	200	11	NSF-assigned award number or NIH-assigned core project number
Fiscal Year	fiscal_year	int	4	4	The 12-month period beginning 1 July and ending 30 June of the following year. This is the 12-month period when a team leader / member is paid on sponsored projects and the same year that falls within the duration of the NSF or NIH award that pays the team leader / member
Institution ID	institution_id	int	4	4	IRIS-generated unique identifier assigned to each IRIS member university for de-identification purposes. Values are four or five digit numbers
Team leader - member pair missing	pair_missing	int	4	4	A binary indicator for team-member pairing associated with an award in a given fiscal year a; in each record (row) if a team leader is associated with other employee (including oneself) through the same grant, it is coded as 1; if there is no team member for a given leader in a given grant, it is coded as 0
Team Leader Employee Number	team_leader_emp_number	varchar	200	32	Unique employee number of team leader (= PIs of NSF or NIH sponsored projects); this employee number is IRIS-generated unique identifier assigned to all personnel being paid by awards

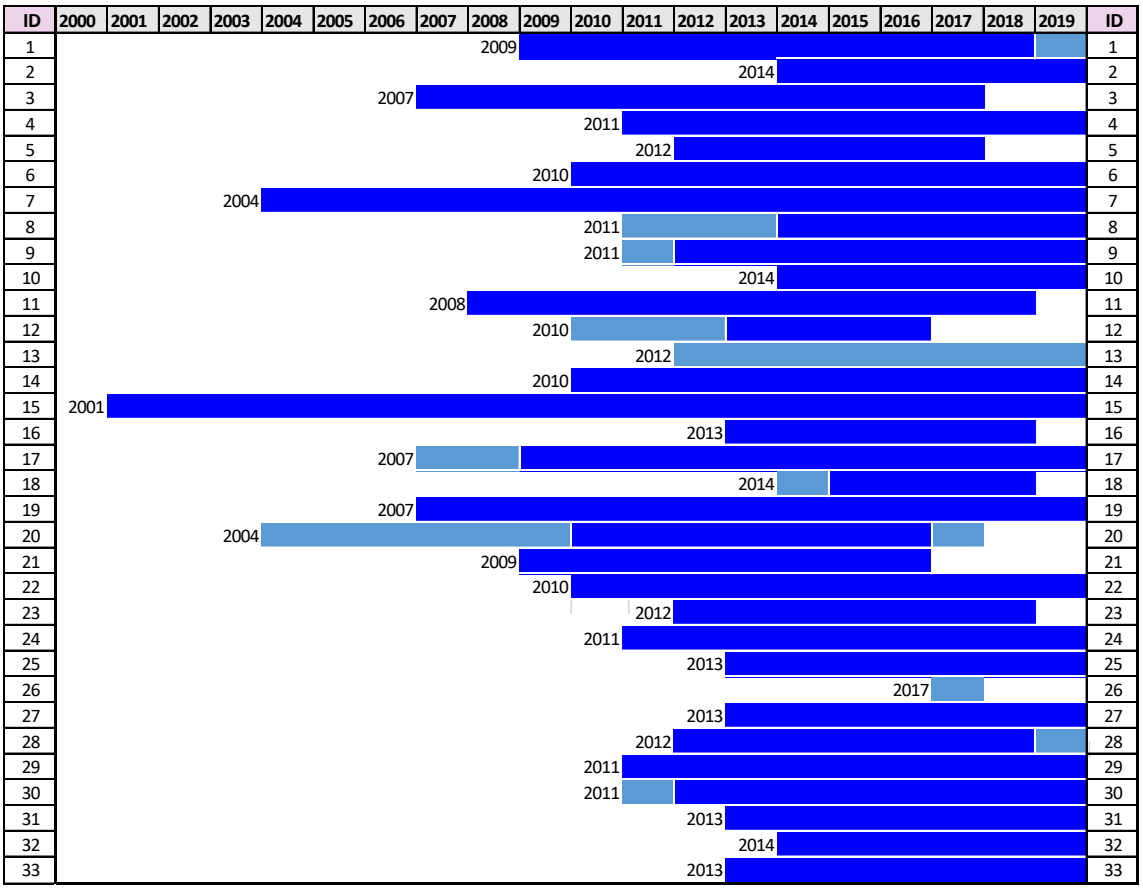
Team Member Employee Number	team_member_emp_number	varchar	200	32	Unique employee number of team members who are paid on sponsored projects during the period of the team leader (PI) record in NSF or NIH award duration data; in some cases the team_leader_emp_number and team_member_emp_number is identical because in UMETRICS data team leaders are recorded in the Employee transaction file as personnel being paid by the award whose PIs are themselves
Unique Award Number	unique_award_number	varchar	200	50	University-generated unique identifier specifying an award and its funding source, made up of the 5-digit funding source code (e.g., CFDA number) and an award identifier. Award identifier may include the awarding agency's federal award ID (e.g., federal grant number, contract number, or loan number) or an internal award ID for non-federal awards. Values may include a space or dash in between them: e.g., “10.310 2010-12345-54321” (USDA example), “47.050 1234567” (NSF example), “93.865 2-R01-DK-012345-15-S1” (NIH example), “00.000 1234567” and “00.200 State Award 1” (Non-federal grant examples)

Data Summary & Descriptive Statistics

File Availability & Data Coverage

University representation and temporal data coverage varies across files and institutions. IRIS asks universities to submit their administrative data in fiscal year format, generally starting on July 1 and ending June 30 in the following year, though some universities run on a slightly different fiscal year, starting in September and ending in August of the following year. Note that the current release contains data through an October 2019 submission by universities, thus may include partial fiscal year data (e.g., until June/July/August 2019).

Figure 4: Temporal Coverage by University



Note: The color-coding in demonstrates coverage by file—deep blue indicates core file availability for all award, employee, vendor, and subaward files while light blue indicates that not all four files are available for a given year.

Basic Record Counts

The following tables show award, employee, vendor, and subaward counts with additional information such as type of awards, sponsoring agencies, employee occupations, gender, ethnicity, vendor types (organization vs. individuals), etc.

Table 23: Award Data Summary Statistics

Number of Universities	Total Number of Unique Awards	Min	Max	Mean	Standard Deviation
33	402,396	467	52,978	12,193	10,348.20

Table 24: Employee Data Summary Statistics

Number of Universities	Total Number of Unique Employees	Min	Max	Mean	Standard Deviation
32	720,679	1,386	84,751	22,521	19,188.82

Table 25: Vendor Data Summary Statistics

Number of Universities	Total Number of Unique Vendors	Min	Max	Mean	Standard Deviation
32	971,194	2,124	102,171	30,349	23,174.22

Table 26: Subaward Data Summary Statistics

Number of Universities	Total Number of Unique Subawards	Min	Max	Mean	Standard Deviation
32	29,629	273	3,072	925	557.13

Basic Data Distribution

Table 27: Employee counts by occupation using the 'systematic_occupational_class' field

Faculty	Staff	Post Graduate Research	Graduate	Under-graduate	Other Student	TOTAL
94,944	291,652	57,350	166,857	77,600	141,868	830,271

Table 28: Employee counts by age

Age Group	YOB Range	Count of People
1	between_1943_and_1947	5344
2	between_1948_and_1952	11031
3	between_1953_and_1957	15809
4	between_1958_and_1962	18092
5	between_1963_and_1967	20976
6	between_1968_and_1972	27568
7	between_1973_and_1977	39412
8	between_1978_and_1982	65208
9	between_1983_and_1987	92898
10	between_1988_and_1992	116008
11	between_1993_and_1999	108580
12	masked	6641
13	na	20916
Total		548483

Table 29: Employee counts by imputed gender

Employee Gender		
548,483		
	Count	***Percentage***
Female	245482	44.8%
Male	228670	41.7%
Unknown (null)	74,331	13.6%

Table 30: Employee counts by imputed ethnicity groups

Imputed Ethnicity	Employee Count	% of Total
ENGLISH	223689	40.8%
CHINESE	64935	11.8%
GERMAN	40409	7.4%
HISPANIC	36349	6.6%
INDIAN	25814	4.7%
KOREAN	15664	2.9%
ARAB	13400	2.4%
SLAV	12025	2.2%
ITALIAN	11658	2.1%
FRENCH	11425	2.1%
NORDIC	7604	1.4%

JAPANESE	4684	0.9%
DUTCH	4552	0.8%
AFRICAN	2800	0.5%
TURKISH	2265	0.4%
ISRAELI	1908	0.3%
GREEK	1541	0.3%
HUNGARIAN	1154	0.2%
ROMANIAN	558	0.1%
THAI	363	0.1%
VIETNAMESE	239	0.0%
BALTIC	187	0.0%
INDONESIAN	66	0.0%
Other (narrowly-defined ethnicity group)	16861	3.1%
Null	48333	8.8%

Table 31: Vendor counts and distribution by vendor type

Unique Vendor Count		
902,370		
	Count	***Percentage***
Organization	297,832	33%
Person	604,538	67%

Table 32: Subaward counts and distribution by vendor type

Unique Subaward Count		
23,552		
	Count	***Percentage***
Organization	21,464	91.1%
Person	2,088	8.9%

Record Linkage Results

Table 33: NSF award match results

Number of Universities	Total Number of Unique NSF Awards	Matched NSF Awards	Match Rate
32	42,366	34,764	82.1%

Table 34: NIH award match results

Number of Universities	Total Number of Unique NIH Awards	Matched NIH Awards	Match Rate
31	97,736	80,651	82.5%

Table 35: NIH-funded publication match results

Number of Universities	Matched Awards (NIH-UMETRICS-PubMed)	NIH Core Project Numbers	NIH-funded publications (PMIDs)
31	67,938	37,619	613,439

Methodology

Data Cleaning & De-identification

IRIS seeks to add value to the data received by universities by identifying and resolving data discrepancies when possible, providing standardized values (e.g., occupational classifications, sub-organization units) as well as through various cleaning processes including name standardization. More importantly, we carefully mask information from the release files in order to minimize the risk of re-identifying universities or individuals from particular data elements. When preparing the dataset for research use, IRIS data processing methods included but were not limited to:

1) Assigning IRIS-generated employee numbers as unique IDs, using the HashBytes function built in the SQL database;
2) Removing university names, campus names and location information of IRIS universities;
3) Removing any personally identifiable information (e.g., any individual names, personal employee identification numbers, and EINs if vendors and contractors (such as individual consulting service providers) are individuals);
4) Replacing any university-submitted identification numbers with randomly assigned numbers for a new set of IDs;
5) Replacing campus-level vendor and subaward recipient's identification numbers with randomly assigned unique identification numbers that help to disambiguate them at the national level;
6) Replacing names of individuals in the unique award number field with a hashed ID instead of masking so that researchers still can link across files if needed—note that the unique award number field usually includes no names with the exception of cases with Intergovernmental Personnel Act (IPA)-related awards;
7) Generating and assigning occupation classification to all personnel paid by awards with two different methods;
8) Cleaning and standardizing names of vendors and subaward recipients;
9) Cleaning and standardizing names of award sponsors;
10) Cleaning unique award numbers across files by removing white spaces and punctuations in the beginning of the string; and,
11) Cleaning and re-formatting CFDA numbers (in the correct form of ##.###) in the CFDA field and replacing with 00.000 if no CFDA or OSF code are not identified in the Unique Award Number field.

Classification

Occupation Classification

As noted earlier, in this year's release we included three occupational classification fields: 1) 'occupational_class', 2) 'umetrics_occupational_class', and 3) 'systematic_occupational_class'. Although the first one is the raw occupational data we received from IRIS universities, the other two are generated by Dr. Bruce Weinberg (an IRIS Co-PI)'s team at the Ohio State University and IRIS in Ann Arbor.

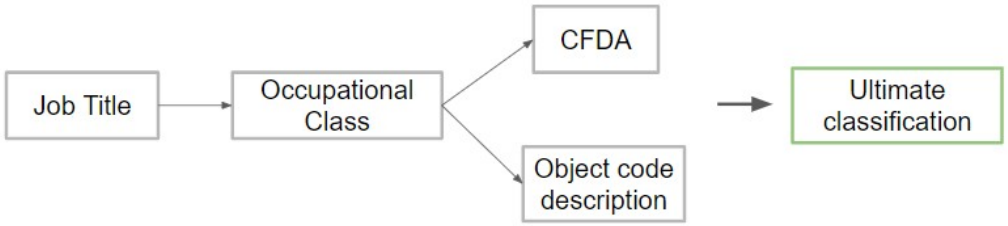
1) 'umetrics_occupational_class'

Dr. Weinberg has led the occupation classification coding project at OSU. In defining occupations more generally, an occupation classification coding rule and practice was carefully developed. The university-assigned job titles are manually reviewed and classified into one of the UMETRICS occupational categories in the umetrics_occupational_class field. The first tier of classification considers six categories of relationships to a university: Faculty, Staff, Postgraduate Research, Graduate Student, Undergraduate, and Other. A second tier considers six categories of job responsibilities to disaggregate the "Staff" titles from the first tier: Clinical Staff, Research Staff, Research Facilitation Staff, Instructional Staff, Technical Support, and Other Staff. For detailed occupational category descriptions, see Appendix D.

2) 'Systematic_occupational_class'

This new occupational classification was initially developed by IRIS Data Production Team for University reporting purposes, and then adopted by IRIS Research Support Team to include in the release file. This new classification maintains most of the existing classification groups, with two major changes: Undergraduate, Graduate Student, Postgraduate Research, Faculty, Other Student (newly added for ambiguous job titles), and Staff. Unlike the UMETRICS occ classification, here we do not emphasize job function, therefore, Staff was not further categorized into other groups (e.g. Technical Support, Research Facilitation, etc.). In addition, the category Staff served almost as an "Other" category to encompass any employees who did not fall into the other groups they are considered staff. This more systematic classification relies on more than one field (job title); instead, multiple fields are being considered when choosing one status / job classification. These include: Job Title, Occupational Classification, CFDA, and Object Code Description. More importantly, key words and phrases in some fields are used to categorize occ. Classification takes place in three steps: 1) Records with no conflicting information are classified; 2) Individual fields of the remaining records are classified; and, 3) The field-level classifications from step 2 are combined to classify remaining records.

Figure 5: New occupational classification steps



Note that with this classification, employees who are categorized as Students are increased by 7% and Graduate Students are increased by 4% if compared to UMETRICS occ classification. See Data Summary section for employee distribution by each occupation group using two different methods.

Imputation

Source Data for Gender and Ethnicity Imputation

In response to increased needs from researchers for more demographic variables, this year's release includes imputation work that was originally introduced by Dr. Bruce Weinberg (IRIS Co-PI)'s team who had collaborated with the research group led by Vetle Torvik at the University of Illinois Urbana-Champaign. IRIS has adopted and adapted their work to apply to the 2020 release employee name data to develop an auxiliary file (release2020.aux_emp_demographics) that includes demographic variables.

The imputation relies on primarily employees' first and last names, thus name matching. This is done in conjunction with datasets on first and last name frequencies, ethnicities, and genders derived from Ethnea, an ethnicity classifier developed by Vetle Torvik et al (for a quick reference, see https://www.ideals.illinois.edu/bitstream/handle/2142/88927/ethnea.pdf?sequence=2&isAllowed=y ; for Ethnea the Ethnicity Predictor, see http://abel.lis.illinois.edu/cgi-bin/ethnea/search.py).

The source data from Ethnea that we matched to UMETRICS employee names include two files: 1) Ethnea Last Name dataset (1,457,761 unique records) and, Ethnea First Name dataset (274,491 rows, 120,715 unique first names).

Name Matching and Imputation Method

Initial processing uses Stata to standardize the column names in the UMETRICS employee name file to match column names to Ethnea datasets, as well as preprocessing to remove leading and trailing whitespace and to convert all fields to uppercase. Ethnicity is derived through a merge of UMETRICS employee names and the Ethnea Last Name dataset. Note that each last name may have more than one ethnicity associated with it. If a given last name is not in the Ethnea dataset, then the matching process ends here and will not be moved to succeeding matching and imputation steps

using Ethnea datasets—this leads to no imputation outcome (therefore, a null value is given in the 'imputed_ethnicity' or 'imputed_gender' field).

If the last name of a given employee from the UMETRICS employee name is matched to one of the Ethnea last name records, then the record linkage will proceed to the next step, i.e., first name matching using the Ethnea First Name dataset. We perform multiple matching steps, using the first name field and one of the ethnicities identified for a specific employee record (noted as eth1, eth2, etc.) as keys. For each matching step, if a first name/ethnicity pair is found in the Ethnea first name dataset, the corresponding gender for the first name/ethnicity pair is recorded.

Next, using the first name field, we merge the UMETRICS employee name dataset with a first name maximum frequency count for each first name/gender pair in the Ethnea First Name dataset. For each match, the corresponding gender listed in the dataset is recorded.

Lastly, after all these steps are merged, the imputed gender and ethnicity identified for each employee name is selected, giving a priority to gender and ethnicity identified earlier in the matching process.

Supplementary Method for Imputing Gender

For all employee names where no imputed gender was identified using the Ethnea methodology, we used the Python package gender_guesser to impute gender. For this methodology, only the first name is used for imputing gender. If the resulting gender is determined as either androgynous (indicated as 'andy' in the package output) or unknown, the imputed gender is reported as "UNK". We replaced these coding with null in the release file. The distribution of employees by gender and ethnicity is provided in the Data Summary section earlier.

Disambiguation

Non-federal (Nonprofit) Awarding Organization Names

As part of cleaning and disambiguating names of funding organizations, we particularly focused on nonprofit organizations for the 2020 release. Our initial step was to compile the source file to which we matched funding source names from the UMETRICS Award file. The source file contains 32,668 nonprofit foundation names and associated details–data were downloaded from the Foundation Directory Online database.

Our initial data pre-processing, applied to both the source and UMETRICS Award file, was converting all strings to lowercase, removing punctuations,and common stopwords, and expanding common abbreviations and acronyms, especially those used to indicate foundation such as 'fndtn', 'fndt', 'fndn', 'fdtn', 'fdt', 'fnd', 'fdn', 'foundatn', 'foundat', etc.

We then used the Python package sklearn implementation of a count vectorizer on the combined set of both foundation names and UMETRICS funding source names, generating a count of

all unique 3-grams found in the corpus. Using this, we created a TF-IDF vectorizer using the counts for all unique 3-grams, and created a weighted vector for each foundation name and funding source name.

We used sparse_dot_topn (https://github.com/ing-bank/sparse_dot_topn), a package developed for fast sparse matrix calculation of cosine similarity, in order to generate the cosine similarity of each pair of foundation name to UMETRICS funding source name. We decided to only keep pairs that met our set threshold of having a cosine similarity of 0.8 or greater, due to a large increase of false positives during testing with lower thresholds.

We were able to generate 5,223 nonprofit foundation names - funding source name pairs that met our set threshold. Upon manual review, the use of TF-IDF in the methodology was able to link funding source names to foundations even when the funding source name had abbreviations that we did not consider. (ex. “Amer” for “America”). In the lower end of the cosine score threshold, we found that the methodology also helps match names that are missing information, or are in a different order, such as foundations and funding source names with people as their namesake.

Linkage

NIH Award Linkage

IRIS has made no changes in the NIH-UMETRICS award linkage process for this current release. Same as last year’s release 2019, we used NIH Core Project numbers as a linking data entity and the code was built on a thorough investigation into university-specific NIH award number formatting in our UMETRICS data for better data extraction for matching. Matching results were reported in the Data Summary section earlier.

The UMETRICS unique award number/NIH core project award linkage process matches extracts of the unique award numbers to exact listed NIH core project award numbers. Most NIH core project award numbers are an eleven-digit string composed of three independent parts:[1]

- a three-character activity starting with a letter (e.g., R01; R23; KL2; U01; ZIA);
- a two-letter institute (e.g., AG, DK, HD);[2] and,
- a six-digit serial number (000000 through, currently, 998095).

[1] In the NIH ExPORTER between 1999 and 2019, only 0.52% of projects have a core project number that does not match this format.

[2] N.B. Due to the NIH process of assigning the core project award number, its two-letter institute does not necessarily match either the funding institute or the administering institute ultimately listed in the NIH ExPORTER.

Any combination of these is theoretically possible. At the time of linkage, NIH ExPORTER contained 320,571 unique core project award numbers. Due to inconsistent treatment of leading zeros, for linkage these core project award numbers are decomposed into a three-character activity string, a two-character institute string, and a serial number integer.

Unique Award Number Cleaning

- **Stopwords removed:** To reduce false positives and improve performance, several common words found in these unique award numbers are removed, for example: 'SIGNED', 'ACTIVITY', and 'NIH'. Regular expressions are used to concisely remove patterns that are not part of standard core project award numbers.[3] Stopwords are removed before and after left and right truncation.
- **Left digits and right letters truncated:** Because all core project award numbers start with a letter and end with a serial number, unique award numbers are stripped of all non-letter characters on the left and all non-numeric characters on the right.
- **Minimum length imposed**: At the conclusion of cleaning, all cleaned award numbers with fewer than six characters are ineligible for further linkage. This reflects the minimum possible match criteria later imposed: two letters for the institute and four for the serial number.

Parsing and Matching

The program parses and attempts to exactly match unique award number extracts to core project award numbers with progressively less restrictive extraction criteria:

1. Full activity: An ideal match would include the full activity, institute, and serial number, exactly replicating the core project number itself. In the various ways that institutions record this information, the institute and serial are typically contiguous while the activity can appear more distant. A three-stage sequence is used to attempt to capture a full activity:

- A valid activity is no more than two characters to the left of institute and serial.
- A valid activity is found to the right of institute and serial.

[3] For example, /HHS[A-Z]?/ captures 'HHS' alone as well as the prefix of contract numbers that does not feature in standard core project award numbers. /DA?TE?D/ captures several common abbreviations of 'DATED'.

- A valid activity is found anywhere to the left of institute and serial.

In each case, only an exact match to one of the 320,571 unique core project award numbers is recorded as successful linkage.[4]

2. Activity initial: For unique award numbers that have not matched on full activity, a similar search attempts to find a valid institute and serial along with an initial that uniquely identifies a core project award number. The match must not be ambiguous between existing core project award numbers. For example, unique award number 'RDK234567' could be linked to core project award number 'R01DK234567'—but only if there is no 'R23DK234567.' Only one- to-one matches are made, and so if both R01 and R23 existed in this example, 'RDK234567' would not be matched at all.

3. No activity: In many cases, no activity has been listed in the unique award number but the combination of institute and serial number can be used to uniquely identify an NIH core project award number. As with the activity initial matching, only a one-to-one match is kept, and any institute-serial combinations that would match multiple awards are not linked.

Final Composition

Each successfully linked unique award number is reconstructed into the standard eleven-character string by concatenating the fields together and with leading zeros in the serial number.

NSF Award Linkage

IRIS has made only one minor change in the NSF-UMETRICS award linkage process for this current release. Same as last year, the National Science Foundation (NSF) agency award data used for linkage were downloaded in January 2020 from the NSF website.

The algorithm is composed of four matching steps, of decreasing accuracy/confidence to the NSF award ID. Step 1 reflects an exact match with the NSF Award ID as provided by the IRIS university. Step 2 reflects exact matches after removal of any leading alpha characters and white spaces. There is an intermediary cleaning step between each matching step, potentially increasing the chance of matching the underlying UMETRICS data element to an NSF Award ID. As such, lower step numbers (beginning with 1) reflect more accuracy/confidence in exact matching, and higher step numbers reflect less accuracy/confidence.

The algorithm derives two distinct formats of award IDs from UMETRICS unique award

numbers for matching:

Full ID (UMETRICS full unique award numbers excluding leading CFDA numbers): This represents all characters after the CFDA number less strings such as “SUB.” Examples are shown below:

Table 36: Examples of Award Number Formats Used for Match

UMETRICS unique award number raw string	Cleaned string (Full ID) for further processing and matching
47.041 1355816	1355816
47.041 1355816 SUB	1355816
47.049 1654485 Amendment 00	1654485
47.049 DMR 13-08584	DMR 13-08584
47.080 RR197-017/4941206	RR197-017/4941206
47.000 Early award	Early award
47.076 490K755	490K755

Core ID: The Full ID minus various numeric characters and/or punctuations and/or strings that can be categorized into Patterns #1-4 as shown in the table below.

For all steps, if a Core ID is unavailable (no numeric characters in the award ID), the Full ID is used instead. Note that in the matching process the same UMETRICS unique award number may be matched more than once across multiple steps. After all four matching steps are complete, the results are concatenated in a Pandas DataFrame, and using the matched NSF award ID, UMETRICS unique award number, and CFDA (parsed token, 47.XXX) as a reference, all duplicated rows are dropped, leaving successful matches at the earliest match step that resulted in a match.

Table 37: Award Matching Steps

<table>
<tr><th>Step</th><th>UMETRICS data element (cleaned string)</th><th></th><th>Matched /Non-matched NSF Award Identifiers</th></tr>
<tr><td>1</td><td>Full ID
e.g., 1355816
e.g., 1644899</td><td>←→</td><td>NSF Award ID e.g., 1355816
(matched)
e.g., 1644899</td></tr>
<tr><td>2</td><td>Core ID Pattern #1
(All characters after alpha characters until whitespace)
e.g., DUE 0631188 → 0631188
e.g., BCS 04-37581 → 04-37581</td><td>←→</td><td>NSF Award ID

e.g., 0631188
(matched)
e.g., 0437581 (no match)</td></tr>
<tr><td>3</td><td>Core ID Pattern #2
(All numeric characters without the “-” character)
e.g., CMMI-1138640 → 1138640 e.g.,
BCS 04-37581 → 0437581</td><td>←→</td><td>NSF Award ID

e.g., 1138640
(matched)
e.g., 0437581
(matched)</td></tr>
<tr><td>4</td><td>Core ID Pattern #3
(First 7 characters of Core ID from Pattern # 1)
<table>
<tr><th>Unique Award Number</th><th>Core ID Pattern 1</th><th>Matching Pattern #4</th></tr>
<tr><td>CHE-1058501/001</td><td>1058501/001</td><td>1058501</td></tr>
<tr><td>CHE-0948017/007/0</td><td>0948017/007/008</td><td>0948017</td></tr>
<tr><td>0838536-BCS</td><td>0838536-BCS</td><td>0838536</td></tr>
</table></td><td>←→</td><td>NSF Award ID

e.g., 1058501
(matched)
e.g., 0948017
(matched)
e.g., 0838536
(matched)</td></tr>
<tr><td></td><td colspan="3">Remove duplicated matches by keeping matched pairs from earlier matching step(s)</td></tr>
</table>

For error detection, the NSF award matching code generates a list of UMETRICS awards that failed to match any existing NSF award numbers during the matching methodology. Upon manual review, we noted an inconsistency in the UMETRICS unique award number format, potentially causing failed matches, affecting 1,850 awards. To check if additional matches can be recovered from the list of non-matched UMETRICS awards, we developed a separate matching methodology, performing minor preprocessing, and using regular expressions to extract all 7-digit sequences in the UMETRICS award number, in line with NSF's 7-digit award ID format. The first 7-digit sequence found for a specific unique award number was recorded as a potential NSF award ID, and used for matching with the NSF

award data. Overall, this new matching method recovered 1,467 matches from the 1,850 affected awards.

NIH-funded publication linkage

This file includes publicly available NIH publication data downloaded from NIH ExPORTER in January 2020. Using the NIH-UMETRICS award crosswalk (thus 31 universities successfully matched), we applied an exact match method to bridge from UMETRICS award data to NIH Core Project Numbers and then to PMIDs (PubMed assigned unique publication identifiers). Note that linking entities (NIH Core Project Number—PMID) are in many-to-many relationships.

NSF / NIH Principal Investigator (PI) and UMETRICS Employee Name Matching

Methodology

We first join UMETRICS employee name dataset with the UMETRICS employee dataset to get the name details for each employee paid under a corresponding UMETRICS unique award number. We then link the results to the UMETRICS-federal award data crosswalk in order to get the corresponding federal award identifier for said UMETRICS unique award numbers (For NSF, this would be the 7-digit award ID, and for NIH, this would be the core award number).

The resulting table is joined with the federal award data using the specific federal award identifier in use. This generates a table of all combinations of paired UMETRICS employee names - Federal Award Dataset name for each award present in both datasets.

Source Files

The datasets to be used for the name matching process are:
2020 Data Release: UMETRICS Employee Name table
2020 Data Release: UMETRICS Employee table
2020 Data Release: NIH/NSF to UMETRICS crosswalks
NIH Award Data - all Principal Investigator names extracted and separated
NSF Award Data - all Principal Investigator names

Name Matching

Matching is done in the following 4 steps and are applied to all UMETRICS-Federal data name pairs. Exact matching looks for all matches where the first name and last name fields are identical in both

datasets. Other methods compare between the concatenation of the first and last name in each dataset, and provide a score ranging from 0 to 1:

1. Exact match
2. Token set ratio (Python fuzzywuzzy implementation)
3. Affine gap (Python py_stringmatching implementation)
4. Soundex (Python py_stringmatching implementation)

Lastly, we set a threshold of 0.8 for all matches. This threshold was chosen upon experimentation on lower thresholds leading to a significant increase in false matches.

Hierarchy of Results

Upon manual review of the results, we noted that UMETRICS - federal data name pairs may be matched multiple times through all 4 algorithms. Therefore, we developed the following hierarchical tagging system to indicate confidence in the match, in descending levels of confidence:

1. Exact: Exact match
2. Consensus: Exact match failed, but all fuzzy matching algorithms indicated name pair match
3. Agreement: Exact match failed, 2 of the 3 fuzzy matching algorithms indicated name pair match
4. TSR: Token set ratio indicated match. Upon manual review of match results, token set ratio provides good matches at the set level of confidence of 0.8

Table 38: Principal Investigator Name Matching Result

Type	NSF	NIH
Unique names	131,706	244,620
Unique UMETRICS pairs	9,741	16,030
Exact	9,125	14,668
Agreement	151	518
Consensus	43	209
TSR	422	635

Future Data Releases

IRIS plans to continue releasing a research dataset annually in Spring/Summer going forward. Each annual release contains updates to the core files based on new submissions from IRIS member universities. The annual release also includes updates for the auxiliary and linkage collection files and any newly developed files. In addition to this annual release, IRIS provides additional supplemental releases throughout the year as they become available, typically new linkage files. From this summer into early fall, IRIS plans to have two supplementary releases focusing on record linkage at the individual level: 1) UMETRICS-PubMed linkage, 2) UMETRICS-Patent linkage.

We also plan to do more verification work for imputed gender and ethnicity data. IRIS continues to devote considerable effort to normalizing and disambiguating vendor names, and made significant improvements in the past year to the script that flags individuals versus organizations.

Appendices

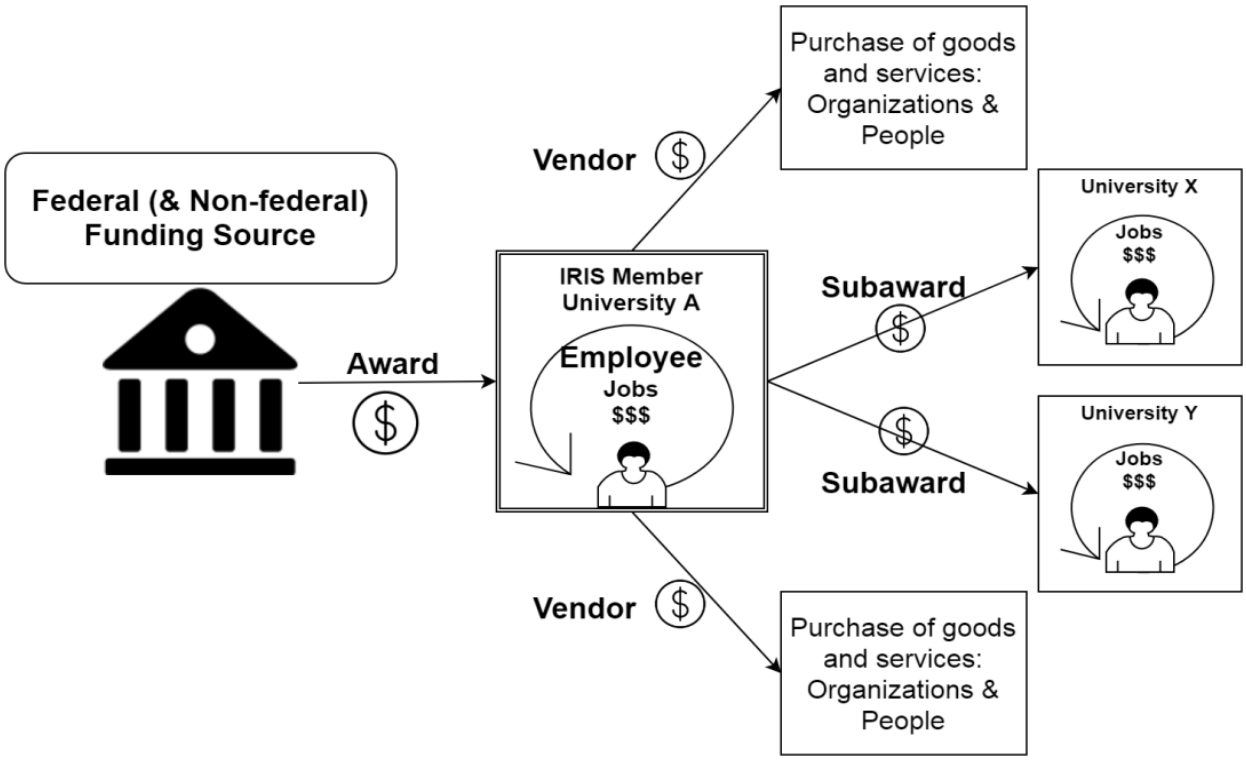
Appendix A Universities in the Release 2020 Dataset

Thirty-three (33) universities represented in the 2020 dataset are listed below. Asterisk (*) indicates inactive members as of June 2020.

- Boston University
- Emory University
- Johns Hopkins University
- Indiana University
- Michigan State University
- New York University
- Northwestern University
- Ohio State University
- Oregon State University*
- Pennsylvania State University
- Princeton University
- Purdue University*
- Rutgers University
- Stony Brook University
- Texas A&M
- University of Arizona*
- University of California - San Diego
- University of Cincinnati*
- University of Colorado, Boulder
- University of Hawaii*
- University of Illinois at Urbana-Champaign
- University of Iowa*
- University of Kansas
- University of Michigan
- University of Missouri
- University of Oregon
- University of Pennsylvania
- University of Pittsburgh
- University of Texas – Austin
- University of Utah
- University of Virginia*
- University of Wisconsin - Madison
- Washington University in St. Louis

Appendix B UMETRICS Core File Relationship in Monetary Flow

Figure below demonstrates the relationship among the four files in the IRIS core collection.



Appendix C Other Funding Source (OFS) Codes

These codes are used in place of an official CFDA code for non-federal awards.

Funding source	Other Funding Source (OFS) code	Definition	Examples of Funders included in this code
Unknown or Generic Nonfederal	**00.000**	Code for use when specific nonfederal funding type cannot be determined	
Federal - Other	**00.070**	Agencies or federal contracts that do not have CFDA numbers	CIA, FEMA, Veterans' Administration, etc.
CIA	00.071	Central Intelligence Agency	CIA
FEMA	00.072	Federal Emergency Management Administration	FEMA
Veterans' Administration	00.073	Veterans' Administration	VA
Institutional Investment	**00.100**	Award provided internally by your own university to support research at your university	
State Funding - Home State	**00.200**	Funding from a state agency in the state where the university is based	For a university in New York, the New York State Division of Veterans' Services
State Funding - Nonresident State	**00.300**	Funding from a state agency in states other than the state where the university is based	For a university outside New York, the New York State Division of Veterans' Services
Specific Nonresident State	00.3xx	Funding from a state agency in a particular state other than the state where the university is based. See separate worksheet below for state-specific codes (xx is the FIPS state numeric code)	03.36 New York State Division of Veterans' Services 03.56 State of Wyoming Game and Fish Department
Local Funding (City/ County)	**00.400**	Funding provided by a local government body; that is, a government body smaller than a state.	City of Seattle Jefferson County Conservation District Menominee Indian Tribe of Wisconsin Cook County Health Department
Business / For Profit	**00.500**	Industrial / Commercial Funders that do business in the United States	American Steamship Company Rembrandt Foods
Nonprofit	**00.600**	Nonprofit funding; use a subcategory if desire, to be more specific	American Heart Association Annenberg Foundation
Philanthropic	00.610	Money given by an individual or corporation that sits in a University Foundation account and is expended to conduct research	Bill & Melinda Gates Foundation Ford Foundation
Public Foundations	00.620	Funding provided by a foundation that accepts public donations	The CDC Foundation The March of Dimes The Pediatric Cancer Research Foundation.
Private Foundations	00.630	Funding that is given by foundations that are privately funded.	the Annenberg Foundation the William & Flora Hewlett Foundation
Private Associations	00.640	Funding provided by professional medical societies, fraternal organizations, and other private associations	Indiana Elks Charities American College of Radiology Kiwanis International
Hospital / Medical Centers	00.700	Funding from a health care facility	Mayo Clinic Hospital for Sick Children Tufts Medical Center
Your university's hospital or medical center	00.710	Funding from your university's hospital or medical center	For the University of Pittsburgh, UPMC
A hospital or medical center other than your university's	00.720	Funding from a different university's hospital or medical center	For the University of Pennsylvania, UPMC
Universities / Colleges	00.800	Any award coming from universities and colleges including flow-through funding	The University of Maryland Black Hills State University London School of Economics
Foreign	00.900	Funding from non-US corporations, associations, or governments	Flax Canada Government of Pakistan
Foreign Government	00.910	Funding from governments outside the U.S.	Government of Pakistan
Foreign Non-Government	00.920	Funding from corporations, associations, etc outside the U.S.	Flax Canada

Appendix D Occupational Category Descriptions (Tier 1 & 2)

Tier 1 Descriptions

Category	Description	Example Job Titles
Faculty	Advanced academic employees who are directly involved in scientific research, clinical care, and/or scientific instruction. Faculty have the highest degree in their area of study and tenure or a tenure-track/"permanent" position at their university (e.g., in the case of medical school faculty)	Academic Administrators, if faculty (e.g., Dean, Provost, Center Director) Adjunct Professor Clinical Faculty Professor (Assistant, Associate, Full and Chaired) Research Faculty Visiting Professor
Post Graduate Research	Individuals holding terminal degrees (PhD, MD) who are in temporary training status	Clinical Fellow Medical Resident/Intern/Fellow Postdoctoral Fellow/Researcher Research Associate

Graduate Student	Students earning advanced degrees	Graduate Student Medical/Dental/Nursing Student Research Assistant
Undergraduate	Students earning baccalaureate/other degrees including full-time, part-time, summer research assistants, and work-study. Includes high school students who would likely be acting in a similar capacity	Intern Nursing Student (BA programs) Student Worker Undergraduate Student
Staff	Individuals in non-faculty roles who are clinical staff, research and research support staff, instructional staff, and technical support staff	(See Table 16 below)
Other	Positions that support general university functions such as undergraduate education and student activities. Employees whose titles cannot be attributed to the scientific research enterprise	

Tier 2 Descriptions

Category	Description	Example Job Titles
Clinical	All non-faculty health care professionals (clinicians)	Clinical Psychologist Dental Hygienist Dietician or Nutritionist Genetics Counselor Nurse Physical Therapist Social Worker
Research	Non-faculty scientists, engineers, analysts and technicians. Research staff are directly involved in conducting research, usually have advanced degrees, and are skilled and specialized in some area of science, technology, equipment or research but are not faculty members	Engineer Lab Manager Medical Technician Regulatory Officer Research Analyst Research Associate, Specialist, or Professional Staff Scientist Statistician Technician

Research Facilitation	Administrative employees who are not specifically employed for scientific research purposes but perform job tasks that support the research enterprise. Individuals who serve as managers/coordinators/facilitators for laboratory studies/clinical trials/large facilities/research programs. Employees who direct and influence scientific research activity from the level of the laboratory up to the level of the university/research center	Administrative Staff Associate Center Director Associate Dean or Provost Communications Specialist Grants Manager Interviewer Lab Coordinator (not Lab Manager) Laboratory Aide Managing Director Operations Manager Project Manager Study Coordinator
Technical Support	Technical employees who are not specifically employed for scientific research purposes but perform job tasks that support the research enterprise	Data Entry/Data Analyst Information Technology Manager Network Support Specialist Programmer Software Engineer
Instructional	Staff members who are instructional or academic specialists	Academic Specialist Instructor Lecturer

Other Staff	All other research staff that do not clearly fall into another category

Appendix E IRIS Research Bibliography
(Last update: March 2020)

PUBLISHED PAPERS

Money for Something: Braided Funding and the Structure and Output of Research Groups

Funk R, Glennon B, Lane JI, Murciano-Goroff R, Ross M
IZA Discussion Paper No. 12762, available online 18 Nov 2019
https://ssrn.com/abstract=3488189

Abstract
In 2017, the federal government invested over $40 billion on university research; another $16 billion came from private sector sources. The expectation is that these investments will bear varied fruits, including outputs like more economic growth, more scientific advances, the training and development of future scientists, and a more diverse pipeline of STEM researchers; an expectation that is supported by the work of recent Nobel Laureate in Economics, Paul Romer. Yet volatility in federal funding, highlighted by a 35 day federal shutdown in early 2019, has resulted in an increased interest on the part of scientists in finding other sources of funding. Understanding the effect of such different funding streams on research outputs is thus of more than academic importance, particularly because there are likely to be tradeoffs, both in terms of the structure of research and in terms of research outputs. For example, federal funding is often intended to affect the structure of research, with explicit goals of training the next generation of scientists and promoting diversity; those goals are less salient for non-federal funding. On the output side, federally funded research may be more likely to emphasize producing purely scientific outputs, like publications, rather than commercial outputs, like patents. The contribution of this paper is to use new data to examine how different sources of financial support – which we refer to as "braided" funding – affect both the structure of scientific research and the subsequent outputs.

Federal Funding of Doctoral Recipients: What Can Be Learned From Linked Data

Chang W, Cheng W, Lane JI, & Weinberg BA
Research Policy available online 14 March 2019
https://doi.org/10.1016/j.respol.2019.03.001

Abstract
This technical note describes the results of a pilot approach to link administrative and survey data to better describe the richness and complexity of the research enterprise. In particular, we demonstrate how multiple funding channels can be studied by bringing together two disparate datasets: UMETRICS, which is based on university payroll and financial records, and the Survey of Earned Doctorates (SED), which is one of the most important US survey datasets about the doctoral workforce. We show how it is possible to link data on research funding and the doctorally qualified workforce to describe how many individuals are supported in different disciplines and by different agencies. We outline the potential for more work as the UMETRICS data expands to incorporate more linkages and more access is provided.

Generating Automatically Labeled Data for Author Name Disambiguation: An Iterative Clustering Method.

Kim, Jinseok, Kim, Jinmo, & Owen-Smith, Jason
Scientometrics **118,** 253–280 (2019). https://doi.org/10.1007/s11192-018-2968-3

Abstract
Many author name disambiguation studies have relied on hand-labeled truth data that are very laborious to generate. This paper shows that labeled data can be automatically generated using information features such as email address, coauthor names, and cited references that are available from publication records. For this purpose, high-precision rules for matching name instances on each feature are learned using an external-authority database. Then, selected name instances in target ambiguous data go through the process of pairwise matching based on the learned rules. Next, they are merged into blocks by a generic entity resolution algorithm. The blocking procedure is repeated over other features until further merging is impossible. Tested on an example of 26,566 name instances, this iterative blocking produced accurately labeled data with near perfect accuracy (pairwise F1 = 0.99). In addition, the labeled data represented the population data of 227K name instances in terms of name ethnicity and co-disambiguating name group size distributions. Several challenges are discussed for applying this method to resolving author name ambiguity in large-scale scholarly data.

The Impact of Imbalanced Training Data on Machine Learning for Author Name Disambiguation

Kim, Jinseok & Kim, Jenna
Scientometrics **117,** 511–526 (2018). https://doi.org/10.1007/s11192-018-2865-9

Abstract

In supervised machine learning for author name disambiguation, negative training data are often dominantly larger than positive training data. This paper examines how the ratios of negative to positive training data can affect the performance of machine learning algorithms to disambiguate author names in bibliographic records. On multiple labeled datasets, three classifiers – Logistic Regression, Naïve Bayes, and Random Forest – are trained through representative features such as author name, coauthor names, and title words extracted from the same training data but with various positive-to-negative training data ratios. Results show that increasing negative training data can improve disambiguation performance but with a few percent of performance gains and sometimes degrade it. Logistic Regression and Naïve Bayes learn optimal disambiguation models even with a base ratio (1:1) of positive and negative training data. Also, the performance improvement by Random Forest tends to quickly saturate roughly after 1:10 ~ 1:20. These findings imply that contrary to the common practice using all training data, name disambiguation algorithms can be trained using part of negative training data without degrading much disambiguation performance while increasing computational efficiency. This study calls for more attention from author name disambiguation scholars to methods for machine learning from imbalanced data.

Evaluating Author Name Disambiguation for Digital Libraries: A Case of DBLP

Kim, Jinseok
Scientometrics **116,** 1867–1886 (2018). https://doi.org/10.1007/s11192-018-2824-5.

Abstract

Equipped with advanced computing techniques, scholars have disambiguated author names in whole digital libraries and tested their performances in various ways. The purpose of this study is to propose a triangulation approach that author name disambiguation for digital libraries can be better evaluated when its performance is assessed on multiple labeled datasets with comparison to baselines for diverse ambiguity dimensions. To illustrate the proposed approach, accuracy of author name disambiguation in DBLP's 3.7M records is evaluated on three types of labeled data containing 5,000 to 6M disambiguated names. Results show that the triangulation method can provide a more holistic, granulated understanding of a disambiguation method's performance than common evaluation practices in prior studies. With the review of strengths and weaknesses of the proposed approach, this study calls for further discussion about consistent frameworks and methodologies for evaluating author name disambiguation so that findings from a variety of studies can be synthesized to produce insights for improving name ambiguity resolution for fast-growing digital libraries.

Proximity and Economic Activity: An Analysis of Vendor-University Transactions

Goldschlag N, Lane JI, Weinberg BA , Zolas, N
***Journal of Regional Science,* 2018: 1-20**

DOI: 10.111/jors.12397 https://onlinelibrary.wiley.com/doi/abs/10.1111/jors.12397

Abstract
This paper uses transaction-based data to provide new insights into the link between the geographic proximity of businesses and associated economic activity. It develops two new measures of, and a set of stylized facts about, the distances between observed transactions between customers and vendors for a research intensive sector. Spending on research inputs is more likely with businesses physically closer to universities than those further away. Firms supplying a university project in one year are more likely to subsequently open an establishment near that university. Vendors who have supplied a project, are subsequently more likely to be a vendor on the same or related project.

Why the U.S. Science and Engineering Workforce is Aging Rapidly

Blau D, & Weinberg BA

***Proceedings of the National Academy of Sciences*, 14 February 2017**
Vol. 114(15), 3879-3884 DOI: 10.1073/pnas.16117481114/-/DCSupplemental
http://www.pnas.org/content/114/15/3879.short

Abstract
The science and engineering workforce has aged rapidly in recent years, both in absolute terms and relative to the workforce as a whole. This is a potential concern if the larger number of older scientists crowds out younger scientists, making it difficult for them to establish independent careers. In addition, scientists are believed to be most creative earlier in their careers, so the aging of the workforce may slow the pace of scientific progress. The authors developed and simulated a demographic model, which shows that a substantial majority of recent aging is a result of the aging of the large baby boom cohort of scientists. However, changes in behavior have also played a significant role, in particular a decline in the retirement rate of older scientists, induced in part by the elimination of mandatory retirement in universities in 1994. Furthermore, the age distribution of the scientific workforce is still adjusting. Current retirement rates and other determinants of employment in science imply a steady-state mean age 2.3 years higher than the 2008 level of 48.6.

STEM Training and Early Career Outcomes of Female and Male Graduate Students: Evidence from UMETRICS Data Linked to the 2010 Census

Buffington C, Cerf B, Jones C, & Weinberg BA

***American Economic Review* May 2016**

106(5), pp. 333–338 DOI: 10.1257/aer.p20161124
https://www.aeaweb.org/articles?id=10.1257/aer.p20161124

Abstract
Women are underrepresented in science and engineering, with the underrepresentation increasing in career stage. We analyze gender differences at critical junctures in the STEM pathway–graduate training and the early career–using UMETRICS administrative data matched to the 2010 Census and W-2s. We find strong gender separation in teams, although the effects of this are ambiguous. While no clear disadvantages exist in training environments, women earn 10% less than men once we include a wide range of controls, most notably field of study. This gap disappears once we control for women's marital status and presence of children.

Wrapping It Up in a Person: Examining Employment and Earnings Outcomes for Ph.D. Recipients

Zolas N, Goldschlag N, Jarmin RS, Stephan P, Owen- Smith J, Rosen RF, McFadden Allen B, Weinberg BA, & Lane JI

***Science* 11 December 2015**
Vol. 350(6266), *pp. 1367-1371*
DOI: 10.1126/science.aac5949
http://www.sciencemag.org/content/350/6266/1367.full
Supplementary material: http://econ.ohio-state.edu/weinberg/Science-aac5949_Zolas-SM-PUBLISHED.pdf

Abstract
In evaluating research investments, it is important to establish whether the expertise gained by researchers in conducting their projects propagates into the broader economy. For eight universities, it was possible to combine data from the UMETRICS project, which provided administrative records on graduate students supported by funded research, with data from the U.S. Census Bureau. The analysis covers 2010–2012 earnings and placement outcomes of people receiving doctorates in 2009–2011. Almost 40% of supported doctorate recipients, both federally and nonfederally funded, entered industry and, when they did, they disproportionately got jobs at large and high-wage establishments in high-tech and professional service industries. Although Ph.D. recipients spread nationally, there was also geographic clustering in employment near the universities that trained and employed the researchers. We also show large differences across fields in placement outcomes.

New Linked Data on Research Investments: Scientific Workforce, Productivity, and Public Value

Lane JI, Owen-Smith J, Rosen RF, & Weinberg BA

Research Policy **December 2014**
Vol. 44(9), pp. 1659-1671 DOI: 10.1016/j.respol.2014.12.013
http://www.sciencedirect.com/science/article/pii/S0048733315000025

Abstract
Longitudinal micro-data derived from transaction level information about wage and vendor payments made by Federal grants on multiple US campuses are being developed in a partnership involving researchers, university administrators, representatives of Federal agencies, and others. This paper describes the UMETRICS data initiative that has been implemented under the auspices of the Committee on Institutional Cooperation. The resulting data set reflects an emerging conceptual framework for analyzing the process, products, and impact of research. It grows from and engages the work of a diverse and vibrant community. This paper situates the UMETRICS effort in the context of research evaluation and ongoing data infrastructure efforts in order to highlight its novel and valuable features. Refocusing data construction in this field around individuals, networks, and teams offers dramatic possibilities for data linkage, the evaluation of research investments, and the development of rigorous conceptual and empirical models. Two preliminary analyses of the scientific workforce and network approaches to characterizing scientific teams ground a discussion of future directions and a call for increased community engagement.

Science Funding and Short-Term Economic Activity

Weinberg BA, Owen-Smith J, Rosen RF, Schwarz L, McFadden Allen B, Weiss RE, & Lane JI

Science **4 April 2014**
Vol. 344(6179), pp. 41-43 DOI: 10.1126/science.1250055
http://www.sciencemag.org/content/344/6179/41.full

Abstract
There is considerable interest among policy-makers in documenting short-term effects of science funding. A multiyear scientific journey that leads to long-term fruits of research, such as a moon landing, is more tangible if there is visible nearer-term activity, such as the presence of astronauts. Yet systematic data on such activities have not heretofore existed. The only source of information for describing the production of most science is surveys that have been called “a rough estimate, frequently based on unexamined assumptions that originated years earlier.

COMMENTARIES

- *Universities are being "short sighted" when chasing partnerships with companies like Amazon* (Michigan Radio's Stateside program March 4, 2019) Jason Owen-Smith
- *Amazon pullout from NYC shows the perils of partnerships between higher education and business* (The Conversation February 26, 2019) Jason-Owen-Smith
- *Building an infrastructure to support the use of government administrative data for program performance and social science research* (ANNALS, AAPSS, Vol 675, Issue 1, January 2018) Julia Lane
- *A roadmap to a nationwide data infrastructure for evidence-based policy making* (ANNALS, AAPSS, Vol 675, Issue 1, January 2018) Andrew Reamer and Julia Lane
- *Tax bill would imperil nation's innovation, future* (Columbus Dispatch December 14, 2017) Bruce Weinberg, Jason Owen-Smith, and Julia Lane
- *Watching the players, not the scoreboard* (Nature: Comment November 2, 2017) Julia Lane
- *The social sciences need to build new foundations* (Significance Magazine June 9, 2017) Julia Lane
- *A call to action to build social science data infrastructure* (Nature Human Behaviour April 7, 2017) Julia Lane
- *Who Feels the Pain of Science Research Budget Cuts?* (The Conversation/Salon March 29, 2017) Bruce Weinberg
- *Fix Incentives* (Nature: Perspective September 1, 2016) Julia Lane

FORTHCOMING

Linking in a Big Data World

Chang W, Emad A, Lane JI, Tokle J, & Weinberg BA

***Science Policy Forum* submission**

Abstract

The increased availability of new types of data means that both social science researchers and statistical agencies are interested in finding new ways of linking survey and other datasets together. A major challenge, however, is that unique identifiers are typically not present, nor are gold standard datasets. This paper describes the process of linking a well curated survey, the Survey of Earned Doctorates to a new source of transaction data which featured all of these challenges. We were able to show that machine learning approaches can successfully be used to improve link quality, and that great use can be made of data that are available in a subset of files. We also extended the literature that uses output measures - analytical utility – in addition to input measures – match quality – as metrics of the success of the linkage effort.

Research Experience as Human Capital in New Business Outcomes

Lane JI, Jarmin RS, Goldschlag N, & Zolas N

American Economic Association Meeting, Philadelphia, January 2018

Session: New Measures of Human Capital and Their Application (moderated by Bruce Weinberg)

Forthcoming in CRIW volume on *The Measurement and Diffusion of Innovation*

Abstract

Human capital is typically cited as an important contributor to the survival, growth and innovative activity of new businesses. This paper contributes to the literature by both developing novel measures of human capital and examining the link between those measures and the outcomes of young firms. It builds on several strands of the literature which emphasize the importance of employee workplace experience as a dimension of human capital. It shows that the effects of work experience differ substantially by where an employee worked and is valued differently by firms in different sectors. This is particularly true for research experience, which is consistent with the notion that on the job training in complex tasks should be valuable to firms with complex production technologies. This paper will be included as a book chapter in the NBER CRIW volume on *The Measurement and Diffusion of Innovation* (Carol Corrado, Javier Miranda, Jonathan Haskel, and Daniel Sichel, eds., University of Chicago Press, forthcoming)

Additional work in development: IRIS Co-PIs and their teams have a variety of papers in process. As of this writing, the following manuscripts are in development:

- IRIS Co-PIs Lane and Weinberg are drafting a manuscript on occupational classification using administrative data, which will be a CES working paper.

- IRIS Co-PIs Lane and Weinberg are drafting a manuscript on the stay rates of foreign-born researchers that is currently being finalized.
- With postdoctoral scholar Valerie Bostwick, IRIS Co-PIs Weinberg and Lane have exploratory results for an extension of the Buffington et al. paper studying gender differences in early career outcomes of STEM Ph.D. recipients.
- A white paper is being written for USPTO by IRIS Co-PI Lane and her team at New York University, and should be completed shortly.
- Co-PI Weinberg and postdoctoral scholar Reza Sattari are working on a paper looking at public research funding and scientific productivity; preliminary results show increases in funding yield additional publications.
- Co-PI Lane, together with colleagues Matt Ross and Ridhima Sodhi at NYU are drafting a manuscript on the relationship between the gender of NIH and NSF staff are related to the gender of researchers supported by NIH and NSF funding.
- Akina Ikudo and Matt Ross are working with Co-PIs Lane and Weinberg on a manuscript focusing on the economic spillovers from science.
- With research investigator Jake Fisher, Co-PI Owen-Smith has drafted a working paper examining the network structure of collaboration across 26 IRIS campuses. That paper, entitled "How Universities Organize Their Science" was presented at the National Academy of Sciences Arthur M. Sackler Colloquium in December 2018. It has been submitted for inclusion in a special issue of *Proceedings of the National Academies of Science* and is under review.

WORKING PAPERS

Occupational Classifications: A Machine Learning Approach

Akina Ikudo, Julia Lane, Joseph Staudt, Bruce Weinberg
NBER Working Paper No. 24951 Issued in August 2018 https://www.nber.org/papers/w24951

Abstract

Characterizing the work that people do on their jobs is a longstanding and core issue in labor economics. Traditionally, classification has been done manually. If it were possible to combine new computational tools and administrative wage records to generate an automated crosswalk between job titles and occupations, millions of dollars could be saved in labor costs, data processing could be sped up, data could become more consistent, and it might be possible to generate, without a lag, current information about the changing occupational composition of the labor market. This paper examines the potential to assign occupations to job titles

contained in administrative data using automated, machine-learning approaches. We use a new extraordinarily rich and detailed set of data on transactional HR records of large firms (universities) in a relatively narrowly defined industry (public institutions of higher education) to identify the potential for machine-learning approaches to classify occupations.

Local Fiscal Multiplier on R&D and Science Spending: Evidence from the American Recovery and Reinvestment Act

Chhabra, Yulia and Levenstein, Margaret C. and Owen-Smith, Jason
Ross School of Business Paper No. 1383
[1]SSRN: https://ssrn.com/abstract=3201136 and http://hdl.handle.net/2027.42/144514. Under review at *American Economic Journal: Economic Policy*.

Abstract

We use the American Recovery and Reinvestment Act (ARRA), a large stimulus package passed into law to combat the Great Recession, to estimate the effect of R&D and science spending on local employment. Unlike most fiscal stimuli, the R&D and science portion of ARRA did not target counties with poor economic conditions but rather was awarded following a peer review process, or based on innovative potential and research infrastructure. We find that, over the program's five-year disbursement period, each one million USD in R&D and science spending was associated with twenty-seven additional jobs. The estimated job-year cost is about $15,000.

How Universities Organize Science

Fisher, Jacob C. & Owen-Smith, Jason
Submitted for publication (2018)

Abstract

Although the results of science -- publications and patents -- have received considerable attention, little work to date has considered how the production of science is organized. Using a unique dataset on grant payments to faculty, staff, and trainees within 23 universities, we explore how universities approach a similar task, developing research findings, in different ways. Drawing on organizational theory that suggests that work is accomplished through a network of collaborations, we examine two complementary processes that cause the organization of science to differ between universities. First, administrators and grantors can control the number and occupation of people involved in the network. Second, individual faculty members can control the specific collaboration relationships, both between faculty members, and among staff and trainees who receive funding from particular grants. In network terms, administrators and grantors control the vertices, and individual faculty control the edges

[1]

between them. We find that the influence of administrators and grantors is most visible along two dimensions: the amount of funding awarded by NIH, and the ratio of trainees to staff. A cluster analysis demonstrates that individual faculty staff grants in one of six ways, which depend on the scale of the grant and the faculty member's preferences. We find that between-university differences in the connectivity of the network can largely be explained by differences in scale, differences in clustering can be explained by faculty preferences, but overall differences in structure of the networks cannot be well-explained by either scale or collaboration preferences.

The link between R&D, human capital and business startups

Goldschlag N, Jarmin RS, Lane JI, & Zolas N

Presented at American Economic Association Meeting, Chicago, January 2017

Session: Using Data Science to Examine the Link Between University R&D and Innovation (moderated by Julia Lane)

NGER CRIW-The Measurement and Diffusion of Innovation (Corrado C, Sichel D, and Miranda J, eds)

Abstract

The reason for the secular decline in entrepreneurship is not well understood. It is evident in all sectors of the economy and almost all regions. One approach to stimulating innovation and entrepreneurship has been to increase investments in science: the U.S. federal government contributed nearly $38 billion for university-based research in Science, Technology, Engineering, and Mathematics (STEM) in 2014. However, there has historically been little evidence about the links between investments in university research and innovation - largely because surveys cannot capture the complex ways in which scientific ideas are created, transmitted and adopted.

This paper examines the relationship between the funding of research teams - in terms of structure, field and type of funding - and the subsequent propensity of members of those teams to start up businesses. It also examines the subsequent survival and productivity growth of those startups.

The work is now possible because of a new data infrastructure resulting from collaborations between the Census Bureau's Innovation Measurement Initiative, the National Science Foundation and the Institute for Research on Innovation and Science at the University of Michigan. The infrastructure links universe data on all people employed on research grants, their funding, and their economic and scientific activities.

This paper is the first to directly trace the pathways from the bench to the workplace at a large scale, using universe data from 25 universities covering about 25% of federal university based

R&D. It is the first to use universe data on workers (the LEHD data) to draw comparison groups of individuals employed both within the university and from other R&D intensive businesses. And it is the first to use universe data on business startups to compare the dynamics of university sourced entrepreneurship with other types of entrepreneurship.

Pathways to Production

Barth E, Davis J, Marschke G, Wang A, Zhou S

Presented at American Economic Association Meeting, Chicago, January 2017
Session: Using Data Science to Examine the Link Between University R&D and Innovation (moderated by Julia Lane)

Abstract

Science funding agencies often require researchers to demonstrate their project's prospects for "development of a diverse, globally competitive STEM workforce," "increase[d] partnerships between academia, industry and others" (NSF, 2016), and other goals beyond the creation of scientific knowledge. This paper attempts to measure these wider impacts of scientific research.

We use the newly created Census data infrastructure that links university grant transaction data to Longitudinal Employer-Household Dynamics (LEHD) data to map employment linkages between universities and industry. First, we ask, what are the flow rates of new STEM workers—post-docs and recent doctorates—into research intensive firms, industries, and regions? startups and established firms? high- and low- productivity firms? local and out-of-state employment?

Second, we estimate the impacts of and returns to university-based human capital accumulation by STEM workers. The sudden increase in science funding under the American Recovery and Reinvestment Act of 2009 (ARRA) increased demand in the academic sector for post-graduate researchers, both lengthening existing post-graduate research engagements in universities, and increasing the likelihood that recent graduates, especially doctorates, obtain post-graduate employment in universities. We estimate the impact of increased university-based research training on career paths, including the likelihood of obtaining a faculty post, and for researchers who enter industry, which firms they match to, and their wage outcomes.
Third, we investigate the extent to which a firm placement depends on the history of previous placements from the same university. Such a correlation could be evidence of "hiring chains", or of specific knowledge links between the research and teaching at a specific university and the production technology of particular firms. The hiring patterns we uncover between universities and industry reveal important features of the labor market for specialized skills, and

increase our understanding of how university research contributes to the diffusion of new ideas in the economy.

Financial Advice and the Entrepreneurial Spillovers of Basic Research

Dacunto F, & Yang L

Presented at American Economic Association Meeting, Chicago, January 2017

Session: Using Data Science to Examine the Link Between University R&D and Innovation (moderated by Julia Lane)

Abstract

We test for the effect of informal financial advice on the establishment and subsequent performance of entrepreneurial ventures that commercialize the results of basic research. To this aim, we construct a unique data set that includes: (i) the characteristics of the faculty recipients of federally-funded grants across 10 large U.S. universities, which produce innovation that can be commercialized through the establishment of startups; (ii) the likelihood that recipients establish a non-employer venture (iLBD) or an employer venture (LEHD), as well as the job growth characteristics of these ventures; and (iii) the network of neighbors in the locations where the recipients' reside, including the occupation titles and demographics of the neighbors (ACS/Decennial Census). We use the presence of financial-sector employees among the faculty's network members (spouses or neighbors) to test for the effect. We compare faculty grant recipients in similar areas of research, obtaining grants of similar sizes in the same rounds of funding, and at similar stages of their academic careers, but belonging to networks with different levels of exposure to informal financial advice from family and friends. Financial advice from one's social network is informal because advisers are not paid fees for providing their service. Therefore, the paper broadly tests for whether advice is a positive externality of one's social networks, which is valuable to the individual entrepreneurs as well as to economic growth.

Research Funding and Subsequent Entrepreneurship: The Role of Underrepresentation

Buffington C, Harris B, Feng F, & Weinberg BA

Presented at American Economic Association Meeting, Chicago, January 2017

Session: Using Data Science to Examine the Link Between University R&D and Innovation (moderated by Julia Lane)

Abstract

Federal funding affects both who does research, and the environment in which research is done. In a recent study, 6 in 10 female doctoral recipients had been supported by federal research funds, compared to 7 in 10 male doctoral recipients. Federal funding also appears to be highly correlated with the pipeline of researchers going into different fields; particularly into R&D fields and the decision to pursue postdoctoral fellowships.

This paper uses rich new Census Bureau data linked to detailed information on the individuals supported by research funding to examine the effect of both the type and structure of federal funding on the outcomes of underrepresented students. It makes use of rich measures on student characteristics, including their race, gender, place of birth, marital status and presence of children. It constructs new network theoretic measures of team environment, based on the characteristics of all individuals working together on research grants. It also includes information about household and family structure in the model. It also examines two types of outcome measures - placement in R&D performing, high technology or young and small firms - as well as the propensity of underrepresented groups to start up businesses.

Nevertheless She Persisted? Gender Peer Effects in Doctoral Stem Programs

Bostwick, Valerie K. and Weinberg, Bruce A.

NBER Working Paper No. 25028, September 2018 http://www.nber.org/papers/w25028

Abstract

We study the effects of peer gender composition, a proxy for female-friendliness of environment, in STEM doctoral programs on persistence and degree completion. Leveraging unique new data and quasi-random variation in gender composition across cohorts within programs, we show that women entering cohorts with no female peers are 11.9pp less likely to graduate within 6 years than their male counterparts. A 1 sd increase in the percentage of female students differentially increases the probability of on-time graduation for women by 4.6pp. These gender peer effects function primarily through changes in the probability of dropping out in the first year of a Ph.D. program and are largest in programs that are typically male-dominated.

BOOKS & BOOK CHAPTERS

Research Universities and the Public Good: Discovery for an Uncertain Future

Owen-Smith, Jason

Stanford University Press, September 2018

Abstract

In a political climate that is skeptical of hard-to-measure outcomes, public funding for research universities is under threat. But if we scale back support for these institutions, we also cut off a key source of value creation in our economy and society. *Research Universities and the Public Good* offers a unique view of how universities work, what their purpose is, and why they are important.

Countering recent arguments that we should "unbundle" or "disrupt" higher education, Jason Owen-Smith argues that research universities are valuable gems that deserve support. While they are complex and costly, their enduring value is threefold: they simultaneously act as *sources* of new knowledge, *anchors* for regional and national communities, and *hubs* that connect disparate parts of society. These distinctive features allow them, more than any other institution, to innovate in response to new problems and opportunities. Presenting numerous case studies that show how research universities play these three roles and why they matter, this book offers a fresh and stirring defense of the research university.

Measuring the Economic Value of Research: The Case of Food Safety

Husbands Fealing K, Lane JI, King J, & Johnson SR (eds.)

***Cambridge University Press*, 2017**

ISBN 9781316671788

Abstract

This innovative study demonstrates new methods and tools to trace the impact of federal research funding on the structure of research, and the subsequent economic activities of funded researchers. The case study is food safety research, which is critical to avoiding outbreaks of disease. The authors make use of an extraordinary new data infrastructure and apply new techniques in text analysis. Focusing on the impact of U.S. federal food safety research, this book develops vital data-intensive methodologies that have a real world application to many other scientific fields.

Chapter 8 Networks: The Basics

Owen-Smith J, in *Big Data and Social Science* pp. 215-240

Foster I, Ghani R, Jarmin RS, Kreuter F, & Lane JI (eds.)

***Chapman and Hall/CRC*, August 9, 2016**
ISBN 9781498751407

Abstract
Noted sociologist and network theorist Jason Owen-Smith provides a primer on network theory, including details on network measures and components.

Big Data and Social Science: A Practical Guide to Methods and Tools

Foster I, Ghani R, Jarmin RS, Kreuter F, & Lane JI (eds.)

***Chapman and Hall/CRC*, August 9, 2016**
ISBN 9781498751407

Abstract
Big Data and Social Science: A Practical Guide to Methods and Tools shows how to apply data science to real-world problems in both research and the practice. The book provides practical guidance on combining methods and tools from computer science, statistics, and social science. This concrete approach is illustrated throughout using an important national problem, the quantitative study of innovation. The text draws on the expertise of prominent leaders in statistics, the social sciences, data science, and computer science to teach students how to use modern social science research principles as well as the best analytical and computational tools. It uses a real-world challenge to introduce how these tools are used to identify and capture appropriate data, apply data science models and tools to that data, and recognize and respond to data errors and limitations.

PRESS

- *Biologists lose out in post-PhD earnings analysis* (Nature: News December 10, 2015)
- *Where new PhD grads find work — and who earns the most with their degree* (Washington Post December 10, 2015)
- *PhDs pay: study reveals economic benefit of funding doctorates* (Times Higher Education December 10, 2015)
- *Science and math PhDs earn about $65,000 — more than double what arts majors do* (Vox December 11, 2015)

- *ProQuest Dissertation Database Provides Critical Information for Research Projects Across the US* (PR Newswire March 22, 2016)
- *Assessment: Academic return* (Nature May 4, 2016)
- *There's a huge gender pay gap for STEM careers — just one year after graduation* (Vox May 11, 2016)
- *Facing Skepticism, colleges set out to prove their value* (PBS Newshour January 22, 2016)
- *The Price of Doing a Postdoc* (Science: Share January 10, 2017)
- *Trump Administration Proposes Big Cuts in Medical Research* (NPR Health Shots March 16, 2017)
- *Communicating the Value of University Research When Science is Under Attack* (Inside Higher Ed April 6, 2017)
- *The Looming Decline of the Public Research University* (Washington Monthly September/October 2017)
- *One big reason why women drop out of doctoral STEM programs* (Ohio State News September 17, 2018
- *One Big Reason Why Women Drop Out of Doctoral STEM Programs* (Communication of the ACM September 17, 2018)
- *An insidious reason women are less likely to get a STEM doctoral degree than men* (Moneyish September 17, 2018)
- *'Nevertheless She Persisted?'* (Inside Higher Ed September 18, 2018)
- *Women In Stem Benefit From Same-Sex Support* (Pacific Standard September 19, 2018)
- *Gender imbalance affects degrees* (Science News at a glance September 28, 2018)
- *When you're the only woman: The challenges for female Ph.D. students in male-dominated cohorts* (Science October 24, 2018)