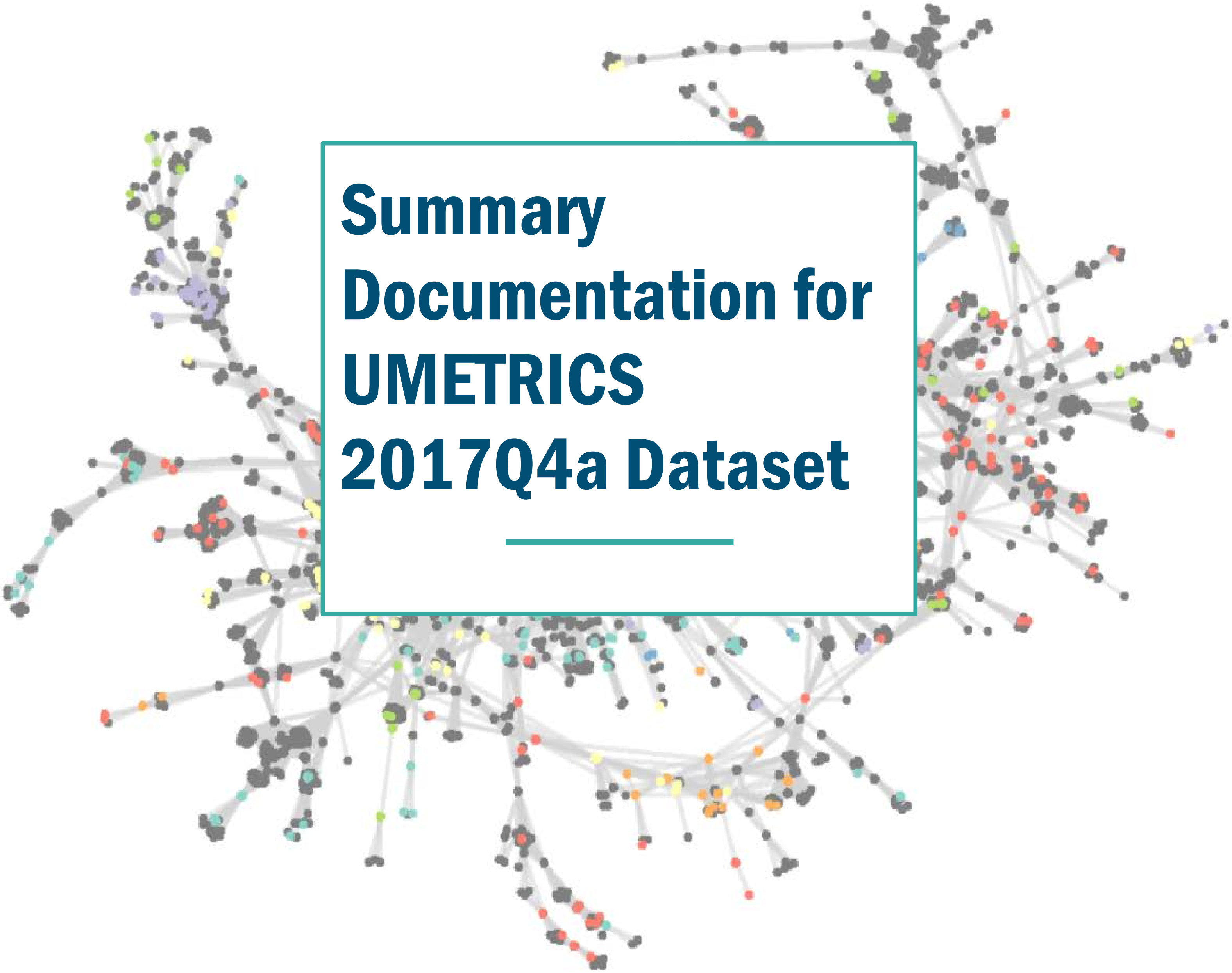
Summary Documentation for UMETRICS 2017Q4a Dataset



April 2018

IRIS Research Team

IRIS
INSTITUTE FOR
RESEARCH ON
INNOVATION & SCIENCE

Data Access Statement

This is the publicly available summary documentation for the UMETRICS 2017Q4a dataset.
Access to the full documentation is restricted to authorized IRIS data users.

Contents

Tables and Figures

Project Overview

Principal Investigators

Ron Jarmin (U.S. Census Bureau)

Julia Lane (New York University)

Jason Owen-Smith (University of Michigan)

Bruce Weinberg (The Ohio State University / National Bureau of Economic Research)

Funding

IRIS is funded primarily by the Ewing Marion Kauffman and Alfred P. Sloan foundations and by IRIS member universities.

Human Subjects Oversight

The IRIS repository received an initial approval determination from the University of Michigan's IRB-Health Sciences and Behavioral Sciences on 24 March 2015. Since then, three continuing review applications have been reviewed and approved by the IRB. The most recent approval is valid through January 2019.

Keywords / Subject Terms

Administrative data, award activity, awards and funding, award expense transaction, collaboration networks, economic and social value of research, graduate students and postdoctoral researchers, research activity, research impact, science of innovation, scientific productivity, and scientific workforce

About This Release

This data release (UMETRICS 2017Q4a dataset) is the second annual IRIS research data release. The first research data release (UMETRICS 2016Q3a dataset) was made available to approved researchers in the IRIS Virtual Data Enclave in March 2017, and a mirror of the dataset with linkages to restricted Census Bureau data was released via the Federal Statistical Research Data Center (FSRDC) system[1] in May 2017.

The documentation for this data release focuses on both descriptions of IRIS data and our process and methodology for record linkage. The dataset includes de-identified IRIS data, public elements of external datasets (e.g., grants and publications), and crosswalk tables to match particular data elements (e.g., awards, awardees, research employees) across IRIS data and external datasets. As with the first release, this research data release is being integrated with U.S. Census Bureau data and will be available in the FSRDC system in May 2018.

Data Access Notes

Data files are available in two environments. First, approved users may access the dataset by logging in to the IRIS Virtual Data Enclave (VDE or enclave hereafter). No downloadable or otherwise publicly accessible data are available outside of the enclave. Access to the IRIS dataset through the enclave is dependent upon approval of a research proposal and receipt of an IRB determination letter and IRIS Data Use Agreement. For more information about access to the IRIS VDE, contact irisdatarequests@umich.edu. The Data Access Application Form and Data Use Agreement are available on the IRIS website.[2]

Second, a copy of the IRIS dataset with additional crosswalks to restricted U.S. Census Bureau data resources is available through the FSRDC system for those researchers with Special Sworn Status. Regarding crosswalks to Census data, the employee and vendor/subaward information was matched to Census Bureau records to enable researchers to utilize Census records to further estimate the economic impacts of research funding. Detailed discussion of the procedures and crosswalks created for that process are available in the UMETRICS RDC sub-folder. Please contact your local RDC administrator for any questions on access to the FSRDC system.

[1] https://www.census.gov/fsrdc

[2] http://iris.isr.umich.edu/research-data/access/

Methodology

Data Contributor(s)

- IRIS member universities
- Data manager(s), curator(s), and distributor(s):
  - IRIS has served as a data manager, curator, and distributor in this round of data release
  - IRIS has prepared the UMETRICS 2017Q4a dataset for distribution for research purposes

Mode of Data Collection

IRIS Principal Investigators (PIs) were not involved in primary data collection, which was initially carried out by each IRIS member university as part of its administrative functions. Twenty-six (26) member universities are represented in the UMETRICS2017Q4a dataset:

- Boston University
- Emory University
- Indiana University
- Michigan State University
- New York University
- Northwestern University
- Ohio State University
- Pennsylvania State University
- Princeton University
- Purdue University
- Rutgers University
- Stony Brook University
- University of Arizona
- University of California - San Diego
- University of Cincinnati
- University of Colorado, Boulder

- University of Hawaii[3]
- University of Illinois at Urbana-Champaign
- University of Iowa
- University of Kansas
- University of Michigan
- University of Missouri
- University of Pennsylvania
- University of Pittsburgh
- University of Virginia
- University of Wisconsin - Madison

Units of Observation

- Each expense transaction is a mix of quarterly, monthly, and daily records.
- The pay / transaction period varies across files and universities.
- Most of the transaction records in the Award File are monthly, whereas the majority of records in the Vendor File are daily (see 2018 Dataset Descriptions for details).

Data Processing

IRIS seeks to add value to the data received by member universities by identifying and resolving data discrepancies when possible, providing standardized occupational classification codes, 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 member universities or individuals from particular data elements. As part of preparing the dataset for research use, IRIS data processing methods have included but are not limited to:

1) Removal of university names and campus location information of IRIS member universities;

[3] The University of Hawaii terminated its IRIS membership in October 2017. Although we continue to include its data in the current and future release datasets, no new data are added to the collection from the University of Hawaii.

2) Removal of any personally identifiable information (e.g., any individual names, personal employee identification numbers, and EINs if vendors, contractors, or subrecipients are individuals);

3) Replacement of any university-submitted identification numbers with randomly assigned numbers for a new set of IDs;

4) Replacement of campus-level vendor and subaward recipient's identification numbers with randomly assigned unique identification numbers that help to disambiguate them at the national level;

5) Generating and assigning occupation classification to all personnel paid by grants;

6) Cleaning and standardizing names of vendors and subaward recipients; and,

7) Linking records using a variety of algorithms.

IRIS Research Data Users

A primary goal of the IRIS research data release is to enable the research community to access and use this dataset, subject to responsible privacy and confidentiality restrictions. IRIS encourages researchers from all disciplines to apply for approval to access IRIS data. As of this data release, 40 research users spanning 15 institutions have active IRIS VDE accounts, including 24 faculty members, research scientists and/or data analysts, four post-doctoral researchers or fellows, and 12 graduate or undergraduate students. Recent IRIS data use cases include studies to determine how research experience shapes the career pathways of students; to examine how federally funded research yields safer and more secure food systems; to analyze gender differences in graduate studies and early career pathways within STEM fields; to explore the way scientific knowledge is translated into society by the public service activities of faculty; and to measure how university vendors produce additional innovations and contribute to regional growth.

Beginning in March 2018, the first cohort of researchers who were awarded the IRIS Researcher Awards (six awardees and their collaborators) will pursue research projects using the IRIS dataset. This first cohort includes three graduate students representing the University of Pennsylvania and the University of Wisconsin-Madison, two early career faculty members representing the George Washington University and North Carolina State University, and one established researcher representing Michigan State University. IRIS Researcher Award recipients will have access to this current release dataset for a one-year term for their research projects. In addition to our first cohort (2017-2018), IRIS plans to offer two more rounds of awards (2018-2019 and 2019-2020).

Selected Publications Based on UMETRICS Data

A list including the following papers, working papers, and books as well as additional IRIS publications is updated regularly at: http://iris.isr.umich.edu/research-data/publications/

Blau, D. M., & Weinberg, B. A. (2017). Why the US science and engineering workforce is aging rapidly. *Proceedings of the National Academy of Sciences*, 114(15), 3879-3884. doi:10.1073/pnas.1611748114

Buffington, C., Cerf, B., Jones, C., & Weinberg, B. A. (2016). STEM Training and Early Career Outcomes of Female and Male Graduate Students: Evidence from UMETRICS Data Linked to the 2010 Census. *American Economic Review*, 106(5), 333-338. doi:10.1257/aer.p20161124

Buffington, C., Harris, B., Feng, F., & Weinberg, B. A. (2017, January 6). *Research Funding and Subsequent Entrepreneurship: The Role of Underrepresentation.* Paper presented at the American Economic Association Meeting, Chicago, IL.

Chang, W., Cheng, W., Jones, C., Lane, J.I., & Weinberg, B. A. (2018). *Federal Funding of Doctoral Recipients: Results from New Linked Survey and Transaction Data.* Submitted for publication.

Chang, W., Emad, A., Lane, J. I., Tokle, J., & Weinberg, B. A. (2018). *Linking in a Big Data World.* Submitted for publication.

D'Acunto, F., & Yang, L. (2017, January 6) *Financial Advice and the Entrepreneurial Spillovers of Basic Research.* Paper presented at the American Economic Association Meeting, Chicago, IL.

Foster, I. T., Ghani, R., Jarmin, R. S., Kreuter, F., & Lane, J. I. (Eds.). (2016). *Big Data and Social Science: A Practical Guide to Methods and Tools*. Boca Raton, FL: Chapman & Hall/CRC Press.

Fisher, J.C. & Owen-Smith, J. (2018). *How Universities Organize Science*. Submitted for publication.

Goldschlag, N., Bianchini, S., Lane, J. I., Sanmartín Sola J., & Weinberg, B. A. (2017). *Research Funding and Regional Economies.* (Working Paper No. 23018). National Bureau of Economic Research. doi: 10.3386/w23018

Goldschlag, N., Jarmin, R.S., Lane, J. I., & Zolas, N. (2017, January 6). *The Link between R&D, Human Capital and Business Startups.* Paper presented at the American Economic Association Meeting, Chicago, IL.

Husbands Fealing K., Lane, J. I., King, J., & Johnson, S. R., eds. (2017). *Measuring the Economic* Value of Research: The Case of Food Safety. Cambridge, UK: Cambridge University Press.

Kim, J. (2018). Evaluating Author Name Disambiguation for Digital Libraries: A Triangulation Approach. Submitted for publication.

Kim, J. & Kim, J. (2018). The Impact of Imbalanced Training Data on Machine Learning for Author Name Disambiguation. Submitted for publication.

Kim, J. & Owen-Smith, J. (2018). Automatic Generation of Labeled Data for Author Name Disambiguation: An Iterative Blocking Method. Submitted for publication.

Lane, J., Owen-Smith, J., Rosen, R., & Weinberg, B. A. (2014). New Linked Data on Research Investments: Scientific Workforce, Productivity, and Public Value. *Research Policy*, 44(9), 1659-1671. doi:10.3386/w20683

Lane, J. I., Jarmin, R.S., Goldschlag, N., & Zolas, N. (2018, January 5). *Research Experience as Human Capital in New Business Outcomes*. Paper presented at the American Economic Association Meeting, Philadelphia, PA.

Weinberg, B. A., Owen-Smith, J., Rosen, R. F., Schwarz, L., Allen, B. M., Weiss, R. E., & Lane, J. (2014). Science Funding and Short-Term Economic Activity. *Science*, 344(6179), 41-43. doi:10.1126/science.1250055

Zolas, N., Goldschlag, N., Jarmin, R., Stephan, P., Owen-Smith, J., Rosen, R. F., McFadden Allen, B., Weinberg, B.A., & Lane, J. I. (2015). Wrapping it up in a person: Examining employment and earnings outcomes for Ph.D. recipients. *Science*, 350(6266), 1367-1371. doi:10.1126/science.aac5949

2018 Dataset Overview

Citation

The Institute for Research on Innovation & Science (IRIS). Summary Documentation for UMETRICS 2017Q4a Dataset. Ann Arbor, MI: IRIS [distributor], 2018-04-10, DOI: 10.21987/R7FT08[4]

Dataset Summary

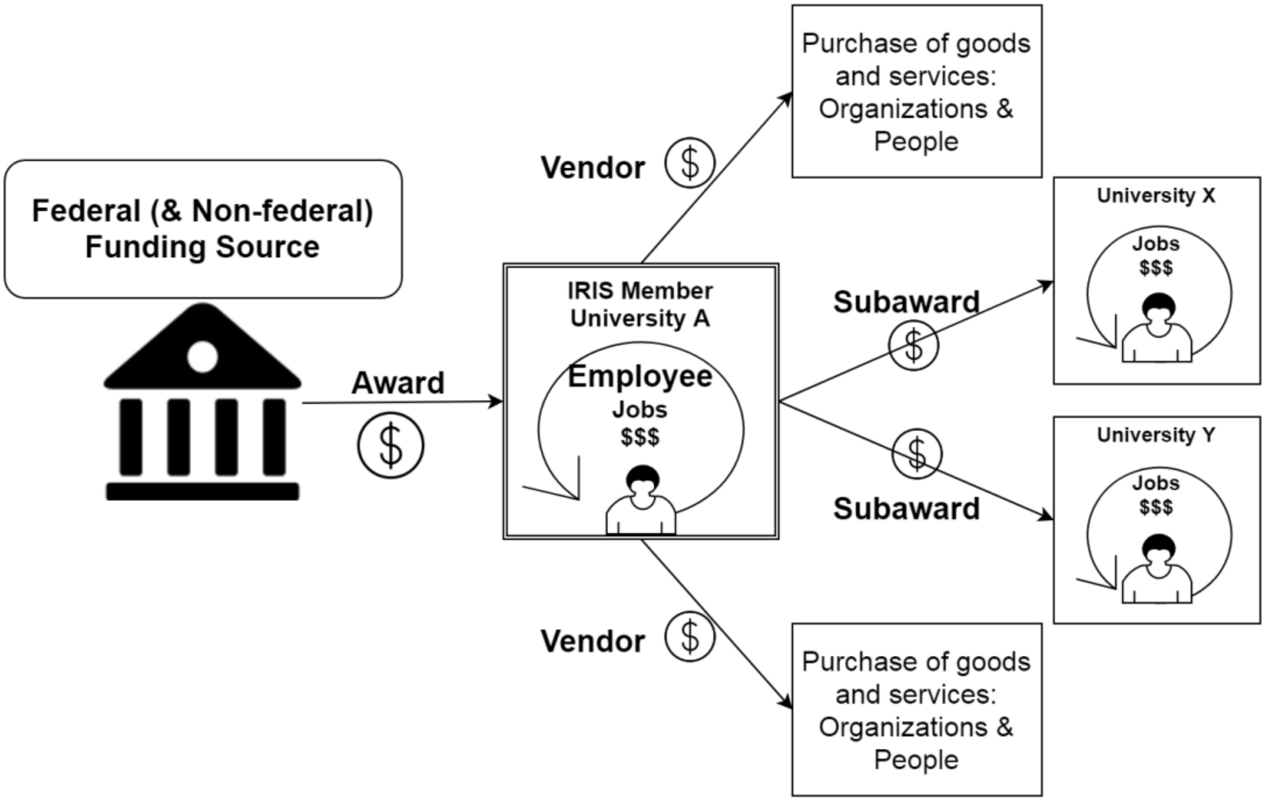
The UMETRICS 2017Q4a dataset for research is based on the fourth quarter 2017 Census data transfer. The release dataset contains a total of over 200 years of annual data from 26 IRIS member universities including coverage between 2001 and 2017. (This coverage varies by institution.) This dataset is comprised of three collections: core, auxiliary, and linkage files.

The first collection includes core files in which researchers will find university financial and personnel administrative data pertaining to sponsored project expenditures at IRIS member universities during a given year. IRIS core files are based on administrative data drawn directly from sponsored projects, procurement, and human resources data systems on each IRIS member university's campus. Individual campus files are de-identified, cleaned and aggregated by IRIS to produce these core files. The 2018 release includes transactions from about 300,000 unique federal and non-federal awards including wage payments to 480,000 individuals as well as transactions to about 600,000 unique vendors (both organizations and individuals). In addition, about 13,000 unique organizations / institutions received subawards from IRIS member universities transferring their prime awards. Approximately 22,000 unique prime awards were used as the funding source to transfer subawards to subrecipients. Vendor and

[4] The full citation is available on the IRIS website at http://iris.isr.umich.edu/research-data/2018datarelease/

subaward payment by grants[5] total $27.2 billion. Figure 1 demonstrates the relationship among the four files in the IRIS core collection.

Figure 1. UMETRICS Core File Relationship in Monetary Flow


Federal (& Non-federal) Funding Source
Award
$
IRIS Member University A
Employee
Jobs
$$$
Vendor
$
Purchase of goods and services: Organizations & People
Subaward
$
University X
Jobs
$$$
Subaward
$
University Y
Jobs
$$$
Vendor
$
Purchase of goods and services: Organizations & People

The auxiliary collection provides researchers with look-up and contextual information on institutions, awards, vendors and subaward recipients. Files can also help to retrieve institutional characteristics as well as to map campus- and sub-organization unit-level codes to more detailed descriptions.

[5] Throughout the documentation, 'grant' and 'award' are interchangeably used as is often the case—e.g., NIH also uses these terms somewhat interchangeably, indicating "award conditions and information for NIH grants." To be precise, an award is defined as "(f)inancial assistance that provides support or stimulation to accomplish a public purpose. Awards include grants and other agreements in the form of money or property in lieu of money, by the federal government to an eligible recipient. The term does not include: technical assistance, which provides services instead of money; other assistance in the form of loans, loan guarantees, interest subsidies, or insurance; direct payments of any kind to individuals; and contracts which are required to be entered into and administered under federal procurement laws and regulations," according to grant terminology at: https://www.grants.gov/web/grants/learn-grants/grant-terminology.html

In addition, IRIS is releasing a linkage collection in which researchers find crosswalks between IRIS data and external datasets (e.g., federal award and publication data) at the individual and award level. The 2018 IRIS data release includes updates for the 2017 data release match tables that: (i) link individual research employees to dissertation data (with a focus on dissertation topics) provided by ProQuest, and (ii) link federal awards from the National Institutes of Health (NIH), National Science Foundation (NSF) and U.S. Department of Agriculture (USDA) to detailed information about the content of grants.

Structure and Naming of Files

The 2018 collection includes fifteen (15) files. For de-identification purposes, the 26 IRIS member universities that contributed data to the UMETRICS 2017Q4a dataset are listed with Institution IDs as a unique identifier throughout the dataset. Data profile information (file name, abbreviation for FSRDC use, file size and record count) is shown in Table 1.

Table 1. UMETRICS 2017Q4a Collections and Files

Collection	File Name	FSRDC File Abbreviation	File Size (in csv format)	Record Count
Core Files (4)	Award Transaction	rawd	1,506,531 KB	7,401,133
	Employee Transaction	remp	2,973,034 KB	16,662,462
	Vendor Transaction	rven	3,381,733 KB	17,537,837
	Subaward Transaction	rsub	118,362 KB	496,545
Auxiliary Files (6)	Sub-organization Units	rsbo	79 KB	1,799
	Object Code	robc	1,150 KB	26,755
	Vendor Lookup	rvlk	87,920 KB	780,242
	Subaward Lookup	rslk	3,015 KB	21,605
	Institution Fastfacts	riff	49 KB	416
	Comprehensive Award List	rcaw	13,331 KB	297,162
Linkage Files (5)	UMETRICS-Federal Agency Award Crosswalk	rawx	13,389 KB	174,609
	UMETRICS-ProQuest Crosswalk	rpqx	2,967 KB	30,246
	NIH Award Details	rnih	4,663,696 KB	1,374,955
	NSF Award Details	rnsf	543,707 KB	230,596
	USDA Award Details	rsda	14,896 KB	70,223

Changes from Last Release

The UMETRICS 2016Q3a research data release in March 2017 included data from 19 IRIS member institutions and was based on the third quarter 2016 Census data transfer. Since the previous release, additional data have been received from member universities, including seven universities that were not represented in the first data release. Key differences between the releases are noted in Table 2 and further described in the 2018 Dataset Descriptions section.

Table 2. Aggregate Data Changes from Last Release

Data Element	2017 Release	2018 Release	Change Rate
Number of Universities	19	26	37%
Number of Files	14	15	7%
Number of Awards	176,971	296,253	67%
Number of Employees	333,944	478,815	43%
Number of Vendors	237,690	582,797	145%
Number of Subawards	9,139	13,262	45%
Number of Prime Awards (as funding source of subawards)	12,282	22,212	81%
Award total (direct total expenditures)	$ 36.4 billion	$ 61.6 billion	70%
Vendor payment total	$ 18.1 billion	$ 18.7 billion	4%
Subaward payment total	$ 6.0 billion	$ 8.5 billion	42%
Number of matched awards (NIH/NSF/USDA)	48,820	82,028	68%
Number of matched dissertation authors	13,660	28,725	110%

Data Coverage

File Availability

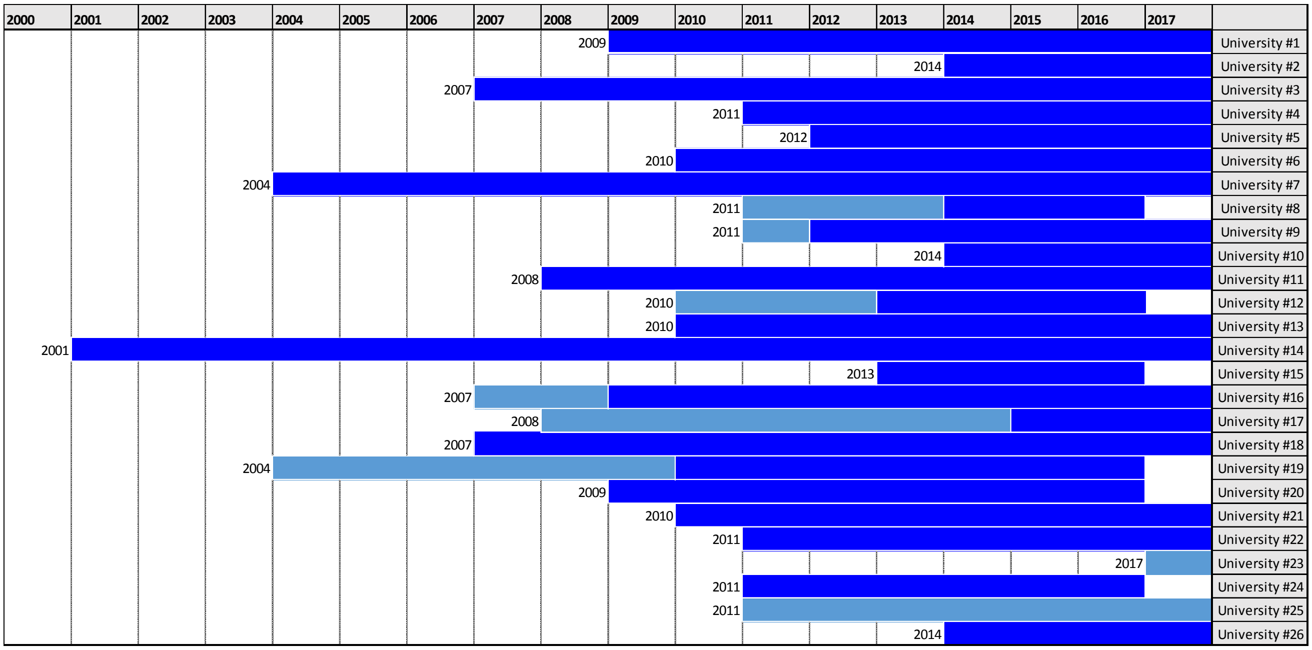
The university representation and temporal data coverage varies significantly across files and institutions as shown in the following tables.

Table 3. University Representation in Files

Collection	Core files				Auxiliary files				Linkage files	
File	Award	Employee	Vendor	Subaward	Institution fastfacts	Object Code Lookup	Sub-Org Units Lookup	Vendor/ Subaward Lookup	Award Crosswalk	ProQuest Crosswalk
Number of Universities	26	25	25	25	26	23	19	26	26	22

The color coding in Figure 2 demonstrates coverage by file—deep blue indicates 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. More detailed information on specific begin- and end-dates in each file for each member university is available in the IRIS Wiki within the IRIS VDE.

Figure 2. Temporal Coverage by University



Missing Data

When dealing with administrative data, it is unavoidable to report missing records in some fields. Detailed information about missing fields in each file are shown in tables in the 2018 Dataset Descriptions section of this documentation. Throughout this documentation, we broadly define ‘missing’ to include: null, blank, ‘ – ‘, ‘#N/A’, and ‘ . ’.

2018 Dataset Descriptions

Core Files

About This File

File Details

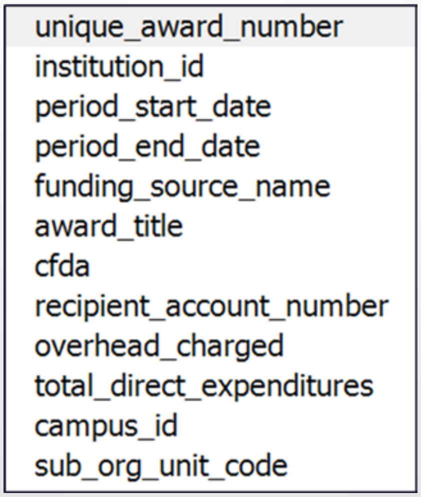
File Name: Core Award
Date Created: 3/15/2018
Record Counts: 7,401,133
Field/Column Counts: 12

File Summary

This file includes all funded awards that IRIS member universities received during a given year. Awards include (but are not limited to):

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

Data Fields



unique_award_number
institution_id
period_start_date
period_end_date
funding_source_name
award_title
cfda
recipient_account_number
overhead_charged
total_direct_expenditures
campus_id
sub_org_unit_code

Data field descriptions are in Appendix A or a separately released Data Dictionary.

Award Transaction

2018 Release Notes

The current release file includes award records from all 26 universities, although some universities lack crucial information such as CFDA (Catalog of Federal Domestic Assistance) numbers, total direct expenditure amounts, and/or award titles.

The "Unique Award Number" is an identifier unique to each award given to a university and serves as the fundamental element by which to associate award information with other transaction data in the employee, vendor, and subaward files. It specifies an award and its funding source, as defined by concatenating the 6-position funding source code (e.g., CFDA number) with an award identifier—either the federal award ID from the awarding Federal Agency (such as the federal grant number, federal contract number, or the federal loan number) or an internal award ID for non-federal awards—with a space or dash in between them. For example:

- 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)

Becoming familiar with both accurately and inaccurately formatted award numbers in the file helps to better map the award information to other data elements within core files, as well as to external data such as federal award details made available by federal agencies, (e.g., NIH, NSF, and USDA).

Most IRIS member universities have started including non-federal grant information in data submissions. Although coverage of federal grants is generally complete with the exception of those universities that hold back sensitive projects, the proportion of coverage of non-federal grants ranges widely from university to university.[6]

Researchers may be interested in accurately identifying award payments used solely for research, separating them from other transactions used for non-research purposes. An example where this is difficult is in work-study program-related payments. It is safe to say some employment and payment based on the work-study program may be unrelated to research, but that is not always the case and it is unclear without more details or background information from member universities. For this current release, we simply note that some universities do include work-study-related payments in their data. Below is a list of CFDA numbers that IRIS believes helps to identify work-study employees and employees paid by Department of Education grant types other than research grants:

- 84.007 Federal Supplemental Educational Opportunity Grants
- 84.033 Federal Work-Study Program
- 84.063 Federal Pell Grant Program
- 84.191 Adult Education National Leadership Activities
- 84.379 Teacher Education Assistance for College and Higher Education (TEACH) Grants
- 84.408 Postsecondary Education Scholarships for Veteran's Dependents
- 84.417 Directed Grants and Awards

[6] According to Grant Terminology provided by Grants.gov (https://www.grants.gov/web/grants/learn-grants/grant-terminology.html), 'Federal Award' is defined as: A. (1) The Federal financial assistance that a non-Federal entity receives directly from a Federal awarding agency or indirectly from a pass-through entity, as described in § 200.101 Applicability of the OMB Uniform Grant Guidance; or (2) The cost-reimbursement contract under the Federal Acquisition Regulations that a non-Federal entity receives directly from a Federal awarding agency or indirectly from a pass-through entity, as described in § 200.101 Applicability of the OMB Uniform Grant Guidance; B. The instrument setting forth the terms and conditions. The instrument is the grant agreement, cooperative agreement, other agreement for assistance covered in paragraph (b) of § 200.40 Federal financial assistance of the OMB Uniform Grant Guidance, or the cost-reimbursement contract awarded under the Federal Acquisition Regulations. (c) Federal award does not include other contracts that a Federal agency uses to buy goods or services from a contractor or a contract to operate Federal government owned, contractor operated facilities (GOCOs).

Any award that is referenced in the employee, vendor, or subaward file should be present in the award file. However, the completeness between the award and three other files (employee, vendor, and subaward varies significantly from university to university (between 58% and 100%) with an average completeness rate of 90%. In other words, approximately 10% of award numbers that appear in either employee, vendor, or subaward file are not present in the current award file. To offset this merge incompleteness, IRIS provides an auxiliary file named Comprehensive Award List.

The "Unique Award Number" field can be used to identify and count unique award numbers given to universities. The descriptive statistics of unique award data (FY 2015) are shown in Table 4 below.

Table 4. Award Data Summary Statistics (FY 2015)

Number of Universities	Total Number of Unique Awards	Min	Max	Mean	Standard Deviation
25	96877	1433	8748	3875	1933.6149

Note: One university was excluded from summary statistics because it submitted data covering only three months (July 2017 - Sept 2017).

Missing Data

Table 5 shows the percentage of records in selected fields in the award file out of the total of 7,401,133 records. Throughout this documentation, we broadly define 'missing' to include: null, blank, ' – ', '#N/A', and ' . '.

Table 5. Missing Records in Award File

Field	Missing	% of Total
Unique Award Number	208,967	2.8
CFDA	514,689	7.0
Total Direct Expenditures	789,986	10.7
Funding Source Name	526,897	7.1
Award Title	711,256	9.6

About This File

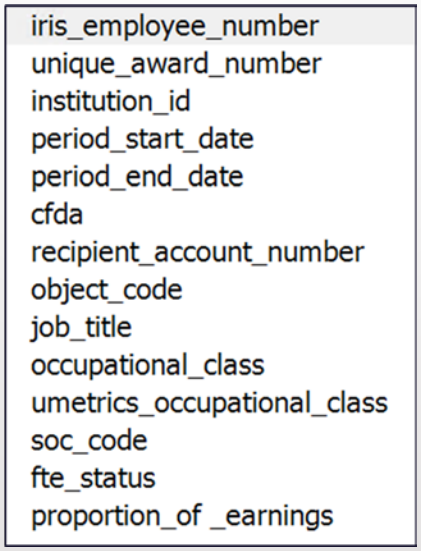
File Details

File Name: Core Employee
Date Created: 3/15/2018
Record Counts: 16,662,462
Field/Column Counts: 14

File Summary

This file includes university payroll transactions for employees paid on any of the sponsored projects in the Award file. While all individuals who charge time to federal or non-federal grants are included in the data, the unit of record is a payment to an individual on a grant in a pay-period. Thus, individuals routinely appear in multiple time periods, on multiple grants.

Data Fields



iris_employee_number
unique_award_number
institution_id
period_start_date
period_end_date
cfda
recipient_account_number
object_code
job_title
occupational_class
umetrics_occupational_class
soc_code
fte_status
proportion_of _earnings

Data field descriptions are in Appendix B or a separately released Data Dictionary.

Employee Transaction

2018 Release Notes

In this current release, IRIS applied a different method (a hash function) to generate a unique identifier (a combination of alpha characters and numbers) to assign to each employee who is paid by grants. Through data hashing, the value in the “IRIS Employee Number” field is generated by reading employee records that cannot be shared and thus are not part of the release file.

When using the Employee Transaction file for research purposes, the oft-used field is “UMETRICS Occupational Class.” Dr. Bruce Weinberg (an IRIS PI) has led the occupation classification coding project at The Ohio State University. Although each IRIS member university provides employee information with a job title and occupational class, these are university-specific. In defining occupations more generally, an innovative occupation classification coding rule and practice was carefully developed (based on performance, research role, professional track, scientific training, and clinical association), and then applied to update the IRIS data for the Employee Transaction file. This classification system was adjusted slightly between the UMETRICS2016Q3a and UMETRICS2017Q4a release and the Employee Transaction file now reflects six broad categories as a first tier of classification and six additional categories for a second tier of classification. This second tier classification further categorizes the Staff titles. Categories that existed in the UMETRICS2016Q3a release have been rolled into the new categories to reflect changes as further discussed below.

The “IRIS Employee Number” field helps to produce unique employee counts. In total, there are 478,815 employees paid by grants across all universities and years. Descriptive statistics of employee data (FY 2015) are shown in Table 6 below.

Table 6. Employee Data Summary Statistics (FY 2015)

Number of Universities	Total Number of Unique Employees	Min	Max	Mean	Standard Deviation
24	178646	2185	17399	7443	3814.1211

Note: Two universities were excluded from the above summary statistics. One university did not submit an employee file, and another submitted data for only three months (July 2017 - Sept 2017).

The “Job Title” field includes data submitted to IRIS from member universities with university-assigned job titles. Because they are university-specific, the number of unique job titles varies across institutions from 9 to 5317 (note that each university's data have different temporal coverage). These university-assigned job titles are classified into one of the 12 UMETRICS occupational categories in the “UMETRICS Occupational Class” field. The current coding has taken a manual approach, as follows:

1) Assign two occupations (primary and secondary);
2) Apply two-level aggregation relationship to university:
    i. The first tier considers six categories of relationships to a university:
        i. Faculty
        ii. Staff
        iii. Post Graduate Research
        iv. Graduate Student
        v. Undergraduate
        vi. Other
    ii. The second tier considers six categories of job responsibilities that disaggregate “Staff” titles from the first tier:
        i. Clinical Staff
        ii. Research Staff
        iii. Research Facilitation Staff
        iv. Instructional Staff

v. Technical Support
vi. Other Staff

3) Reflect classification coding decisions to the IRIS data.

The fifteen categories that existed in the UMETRICS2016Q3a release have been rolled into the new categories to reflect changes as noted in Tables 7 and 8. The more general UMETRICS classification helps to generate unique employee counts by occupational classification.

Table 7. OCC Naming Changes for Tier 1 Categories

Tier 1 Categories	
UMETRICS 2016Q3a categories	**UMETRICS 2017Q4a categories**
1. Faculty	1. Faculty
2. Post Graduate Research	2. Post Graduate Research
3. Graduate Student	3. Graduate Student
4. Undergraduate	4. Undergraduate
5. Staff	5. Staff
	6. Other

Table 8. OCC Naming Changes for Tier 2 Categories

Tier 2 Categories	
UMETRICS 2016Q3a categories	**UMETRICS 2017Q4a categories**
1. Clinician	1. Clinical
2. Staff Scientist	2. Research (combines former categories 2-4)
3. Research Analyst	
4. Technician	
5. Research Support	3. Research Facilitation (combines former categories 5-7)
6. Research Administration	
7. Research Coordinator	
8. Technical Support	4. Technical Support
9. Instructional	5. Instructional
10. Staff Other	6. Other Staff

Missing Data

Table 9 shows the percentage of missing records in selected fields in the employee transaction file out of the total of 16,662,462 records. Throughout this documentation, we broadly define 'missing' to include: null, blank, ' – ', '#N/A', and ' . '.

Table 9. Missing Records in Employee File

Field	Missing	% of Total
Unique Award Number	2,247,206	13.50
CFDA	2,874,103	17.25
Recipient account number	31,766	0.19
Object code	2,434,156	14.61
Job Title	115,506	0.69
SOC code	8,783,249	52.71
FTE status	1,688,357	10.13
Proportion of earnings	2,122,831	12.74
Occupational class	6,291,514	37.76

About This File

File Details

File Name: Core Vendor
Date Created: 3/15/2018
Record Counts: 17,537,837
Field/Column Counts: 19

File Summary

This file includes payments from universities to vendors for goods and services. Vendor transactions can include very small transactions, payments to individuals, and internal fund transfers between units as well as larger purchases from external organizations.

Data Fields

- iris_vendor_id
- unique_award_number
- institution_id
- period_start_date
- period_end_date
- cfda
- recipient_account_number
- object_code
- vendor_ein
- vendor_duns
- vendor_payment_amount
- vendor_name
- vendor_address
- vendor_city
- vendor_state
- vendor_domestic_zipcode
- vendor_foreign_zipcode
- vendor_country
- person_organization_flag

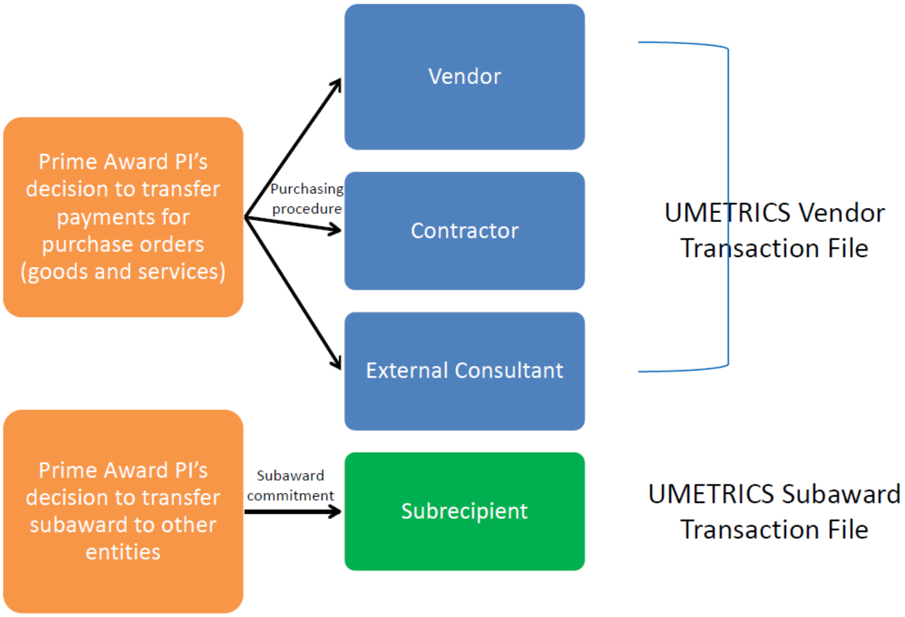
Data field descriptions are in Appendix C or a separately released Data Dictionary.

Vendor Transaction

2018 Release Notes

Vendor and subaward records should be mutually exclusive (Figure 3). No record that appears in one file should appear in the other. If it does, it is considered as duplication of data or, the subaward and contract (i.e., goods or service provider) may not have been clearly differentiated at the time of data submission.

Figure 3. Vendor vs. Subaward File Transactions


Vendor
Prime Award PI's decision to transfer payments for purchase orders (goods and services)
Purchasing procedure
Contractor
UMETRICS Vendor Transaction File
External Consultant
Prime Award PI's decision to transfer subaward to other entities
Subaward commitment
Subrecipient
UMETRICS Subaward Transaction File

In this current release, IRIS applied a similar method (but not quite the same as last release) when generating a unique identifier for vendors (i.e., organizations and people) which / who provided or delivered services and goods to IRIS member universities paid out of research grants. Significant effort was devoted to vendor name cleaning, which enabled IRIS to assign an identifier (a combination of alpha characters and numbers) to unique, more accurately grouped vendors, if not more appropriately disambiguated.[7]

There are other fields that help researchers identify (and characterize) distinct organizations or individuals who were paid by awards; however there are significant numbers of missing records in these fields: e.g., Vendor EIN (62% missing), Vendor DUNS (65% missing), and Vendor Name (34% missing).

In addition to the many fields submitted by member universities, IRIS generates a useful variable ("Person Organization Flag") to differentiate companies, organizations and/or entities from individuals. In assigning this dichotomous category, the algorithm examines every cleaned vendor name as to whether it contains a name in either of the two resources that IRIS has built over the last two years—these two reference tables include over 2 million individual and company names. This method has dramatically improved the quality and outcome of this coding.

For this data release, any personally identifiable information is removed from the Vendor file, e.g., any individual names, personal employee identification numbers, and EINs if vendors and contractors are individuals.

As noted above, IRIS has generated a unique vendor ID, which helps to produce unique vendor counts. There are 582,797 vendors paid by grants. The breakdown by organization-individual flag is shown in Table 10. Descriptive statistics of vendor data (FY 2015) are shown in Table 11.

Table 10. Unique Vendor Counts and Distribution by Vendor Type

Unique Vendor Counts 582,797	
Organization	**Individual**
31%	**69%**

[7] More detail about the work done to clean vendor names is available in documentation for the vendor lookup table.

Table 11. Vendor Data Summary Statistics (FY 2015)

Number of Universities	Total Number of Unique Vendors	Min	Max	Mean	Standard Deviation
25	132032	390	9849	5281	2438.2830

Note: One university was excluded from the summary statistics in this table because the university did not submit a vendor file.

Missing Data

Table 12 shows the percentage of missing records in selected fields in the vendor transaction file out of the total of 17,537,837 records. Throughout this documentation, we broadly define 'missing' to include: null, blank, ' – ', '#N/A', and ' . '.

Table 12. Missing Records in Vendor File

Field	Missing	% of Total
Unique Award Number	593,302	3.38
CFDA	1,384,757	7.90
Vendor Name	5,941,154	33.88
Vendor EIN	10,803,188	61.60
Recipient Account Number	8431	0.05
Vendor DUNS	11428763	65.17
Object Code	197701	1.13
Vendor Payment Amount	1	5.70
Vendor Domestic Zip Code	6627264	37.79

Note that if vendors and contractors are individuals IRIS removes personally identifiable information from the two data fields, Vendor Name and Vendor EIN, and replaces the record with 'masked.' These masked records are not considered as missing and thus are not counted in Table 12.

About This File

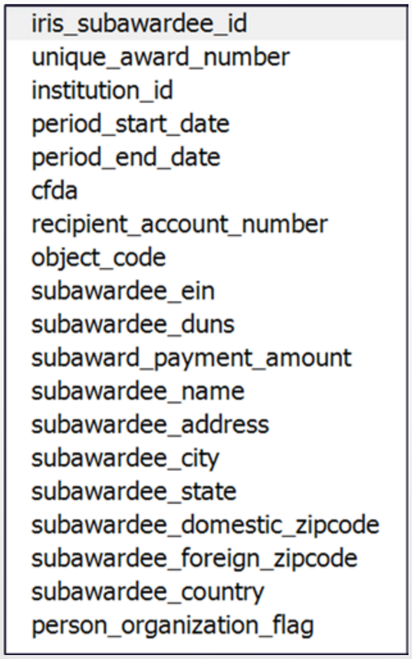
File Details

File Name: Core Subaward
Date Created: 3/15/2018
Record Counts: 496,545
Variable Counts: 19

File Summary

This file includes university expenditure transactions of subawards to another institution or organization provided by IRIS member universities.

Data Fields



iris_subawardee_id
unique_award_number
institution_id
period_start_date
period_end_date
cfda
recipient_account_number
object_code
subawardee_ein
subawardee_duns
subaward_payment_amount
subawardee_name
subawardee_address
subawardee_city
subawardee_state
subawardee_domestic_zipcode
subawardee_foreign_zipcode
subawardee_country
person_organization_flag

Data field descriptions are in Appendix D or a separately released Data Dictionary.

Subaward Transaction

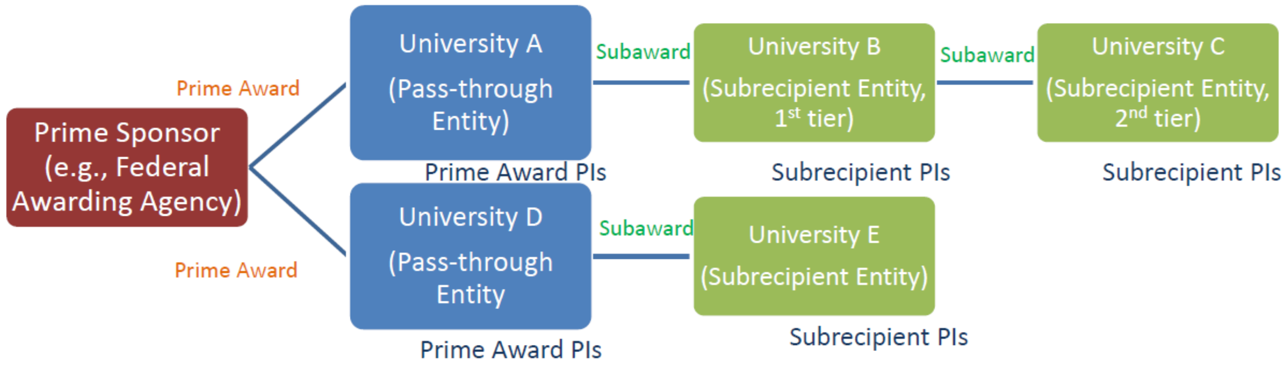
2018 Release Notes

A subaward is a legally binding executed agreement that transfers or delegates a portion of research or substantive intellectual effort of a prime award to another institution or organization. Subawards are not written to individuals. The term subgrant is used when the prime award is a grant and the term subcontract is used when the prime award is a contract.

A prime award could be either federal or non-federal. If federal, a subaward is the award provided by IRIS member universities (as a pass-through entity) to a subrecipient for the subrecipient to carry out part of a federal award received by the pass-through entity. A subaward may be provided through any form of legal agreement, including an agreement that the pass-through entity considers a contract. A subrecipient receives an award of assistance from a pass-through entity and conducts its own scope of work. The subrecipient may also be referred to as a subawardee, subgrantee or lower-tier institution. (See Figure 4.)

In the UMETRICS dataset, records of transactions that meet the aforementioned definition of a subaward as well as, broadly-defined subawards (thus, including subgrants and subcontracts) could appear in this Subaward file. Given that all IRIS member universities uniformly follow the Federal Funding Accountability and Transparency Act (FFATA) Subaward reporting, records in this file should not include payments to a contractor or to an individual that is a beneficiary of a federal program if under federal prime awards.

Figure 4. Award and Subaward Relationship



Vendor and subaward records should be mutually exclusive. No record that appears in one file should appear in the other. If it does, it is considered as duplication of data or, a subaward and contract (i.e., goods or service provider) may not have been clearly differentiated at the time of data submission.

In this current release, IRIS applied a similar method (but not quite the same as last release) in generating a unique identifier (a combination of alpha characters and numbers) for subaward recipients. IRIS has put significant effort into subawardee name cleaning, which enabled IRIS to assign an identifier to unique, more accurately grouped subaward recipients, if not more appropriately disambiguated.

There are other fields that help researchers identify (and characterize) distinct organizations or individuals who are contracted via subaward with IRIS universities; however there are missing records in these columns: e.g., Subaward EIN (31% missing), Subaward DUNS (50% missing), and Subawardee name (10% missing).

In addition to the many fields submitted by member universities, IRIS generates a useful variable ("Person Organization Flag") to differentiate companies, organizations, and/or entities from individuals as subaward recipients. In assigning this dichotomous category, the algorithm examines every cleaned subawardee name as to whether it contains a name in either of the two resources that IRIS has built over the last two years—these two reference tables include over 2 million individual and company names. This method has dramatically improved the quality and outcome of this coding.

For this data release, any personally identifiable information is removed from the Subaward file, e.g., any individual names, personal employee identification numbers, and EINs if subrecipients are individuals.

As noted above, IRIS has generated a unique subaward ID, which helps to produce unique subaward recipient counts. About 13,000 subawards were transferred to organization or institutions from IRIS member universities using their prime awards. Of these 13,000 unique subrecipients, about 20% include ‘university/regent’ in subawardee names.

In theory, subaward recipients should not be individuals, but the IRIS dataset includes individuals (approximately 13%) in the Subaward file as shown in Table 13. Descriptive statistics of subaward data (FY 2015) are shown in Table 14.

Table 13. Unique Subawardee Counts and Distribution by Subawardee Type

Unique Subaward Counts 13,262	
Organization	**Individual**
87%	**13%**

Table 14. Subaward Data Summary Statistics (FY 2015)

Number of Universities	Total Number of Unique Subawards	Min	Max	Mean	Standard Deviation
25	7490	94	824	299	147.2625

Note: One university was excluded from summary statistics in this table because the university did not submit a subaward file.

Missing Data

Table 15 shows the percentage of missing records in selected fields in the subaward transaction file out of the total of 496,545 records. Throughout this documentation, we broadly define ‘missing’ to include: null, blank, ‘ – ‘, ‘#N/A’, and ‘ . ’.

Table 15. Missing Records in Subaward File

Field	Missing	% of total
Unique Award Number	4511	0.91
CFDA	24,499	4.93
Recipient Account Number	432	0.09
Object Code	8,106	1.63
Subawardee EIN	152,315	30.67
Subawardee DUNS	248,682	50.08
Subaward payment amount	0	0
Subawardee name	485	0.10
Subaward domestic zip code	30,068	6.06

Note that if subaward recipients are individuals IRIS removes personally identifiable information from the two data fields, Subawardee Name and Subawardee EIN and replaces the record with ‘masked.’ These masked records are not considered as missing and thus are not counted in Table 15.

Auxiliary Files

About This File

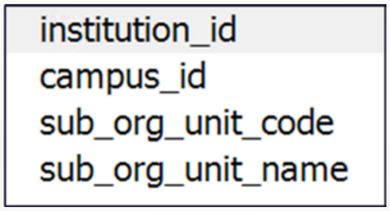
File Details

File Name: Sub-organization Units
Date Created: 3/15/2018
Record Counts: 1,799
Field/Column Counts: 4

File Summary

This file includes a list of sub-organization unit IDs and their names for each IRIS member institution. This should help to map sub-organization unit codes that appear in the Award Transaction file.

Data Fields



institution_id
campus_id
sub_org_unit_code
sub_org_unit_name

Data field descriptions are in Appendix E or a separately released Data Dictionary.

Sub-organization Units

2018 Release Notes

This file was prepared as a lookup table for researchers to map among

i) campus ID (e.g., 10151-03);
ii) sub-organization unit code (e.g., 012); and,
iii) sub-organization unit name (e.g., Engineering, Hospital, etc.).

This file can be used to map fields that appear in the Award Transaction file (in the Core collection) to name descriptions—the award file includes campus ID and sub-org unit code without any identifiable information. When preparing this lookup table, IRIS removed campus names for de-identification purposes and also masked some descriptions of sub-organization unit names if names happened to include the name of campus buildings or anything that is very specific to universities. Note that seven universities do not provide detailed names of their sub-organization units.

About This File

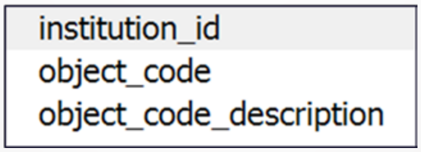
File Details

File Name: Object Code
Date Created: 3/15/2018
Record Counts: 26,755
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



institution_id
object_code
object_code_description

Data field descriptions are in Appendix F or a separately released Data Dictionary.

Object Code

2018 Release Notes

This lookup table was prepared by IRIS to help researchers associate the internal object code or expense type category assigned to each transaction. This file helps to identify payment purposes or resources, although how object codes are grouped and applied depends on university decisions and practices. Object codes are included at the discretion of member institutions. Some universities do not submit an object code lookup file.

Some object (payment purpose) descriptions indicate that internal payments and transactions (e.g., payment to its own university facility and duplication services and university personnel salary) within the institution are also included in the Vendor file.

The two fields, object code and object code description, are not mapped in a 1-to-1 relationship. In some cases, it is one to many and many to one. This is because: 1) Some universities use the same object code for multiple descriptions; 2) Others map one object code to more than one description. For this reason, we did not consider repeated object code numbers as duplicates.

When preparing this file, IRIS removed and masked any potentially re-identifiable information as some universities include university- or campus-names and associated information in object descriptions.

About This File

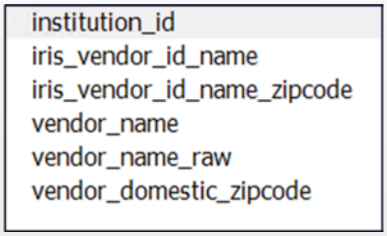
File Details

File Name: Vendor Lookup
Date Created: 3/15/2018
Record Counts: 780,242
Field/Column Counts: 6

File Summary

This file contains two IRIS-generated vendor IDs, vendor names (raw and cleaned), and the vendor location information (based on domestic zip code) if such information was available from the university-submitted data. The underlying information that maps one another in this lookup table originates in the Vendor Transaction file in the core collection.

Data Fields



institution_id
iris_vendor_id_name
iris_vendor_id_name_zipcode
vendor_name
vendor_name_raw
vendor_domestic_zipcode

Data field descriptions are in Appendix G or a separately released Data Dictionary.

Vendor Lookup

2018 Release Notes

This file is new to the current release collection. It contains two IRIS-generated vendor IDs, vendor names (both raw and cleaned), and the vendor location information (based on domestic zip code) if such information was available from the university-submitted data. Each vendor ID is created using one data element (cleaned vendor name) and the combination of the two (cleaned vendor name and vendor location).

In creating this file, IRIS has neither filled missing records nor updated records with additional vendor information except for cleaning vendor name and generating IDs based on cleaned names.

Researchers will benefit from this vendor lookup table in using different unique identifiers to consider location (identified by its domestic zip code if available). Also, one can know raw vendor names and how these names were cleaned by IRIS. See below for details on the vendor name cleaning process.

Vendor Name Cleaning Process

The following steps were applied to vendor name fields as part of data processing and cleaning for the release.

- convert all strings to lowercase and remove the whitespace from the beginning and end of the string
- remove trailing numbers

- remove numbers before asterisk *
- remove string after asterisk *
- remove leading punctuation
- remove any type of punctuation
- remove duplicated blank
- remove abbreviation of business type: e.g., 'inc', 'incorporated', 'ltd','limited', 'co', 'company', 'corp', 'corporation', 'llc', 'liability', 'unltd','unlimited'
- remove trailing and leading spaces from vendor names
- Non-business abbreviations (e.g., "Univ") have been expanded to their complete form (e.g., "University")
- When DBA (doing business as) is included in a name, the name the company is “doing business as” is included and the other name is eliminated.
- When FKA (formerly known as) is included in a name, the newer name is included and the old name is eliminated.
- Potential SSNs are eliminated
- Various phrases used to indicate extraneous vendor data were eliminated, and any accompany extra data was eliminated as well. Example phrases include "See XXX", "do not use", "1 of 2", "Attn:" etc.
- Extraneous information contained in parentheses was removed.
- Removed trailing asterisks, hyphens, and pound symbols
- Eliminated dates in names

About This File

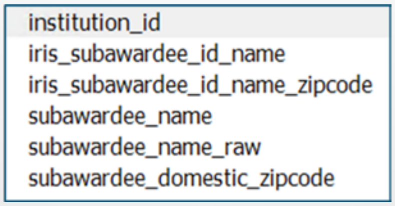
File Details

File Name: Subaward Lookup
Date Created: 3/15/2018
Record Counts: 21,605
Field/Column Counts: 6

File Summary

This file contains two IRIS-generated subawardee IDs, subawardee names (raw and cleaned), and the subaward recipient's location information (based on domestic zip code) if such information was available from the university-submitted data. The underlying information that maps one another in this lookup table originates in the Subaward Transaction file in the core collection.

Data Fields



institution_id
iris_subawardee_id_name
iris_subawardee_id_name_zipcode
subawardee_name
subawardee_name_raw
subawardee_domestic_zipcode

Data field descriptions are in Appendix H or a separately released Data Dictionary.

Subaward Lookup

2018 Release Notes

This file is new to the current release collection. This file contains two IRIS-generated subawardee IDs, subawardee names (both raw and cleaned), and the subaward recipient's location information (based on domestic zip code) if such information was available from the university-submitted data. Each subawardee ID is created using one data element (cleaned subawardee name) and the combination of the two (cleaned subawardee name and location).

In creating this file, IRIS has neither filled missing records nor updated records with additional information about subawardees except for name cleaning and generating IDs based on the cleaned names.

Researchers will benefit from this subaward lookup table in using different unique identifiers to consider location (identified by its domestic zip code if available). Also, one can know raw subawardee names and how these names were cleaned by IRIS. See below for details on the subawardee name cleaning process.

Subaward Name Cleaning Process

The following steps were applied to subawardee name fields as part of data processing and cleaning for the release.

- convert all strings to lowercase and remove the whitespace from the beginning and end of the string

- remove trailing number
- remove numbers before asterisk *
- remove string after asterisk *
- remove leading punctuation
- remove any type of punctuation
- remove duplicated blank
- remove abbreviation of business type: e.g., 'inc', 'incorporated', 'ltd','limited', 'co', 'company', 'corp', 'corporation', 'llc', 'liability', 'unltd','unlimited'
- remove trailing and leading spaces from vendor names
- Non-business abbreviations (e.g., "Univ") have been expanded to their complete form (e.g., "University")
- When DBA (doing business as) is included in a name, the name the company is "doing business as" is included and the other name is eliminated.
- When FKA (formerly known as) is included in a name, the newer name is included and the old name is eliminated.
- Potential SSNs are eliminated
- Various phrases used to indicate extraneous subaward data were eliminated, and any accompany extra data was eliminated as well. Example phrases include "See XXX", "do not use", "1 of 2", "Attn:" etc.
- Extraneous information contained in parentheses was removed.
- Removed trailing asterisks, hyphens, and pound symbols
- Eliminated dates in names

About This File

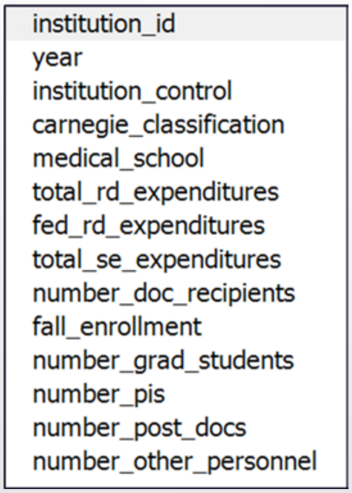
File Details

File Name: Institutional Fastfacts
Date Created: 3/15/2018
Record Counts: 416
Field/Column Counts: 14

File Summary

This file contains information on 26 IRIS member universities, characterizing each institution by its R&D expenditures, student enrollment, number of PIs and research personnel, etc. For de-identification purposes, this file was created by IRIS, retrieving information from different sources via WebCASPAR.

Data Fields



institution_id
year
institution_control
carnegie_classification
medical_school
total_rd_expenditures
fed_rd_expenditures
total_se_expenditures
number_doc_recipients
fall_enrollment
number_grad_students
number_pis
number_post_docs
number_other_personnel

Data field descriptions are in Appendix I or a separately released Data Dictionary.

Institutional Fastfacts

2018 Release Notes

This file was created by IRIS in order to provide institutional context for IRIS member universities. When retrieving institutional data and compiling this file, IRIS had several data processing steps. Most of the data are publicly available (and downloadable) from WebCASPAR (https://ncsesdata.nsf.gov/webcaspar/) including the NSF Higher Education R&D Survey (NSF HERD), NSF-NIH Survey of Graduate Students & Post-doctorates in Science and Engineering, and the Integrated Postsecondary Education Data System (IPEDS) Enrollment Survey. IRIS data users should note that data availability in some fields and campus-specific data may not be available for all 26 member universities. When only campus-level data are available, IRIS aggregated such data to produce the institutional data. In the form of a survey conducted by the National Science Foundation (NSF HERD), universities report their institutional characteristics, but are not obligated to report in each field. For example, some universities have not submitted the latest fall enrollment data, thus it is unavailable from WebCASPAR.

When preparing the current institutional fastfacts file, IRIS collected data for the seven new member universities (2001 -2016) and, added the 2016 data for existing IRIS members previously included in last year's release file. As part of this revision, IRIS dropped two fields (NSF Region and Highest Degree) that were in the 2017 release due to the lack of heterogeneity.

About This File

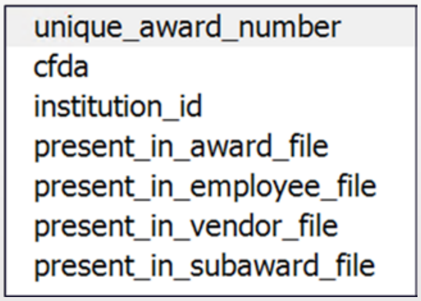
File Details

File Name: Comprehensive Award List
Date Created: 3/15/2018
Record Counts: 297,162
Field/Column Counts: 7

File Summary

This file contains all awards that appear in Employee, Vendor, and Subaward Files. Ideally, each IRIS member institution should submit to IRIS an award file that includes all awards that are present in other files, but some awards were missing from original files. Therefore, IRIS has compiled a comprehensive list of awards that appear in any relevant core file.

Data Fields



unique_award_number
cfda
institution_id
present_in_award_file
present_in_employee_file
present_in_vendor_file
present_in_subaward_file

Data field descriptions are in Appendix J or a separately released Data Dictionary.

Comprehensive Award List

2018 Release Notes

As noted in the Core File section, merge completeness across files (award, employee, vendor, and subaward) is not 100%. This means that some award numbers do not appear in the core Award file. When preparing this comprehensive award list, IRIS examined and decided what level of award number cleaning was necessary without causing any potential 'over-cleaning' or lowering the level of merge completeness. For this reason, IRIS made sure to clean award numbers consistently across files if award numbers appear across multiple files.

This cautious step was particularly important for cases in which universities modify award numbers slightly differently across files; despite the same award number, for example, some universities add extra notes or information that leads or trails award numbers in the unique award number field. In such cases, trailing numbers (e.g., seemingly employee IDs or employee recipient account information) were removed as part of data cleaning when compiling this file.

This list is useful for particular research purposes; for example, when linking federal award data to UMETRICS award data without missing any award numbers that are present in core files.

Linkage Files

About This File

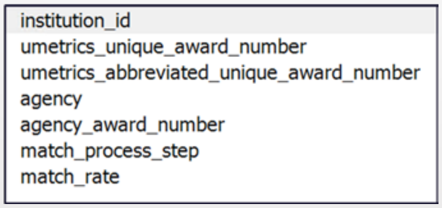
File Details

File Name: UMETRICS-Federal Agency Award Crosswalk
Date Created: 3/15/2018
Record Counts: 174,609
Field/Column Counts: 7

File Summary

This file includes records from a crosswalk between UMETRICS and federal agency award data. This crosswalk table lists all matches including duplicates and possible false positive results generated by award matching code that uses multiple match thresholds. The agency award data used for the record linkage work can be found in the NIH, NSF, and USDA award files in the same Linkage collection.

Data Fields



institution_id
umetrics_unique_award_number
umetrics_abbreviated_unique_award_number
agency
agency_award_number
match_process_step
match_rate

Data field descriptions are in Appendix K or a separately released Data Dictionary.

UMETRICS – Federal Agency Award Crosswalk

2018 Release Notes

IRIS has made no major changes in the award linkage process for this current release—the fundamental algorithm remains the same. Details are discussed below in the Linkage Methodology section. A few notable changes include the update on the configuration file and code application in response to seven new member institutions, as well as, increased effort towards university-specific award data cleaning in UMETRICS for a better match rate.

For this current release, IRIS has done some data cleaning before running the code to match UMETRICS records to three major agency award records.

IRIS has generated a UMETRICS comprehensive award file for award matching in the current release. The federal agency award data used for record matching was downloaded in January 2018 for the most recent publicly available award data. The award data were downloaded from agency websites:

- NSF: https://www.nsf.gov/awardsearch/download.jsp
- NIH:
https://exporter.nih.gov/ExPORTER_Catalog.aspx
- USDA: https://nifa.usda.gov/data

Each agency makes their award information available in a similar but slightly different way in terms of data organization and formatting. NIH and NSF organize and make an annual data file available covering October 1 – September 30. USDA provides their entire award data via the USDA data portal. For particular temporal coverage, one needs to search and filter to download a subset in csv format. NIH makes award data available in both XML and csv format, whereas NSF is available only in XML. IRIS downloaded award data, reformatted, and compiled annual data into one file in order to upload a single file per agency into the SQL Database, making sure to align with the temporal coverage of the UMETRICS data (2001-2017). For USDA, data are not organized per fiscal year, IRIS thus selected out irrelevant cases by removing those with the field ‘project start fiscal year’ indicating anything prior to 2001. For all three agencies, IRIS downloaded federal award data up to the most recent fiscal year (FY2018).

Once the federal agency award dataset was compiled and prepared, IRIS applied the award match code to the UMETRICS award data. Table 16 shows unique award counts from source data and Table 17 provides summary statistics of award data that 26 IRIS member institutions received from the three particular agencies.

Table 16. Number of Awards Used for Linkage

UMETRICS 2017Q4a Award Data	296,253
NIH (core project number)	294,049
NIH (full project number)	1,090,715
NSF (Award ID)	220,421
USDA (Grant Number)[8]	19,281
USDA (Project Number)[9]	16,171

[8] Publicly available USDA award data have a significant number of missing records in both the grant number and project number fields, 72% and 76% respectively. These two fields are crucial linking assets, affecting match rates. Last year, grant numbers were missing in 73% of USDA award data. This continues to be an issue, imposing a challenge for IRIS to improve low match rates across member universities without using an alternative data element to match records.

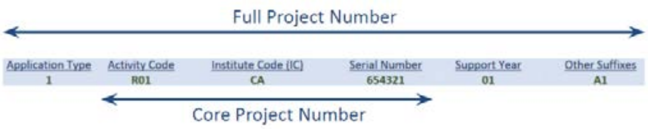
[9] See foot note 8.

Table 17. Award Data Summary Statistics (NIH, NSF, and USDA)

Agency as a Funding Source	Coverage (varies by university)	Number of Universities	Total Number of Unique Awards	Min	Max	Mean	Standard Deviation
NIH	2001-2017	26	79,728	35	13,631	3066.46	3029.90
NSF	2001-2017	26	31,733	192	4,938	1220.5	1027.00
USDA	2001-2017	26	9,554	4	1,923	367.4615	518.88

Linkage Methodology and Findings

Table 18. Linking Elements for UMETRICS-Federal Agency Award Crosswalk

UMETRICS data element		Award file element
Unique Award Number	⟷	NIH Full Project Number  Full Project Number Application Type 1 Activity Code R01 Institute Code (IC) CA Serial Number 654321 Support Year 01 Other Suffixes A1 Core Project Number Core Project Number
Unique Award Number	⟷	NSF Award ID
Unique Award Number	⟷	USDA Grant Number Project Number

1. The code is written to capture both awards that are directly and indirectly funded (subcontracts or transfers).
2. The code is written to filter unique award numbers from UMETRICS that have less than or equal to 5 characters in substrings after CFDA code (or after the 7th position in Python; 8th position in SQL).

3. As to direct funding, the code is written to filter unique award numbers from UMETRICS not starting with ‘47’ ‘93’ or ‘10’ in this particular analysis—the focus is only on NSF, NIH, and USDA.
4. The code is written to map agencies as follows: “agency map” = {'93':'NIH','47':'NSF','10':'USDA'}
5. The code is written not to filter out unique award numbers from UMETRICS starting with ‘00’[10] in case subaward do not carry the same unique award number trailed by the same CFDA number as prime awards.
6. The code is written to filter out unique award numbers from UMETRICS where the substring after the CDFA number (or a pseudo CFDA) in the form of ‘XX.XXX’ is shorter than 7 characters.
7. There are multiple steps (see Figure 5) in the matching procedure to apply both exact and partial matching:
    a. Step 1: Exact Match between each agency data and UMETRICS data.
    b. Step 2: Exact Match after standardization between each agency data and UMETRICS data, i.e., applying ‘punctuation’ dictionary in Python to remove space. This does not filter many records.
    c. Step 3: Matching in such a way that the UMETRICS award number (partially) matched; the agency award number may be longer and the length difference is set for <=5.
    d. Step 4: Reverse of Step 3; Matching in such a way that the UMETRICS award number matched but the agency number may be shorter; the length difference is set for <=5.
    e. Step 5.1: Matching is done by finding a substring of the “Unique Award Number” (from UMETRICS) (90% of it) in agency grant numbers, using the newly compiled public award data table. This applies to NIH and USDA, not to NSF.
    f. Step 5.2: Matching is done for records that still remained after a new Step 5.1 by finding a core project number from NIH within a substring of the UMETRICS “Unique Award Number”. Thus this is NIH specific.
8. The code is written to save output in separate files in csv format for both matched and unmatched results.

[10] If IRIS member universities strictly follow the coding rule that IRIS recommends, federal awards in the unique award number field should not start with 00. However, there are some data inconsistencies.

9. The code is written to record the proportion of string match for every matched record, indicating the match quality. The value ranges between 1 and 0, for instance, if an award from UMETRICS is found exactly matched to a NSF award data, the proportion is 100% and indicator is 1. Results from Steps 1 and 2 have the match quality as 1. Those from Steps 3 and onward have the value less than 1 as partially matched.

Figure 5. Award Matching Process

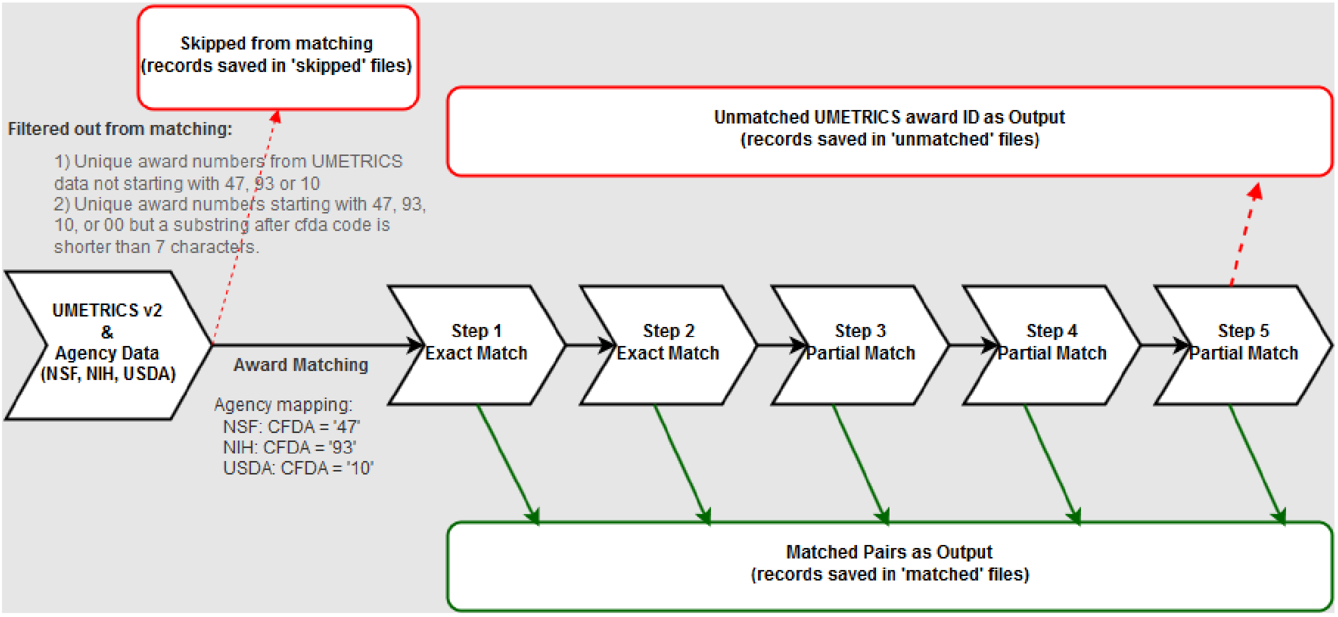

Skipped from matching
(records saved in 'skipped' files)
Filtered out from matching:
1) Unique award numbers from UMETRICS data not starting with 47, 93 or 10
2) Unique award numbers starting with 47, 93, 10, or 00 but a substring after cfda code is shorter than 7 characters.
Unmatched UMETRICS award ID as Output
(records saved in 'unmatched' files)
UMETRICS v2
&
Agency Data
(NSF, NIH, USDA)
Award Matching
Agency mapping:
NSF: CFDA = '47'
NIH: CFDA = '93'
USDA: CFDA = '10'
Step 1
Exact Match
Step 2
Exact Match
Step 3
Partial Match
Step 4
Partial Match
Step 5
Partial Match
Matched Pairs as Output
(records saved in 'matched' files)

Table 19 shows how many NSF, NIH, and USDA award numbers are matched with unique UMETRICS unique award numbers. To generate this table, IRIS focused on cases in which CFDA numbers apparently indicate federal awards coming from NIH, NSF, or USDA. Table 19 shows how many matched records in the release file included awards including (or not including) a correct CFDA number indicating the three agencies as the funding source.

Table 19. Unique Matched Record Counts and Indication of CFDA Numbers

Agency	Total Number of Matched Records	CFDA Indicating the Agency as Funding Source		CFDA *not* Indicating the Agency as Funding Source	
		Number of Matched Records	Percentage	Number of Matched Records	Percentage
NSF	31792	25206	79%	6586	21%
NIH	138018	118362	86%	19656	14%
USDA	4799	2511	52%	2288	48%
Total	174609	146079	84%	28530	16%

About This File

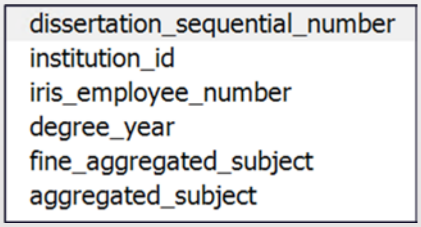
File Details

File Name: UMETRICS – ProQuest Crosswalk
Date Created: 3/15/2018
Record Counts: 30,246
Field/Column Counts: 6

File Summary

This file includes one-to-one match results of the crosswalk between UMETRICS employee names, employee transaction records, and ProQuest publication (dissertation) data with a focus on dissertation subjects. Due to personally identifiable information, the underlying data, i.e., UMETRICS employee names, are not released. Also, due to the terms of the research contract between IRIS and ProQuest, dissertation (publication) IDs originated in the ProQuest database are not made available.

Data Fields



dissertation_sequential_number
institution_id
iris_employee_number
degree_year
fine_aggregated_subject
aggregated_subject

Data field descriptions are in Appendix L or a separately released Data Dictionary.

UMETRICS – ProQuest Crosswalk

2018 Release Notes

Although the original ProQuest data are behind a paywall to most researchers, IRIS has been able to access and use the ProQuest data through a pilot study with ProQuest. This pilot aims to develop programming code to parse publication data and to structure and load the data to a database for crosswalk.

IRIS has made no major changes in the process of linking individual names that appear in the UMETRICS and ProQuest dissertation datasets. Once IRIS completed the initial linkage to individual names, a de-identified crosswalk was constructed and it was made available to other IRIS researchers to work on a subject-focused analysis.

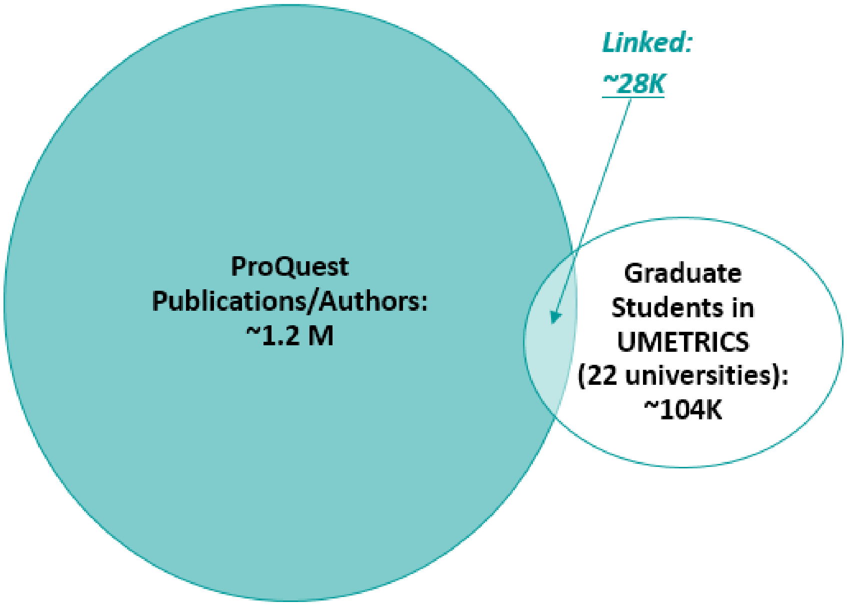
For this current release, the fundamental algorithm remained the same from last year. Unlike last year, data cleaning in the field of individual names was applied prior to the linkage process. The release file includes a linking element from UMETRICS (“IRIS Employee Number”) and ProQuest dissertation records, particularly dissertation topic information, through matched records.

ProQuest Dissertation Data Used for Linkage

For this current release, IRIS first downloaded the latest ProQuest dissertation data in the fall of 2017 to update the publication file previously compiled and used for linkage. The data download (transfer from ProQuest’s server) used ftp and secured login credentials provided by ProQuest to IRIS. ProQuest makes publication data available only in a marc file. Each marc file is first converted to a csv file, and then all is merged into one csv file with 20 variables that include: publication number, author first, middle, and last names, title, subject, advisor, degree year, corporate name, etc. Data preparation also includes retrieval of publication records for IRIS member universities. Using the school code field in the ProQuest database, IRIS retrieved the dissertation data of 22 member universities depending on when the match results were updated and the availability of employee name files from the universities.

- UMETRICS 2017Q4a
  - Dates: 2001-2017
  - Total number of graduate students from 22 IRIS member universities: 103,728
- ProQuest
  - Dates: 1999-2017
  - Total number of dissertation authors: 1,208,844

Figure 6. Unit of Analysis and Matched Record Counts



Linkage Methodology and Findings

Table 20. Linking Elements for UMETRICS-ProQuest Crosswalk

	UMETRICS		ProQuest
Linkage Step 1	IRIS Employee Name IRIS Employee Number	⟵⟶	Dissertation Author Name Publication Number
Linkage Step 2	IRIS Employee Number	⟵⟶	Publication Number Dissertation Subjects

Note: Linking Step 1: Linking Names between UMETRICS and ProQuest
Linking Step 2: Mapping Dissertation Subjects

This file includes 28,725 unique dissertation authors (graduate students) matched to UMETRICS, increased from 13,660 dissertation authors from last year's match results. The unit of record in this crosswalk file is a de-identified individual dissertation, or to put it differently, a de-identified individual dissertation author. Individuals (graduated PhD students) from UMETRICS 2017Q4a, based on their first and last names, are matched to the ProQuest dissertation data prior to selecting dissertation subjects categorized into two sets of 13 groups. Of 26 categories, only the two most aggregated subject categories are released.

There is at least one dissertation subject listed for 99.9%. In the cases where there are multiple topics listed, we have employed two methods to assign a single topic to each degree recipient. First, the topics are listed in order and we have classified people using the first topic listed. We have also classified topics using the frequency with which topics appear, by computing the most common subject listed for each degree recipient. We have compared the most common subject to the first subject listed. Table 21 shows the distribution of subjects associated with (all or matched) dissertations.

Table 21. Common Fine and First Fine Subjects for PQ Dissertation Data

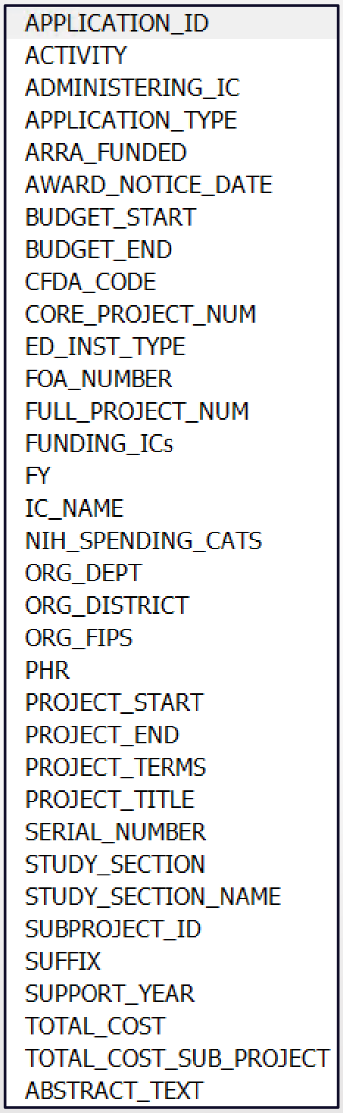
Subject	All Dissertations		Matched Dissertations	
	First Fine Subject	Common Fine Subject	First Fine Subject	Common Fine Subject
Agriculture	1,123	1,060	4,006	3,837
Architecture	42	39	444	381
Area, Ethnic, and Gender studies	145	255	1,279	1,829
Behavioral Sciences	1,045	1,065	6,906	6,998
Biological Sciences	4,579	4,809	17,307	18,044
Business	342	334	2,977	2,900
Chemistry	1,913	1,868	6,495	6,358
Communication and Information sciences	408	382	3,098	2,945
Ecosystem Sciences	246	214	1,323	1,148
Education	1,306	1,371	12,849	13,265
Engineering	9,724	9,568	28,145	27,815
Environmental Sciences	466	472	1,769	1,765
Fine and Performing arts	215	210	5,748	5,467
GeoSciences	885	910	3,293	3,402
Health and Medical Sciences	2,222	2,216	10,224	10,220
History	149	134	3,060	3,038
Interdisciplinary	124	109	421	387
Language and Literature	600	549	8,030	7,673
Law and Legal Studies	7	6	102	88
Mathematics	1,269	1,250	4,736	4,654
Philosophy and Religion	184	190	2,495	2,532
Physical Sciences	1,664	1,696	5,781	5,794
Social Sciences	1,561	1,696	13,190	13,138
TOTAL	30,219	30,403	143,678	143,678

About This File

File Details

File Name: NIH Award Details
Date Created: 3/15/2018
Record Counts: 1,374,955
Field/Column Counts: 34

Data Fields



APPLICATION_ID
ACTIVITY
ADMINISTERING_IC
APPLICATION_TYPE
ARRA_FUNDED
AWARD_NOTICE_DATE
BUDGET_START
BUDGET_END
CFDA_CODE
CORE_PROJECT_NUM
ED_INST_TYPE
FOA_NUMBER
FULL_PROJECT_NUM
FUNDING_ICs
FY
IC_NAME
NIH_SPENDING_CATS
ORG_DEPT
ORG_DISTRICT
ORG_FIPS
PHR
PROJECT_START
PROJECT_END
PROJECT_TERMS
PROJECT_TITLE
SERIAL_NUMBER
STUDY_SECTION
STUDY_SECTION_NAME
SUBPROJECT_ID
SUFFIX
SUPPORT_YEAR
TOTAL_COST
TOTAL_COST_SUB_PROJECT
ABSTRACT_TEXT

Data field descriptions are in Appendix M or a separately released Data Dictionary.

NIH Award Details

2018 Release Notes

This file includes all publicly available NIH award data downloaded from NIH. Of 42 original data fields available from NIH, IRIS dropped seven fields, including PI and funded institution names and location, for de-identification purposes. Data coverage is for years 2001 - 2018 (up to FY 2018 Q2).

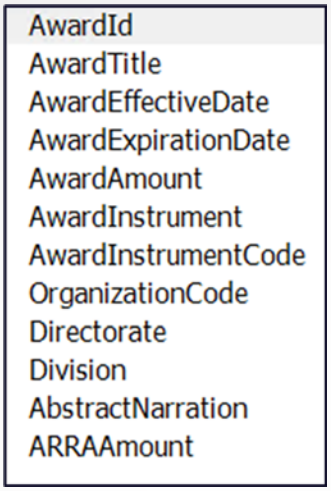
To link UMETRICS award data to NIH data, IRIS used two fields (core project number and full project number) as linking data entities.

About This File

File Details

File Name: NSF Award Details
Date Created: 3/15/2018
Record Counts: 230,596
Field/Column Counts: 12

Data Fields



AwardId
AwardTitle
AwardEffectiveDate
AwardExpirationDate
AwardAmount
AwardInstrument
AwardInstrumentCode
OrganizationCode
Directorate
Division
AbstractNarration
ARRAAmount

Data field descriptions are in Appendix N or a separately released Data Dictionary.

NSF Award Details

2018 Release Notes

This file includes all publicly available NSF award data downloaded from NSF. Of 38 original data fields available from NSF, IRIS dropped 26 fields, including PI information, funded institution’s name and location, for de-identification purposes. Data coverage is for years 2001 - 2018 (up to FY 2018 Q2).

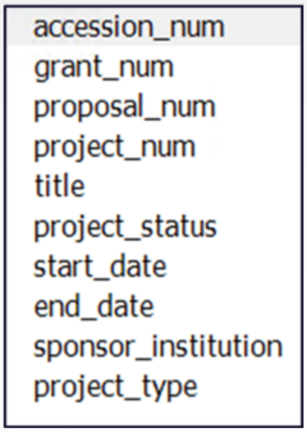
To link UMETRICS award data to NSF data, IRIS used the field “Award ID” as a linking data entity.

About This File

File Details

File Name: USDA Award Details
Date Created: 3/15/2018
Record Counts: 70,223
Field/Column Counts: 10

Data Fields



accession_num
grant_num
proposal_num
project_num
title
project_status
start_date
end_date
sponsor_institution
project_type

Data field descriptions are in Appendix O or a separately released Data Dictionary.

USDA Award Details

2018 Release Notes

This file includes all publicly available USDA award data downloaded from USDA. Of 13 original data fields available from USDA, IRIS dropped 3 fields, including PI information and funded institution's name and location for de-identification purposes. Data coverage is for years 2001 - 2017.

To link UMETRICS award data to USDA data, IRIS used the following two fields: Grant Number and Project Number as linking data entities.

As noted in the award match crosswalk section, the publicly available USDA data has a significant number of missing records in both Award Number and Project Number fields.

Appendix A. Award Transaction Data Fields

Field Name	Data Type	Max Length	Field Definition
Unique Award Number	Char	100	Unique identifier specifying an award and its funding source, as defined by concatenating the 6-position funding source code (e.g., CFDA code) with an award identifier—either the federal award ID from the awarding Federal Agency (such as the federal grant number, federal contract number, or the federal loan number) or an internal award ID for non-federal awards—with 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 ID	Num	8	Unique identifier (a four or five digit number) assigned to each IRIS member university for de-identification purpose
Period Start Date	Num	8	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, 10/1/2015
Period End Date	Num	8	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, 12/31/2014
Funding Source Name	Char	200	Funding source assigned to each project
Award Title	Char	500	Title of award
CFDA	Char	10	A five-digit CFDA (Catalog of Federal Domestic Assistance) number assigned to awards that represents the source of funding; code is retrieved from the unique award number
Recipient Account Number	Char	50	Unique identifier for each project; the account number / code is internal to the university
Overhead Charged	Num	8	Actual overhead dollars charged to the award in the specified period
Total Direct Expenditures	Num	8	Total direct expenditures charged to the award in the specified period

Campus ID	Char	50	Unique identifier of campus to which each award is made; some IRIS universities provide their award data from one campus and others from multiple campuses—this helps to identify the number of campuses (data sources) in this file for a given university; this campus ID was created and reassigned by IRIS for de-identification purposes—each ID is a combination of Institution ID and serial number
Sub-organization Unit Code	Char	50	Sub-organizational unit code to which each funded project is assigned, such as a particular college (not at the level of individual departments) within a given IRIS member university; Universities' codes for their own sub-organization units that have received awards are present in the sub-organization unit field. IRIS has kept raw data submitted by universities and, if universities provided actual sub-organization unit names in this field, then IRIS masked some particular text for de-identification purposes if such names contained potentially re-identifiable information.

Appendix B. Employee Transaction Data Fields

Field Name	Data Type	Max Length	Field Definition
IRIS Employee Number	Char	200	Unique employee ID (random number) assigned by IRIS for grant funded personnel
Institution ID	Num	8	Unique identifier (a four or five digit number) assigned to each IRIS member university for de-identification purpose
Period Start Date	Num	8	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, 10/1/2015
Period End Date	Num	8	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, 12/31/2014
Unique Award Number	Char	100	Unique identifier specifying an award and its funding source, as defined by concatenating the 6-position funding source code (e.g., CFDA code) with an award identifier—either the federal award ID from the awarding Federal Agency (such as the federal grant number, federal contract number, or the federal loan number) or an internal award ID for non-federal awards—with 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)
CFDA	Char	10	A five-digit CFDA (Catalog of Federal Domestic Assistance) number assigned to awards that represents the source of funding; code is retrieved from the unique award number
Recipient Account Number	Char	50	Unique identifier for each project; the account number / code is internal to the university
Object Code	Char	50	Internal object code or other expense type category assigned to a transaction to identify payment purposes or resources

Job Title	Char	200	Job / Occupation title assigned to the funded personnel by IRIS member universities
Occupational Class	Char	50	Job classification provided by IRIS member universities
Umetrics Occupational Class	Char	50	Job classification (12 categories) generated by IRIS; jobs are categorized into 6 major aggregate groups and 6 sub-categories for staff: 1) Macro-level (Tier one) classification (Faculty, Staff, Post Graduate Research, Graduate Student, Undergraduate, and Other); 2) Micro-level (Tier Two) classification for Staff, including: Clinical, Research, Research Facilitation, Technical Support, Instructional, Other Staff
SOC Code	Char	50	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
FTE Status	Num	8	Designation of the status of the funded personnel (full time = 1.0, half time = .5); FTE is a university specific, not an award specific field; the value ranges between 0 and 1
Proportion of Earnings	Num	8	Calculated portion of earnings charged by funded personnel to the award in the specified period, not actual salary or dollar amounts; depending on how much of their salary is derived from an award, the value ranges between 0 and 1

Appendix C. Vendor Transaction Data Fields

Field Name	Data Type	Max Length	Field Definition
IRIS Vendor ID	Char	50	Unique identifier of the vendor (an organization or individual) that provides goods or services paid by an IRIS member university's research grant; after IRIS cleans vendor name records from the data submitted by universities, the identifier is generated by IRIS to uniquely identify vendors based on their cleaned names
Institution ID	Num	8	Unique identifier (a four or five digit number) assigned to each IRIS member university for de-identification purpose
Period Start Date	Num	8	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, 10/1/2015
Period End Date	Num	8	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, 12/31/2014
Unique Award Number	Char	100	Unique identifier specifying an award and its funding source, as defined by concatenating the 6-position funding source code (e.g., CFDA code) with an award identifier—either the federal award ID from the awarding Federal Agency (such as the federal grant number, federal contract number, or the federal loan number) or an internal award ID for non-federal awards—with 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)
CFDA	Char	10	A five-digit CFDA (Catalog of Federal Domestic Assistance) number assigned to awards that represents the source of funding; code is retrieved from the unique award number
Recipient Account Number	Char	50	Unique identifier for each project; the account number / code is internal to the university
Object Code	Char	50	Internal object code or other expense type category assigned to a transaction to identify payment purposes or resources
Vendor EIN	Char	50	The vendor's nine-digit Employer Identification Number (EIN); IRIS has deleted information from this field that would personally identify vendors who are individuals
Vendor DUNS	Char	50	The vendor's nine-digit (DUNS) number to identify business entities on a location-specific basis—the Data Universal Numbering System or D-U-N-S Number is copyrighted and provided by Dun & Bradstreet (D&B)

Vendor Payment Amount	Num	8	The funds charged to the award by the vendor in the specified period
Vendor Name	Char	200	Name of the vendor; IRIS has deleted information from this field that would personally identify vendors who are individuals
Vendor Address	Char	200	Address of the vendor
Vendor City	Char	100	City of the vendor
Vendor State	Char	200	State of the vendor
Vendor Domestic Zipcode	Char	50	US zip code of vendor
Vendor Foreign Zipcode	Char	50	Foreign zip code of vendor
Vendor Country	Char	100	Country of the vendor
Person Organization Flag	Char	1	A binary code (P or O) to differentiate type of vendors: Person (P) or Organization (O)—this dichotomous category was utilized to delete information from fields that would personally identify vendors who are individuals

Appendix D. Subaward Transaction Data Fields

Field Name	Data Type	Max Length	Field Definition
IRIS Subawardee ID	Char	50	Unique identifier of the subawardee to which an IRIS member university (as a pass-through entity) provides program awards / subgrants /subcontracts; after IRIS cleans subawardee name records from the data submitted by member universities, the identifier is generated to uniquely identify subawardees based on their cleaned names
Institution ID	Num	8	Unique identifier (a four or five digit number) assigned to each IRIS member university for de-identification purpose
Period Start Date	Num	8	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, 10/1/2015
Period End Date	Num	8	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, 12/31/2014
Unique Award Number	Char	100	Unique identifier specifying an award and its funding source, as defined by concatenating the 6-position funding source code (e.g., CFDA code) with an award identifier—either the federal award ID from the awarding Federal Agency (such as the federal grant number, federal contract number, or the federal loan number) or an internal award ID for non-federal awards—with 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)
CFDA	Char	10	A five-digit CFDA (Catalog of Federal Domestic Assistance) number assigned to awards that represents the source of funding; code is retrieved from the unique award number
Recipient Account Number	Char	50	Unique identifier for each project; the account number / code is internal to the university
Object Code	Char	50	Internal object code or other expense type category assigned to a transaction to identify payment purposes or resources
Subawardee EIN	Char	50	The subawardee's nine-digit Employer Identification Number (EIN); IRIS has deleted information from this field that would personally identify subawardees who are individuals

Subawardee DUNS	Char	50	The subawardee's nine-digit (DUNS) number to identify business entities on a location-specific basis—the Data Universal Numbering System or D-U-N-S Number is copyrighted and provided by Dun & Bradstreet (D&B)
Subaward Payment Amount	Num	8	The funds charged to the award by the subaward recipient in a specified period
Subawardee Name	Char	200	Name of subaward recipient; IRIS has deleted information from this field that would personally identify subawardees who are individuals
Subawardee Address	Char	200	Address of subaward recipient
Subawardee City	Char	100	City of subaward recipient
Subawardee State	Char	100	State of subaward recipient
Subawardee Domestic Zipcode	Char	50	US zip code of subaward recipient
Subawardee Foreign Zipcode	Char	50	Foreign zip code of subaward recipient
Subawardee Country	Char	100	Country of subaward recipient
Person Organization Flag	Char	1	A binary code (P or O) to differentiate type of subaward recipients: Person (P) or Organization (O); this dichotomous category was utilized to delete information from fields that would personally identify subaward recipients who are individuals

Appendix E. Sub-organization Units Data Fields

Field Name	Data Type	Max Length	Field Definition
Institution ID	Num	8	Unique identifier (a four or five digit number) assigned to each IRIS member university for de-identification purpose
Campus ID	Char	50	Unique identifier of campus to which each award is made; some IRIS universities provide their award data from one campus and others from multiple campuses—this helps to identify the number of campuses (data sources) in this file for a given university; this campus ID was created and reassigned by IRIS for de-identification purposes—each ID is a combination of Institution ID and serial number
Sub-organization Unit Code	Char	50	Sub-organizational unit code to which each funded project is assigned, such as a particular college (not at the level of individual departments) within a given IRIS member university; Universities' codes for their own sub-organization units that have received awards are present in the sub-organization unit field. IRIS has kept raw data submitted by universities and, if universities provided actual sub-organization unit names in this field, then IRIS masked some particular text for de-identification purposes if such names contained potentially re-identifiable information.
Sub-organization Unit Name	Char	64	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 included identifiable information, IRIS removed such information from this field

Appendix F. Object Code Data Fields

Field Name	Data Type	Max Length	Field Definition
Institution ID	Num	8	Unique identifier (a four or five digit number) assigned to each IRIS member university for de-identification purpose
Object Code	Char	50	Internal object code or other expense type category assigned to a transaction to identify payment purposes or resources
Object Code Description	Char	70	Description of internal object code or other expense type category assigned to a transaction—it maps to object code

Appendix G. Vendor Lookup Data Fields

Field Name	Data Type	Max Length	Field Definition
Institution ID	Num	8	Unique identifier (a four or five digit number) assigned to each IRIS member university for de-identification purpose
IRIS Vendor ID Name	Char	50	Unique identifier of the vendor (an organization or individual) that provides goods or services paid by an IRIS member university's research grant; after IRIS cleans vendor name records from the data submitted by universities, the identifier is generated by IRIS to uniquely identify vendors based on their cleaned names; this ID is the same ID included in the Core Vendor Transaction File
IRIS Vendor ID Name Zipcode	Char	50	Unique identifier of the vendor (an organization or individual) that provides goods or services paid by an IRIS member university's research grant; after IRIS cleans vendor name records from the data submitted by universities, the identifier is generated by IRIS to uniquely identify vendors based on the combination of the two data elements (cleaned vendor names and domestic zip code)
Vendor Name	Char	200	Name of the vendor; IRIS has deleted information from this field that would personally identify vendors who are individuals
Vendor Name Raw	Char	200	Original name of the vendor submitted by IRIS member universities–raw data without any data cleaning applied by IRIS
Vendor Domestic Zipcode	Char	50	US zip code of the vendor

Appendix H. Subaward Lookup Data Fields

Field Name	Data Type	Max Length	Field Definition
Institution ID	Num	8	Unique identifier (a four or five digit number) assigned to each IRIS member university for de-identification purpose
IRIS Subawardee ID Name	Char	50	Unique identifier of the subawardee to which an IRIS member university (as a pass-through entity) provides program awards / subgrants /subcontracts; after IRIS cleans subawardee name records from the data submitted by member universities, the identifier is generated to uniquely identify subawardees based on their cleaned names; this ID is the same as the ID included in the Core Subaward Transaction File
IRIS Subawardee ID Name Zipcode	Char	50	Unique identifier of the vendor (an organization or individual) that provides goods or services paid by an IRIS member university's research grant; after IRIS cleans vendor name records from the data submitted by universities, the identifier is generated by IRIS to uniquely identify vendors based on the combination of the two data elements (cleaned vendor names and domestic zip code)
Subawardee Name	Char	200	Name of subaward recipient; IRIS has deleted information from this field that would personally identify subawardees who are individuals
Subawardee Name Raw	Char	200	Original name of the subaward recipient submitted by IRIS member universities—raw data without any data cleaning applied by IRIS
Subawardee Domestic Zipcode	Char	50	US zip code of subaward recipient

Appendix I. Institutional Fastfacts Data Fields

Field Name	Data Type	Max Length	Field Definition
Institution ID	Num	8	Unique identifier (a four or five digit number) assigned to each IRIS member university for de-identification purpose
Year	Num	8	Year (ranging between 2001 and 2016) 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)
Institution Control	Char	50	Defined for academic institutions as private or public (not applicable to biomedical institutions); values include: Public, Private, and Not Applicable; data retrieved from NSF HERD
Carnegie Classification	Char	100	Derived from the 2010 Basic Classification update of the traditional Carnegie Classification Framework for each academic institution; possible values include, e.g., Research Universities-Very High Research Activity, Research Universities-High Research Activity, Doctoral/Research Universities, Master's Colleges and Universities, etc.; data retrieved from NSF HERD and Carnegie Classification of Institutions of Higher Education website
Medical School	Char	100	Indicator for each institution having a medical school included as part of its reporting unit; data retrieved from NSF HERD
Total R&D Expenditures in All Fields	Num	8	R&D expenditures in all fields from current operating funds that are separately budgeted and accounted for—data available from 2003 and beyond; R&D includes expenditures used for: 1) Sponsored research (including federal and nonfederal sponsors); 2) University research (institutional funds that are separately budgeted for individual R&D projects); 3) Other accounts funded by the institution that are only used for research; 4) Recovered and unrecovered indirect costs; 5) Equipment purchased from R&D project accounts; 6) R&D funds passed through to a subrecipient organization, educational or other; 7) Clinical trials, Phases I, II, or III, and; 8) Research training grants

			funding work on organized research projects; data retrieved from NSF HERD
Federally Financed R&D Expenditures in All Fields	Num	8	R&D expenditures in all fields, including direct and recovered indirect costs, funded by all agencies of the Federal government—data available for 2003-16; data retrieved from NSF HERD
Total Higher Education R&D Expenditures for S&E	Num	8	R&D expenditures in science and engineering (S&E) fields from current operating funds that are separately budgeted and accounted for—data available for 2001-16; data retrieved from NSF HERD
Number of Doctorate Recipients	Num	8	All earned doctorates granted by universities—data available for 2001-16; data retrieved from the NSF Survey of Earned Doctorates/Doctorate Records File
Fall Enrollment	Num	8	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 available for 2001-15; data retrieved from the Integrated Postsecondary Education Data System (IPEDS) Enrollment Survey
Number of Graduate Students	Num	8	The number of graduate students enrolled in GSS-eligible science, engineering, and health (SEH) units in the fall of the data collection year—data available for 2001-16; data retrieved from the NSF-NIH Survey of Graduate Students & Postdoctorates in Science and Engineering
Number of Principal Investigators	Num	8	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 available for 2010-16; data retrieved from NSF HERD

Number of Postdocs	Num	8	All personnel paid from 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) that are categorized as postdocs—postdocs are defined by NSF as meeting both of the following qualifications: 1) Holds a recent doctoral degree, generally awarded within the last 5 years; 2) Has a limited-term appointment, generally no more than 5-7 years; data retrieved from NSF HERD
Number of Other Personnel	Num	8	All other 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 are not categorized as principal investigators—data available for 2010-16; data retrieved from NSF HERD

Appendix J. Comprehensive Award List Data Fields

Field Name	Data Type	Max Length	Field Definition
Institution ID	Num	8	Unique identifier (a four or five digit number) assigned to each IRIS member university for de-identification purposes
Unique Award Number	Char	100	Unique identifier specifying an award and its funding source, as defined by concatenating the 6-position funding source code (e.g., CFDA code) with an award identifier—either the federal award ID from the awarding Federal Agency (such as the federal grant number, federal contract number, or the federal loan number) or an internal award ID for non-federal awards—with 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)
CFDA	Char	10	A five-digit CFDA (Catalog of Federal Domestic Assistance) number assigned to awards that represents the source of funding; code is retrieved from the unique award number
Present in Award File	Num	8	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
Present in Employee File	Num	8	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
Present in Vendor File	Num	8	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
Present in Subaward File	Num	8	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

Appendix K. UMETRICS-Federal Agency Award Crosswalk Fields

Field Name	Data Type	Max Length	Field Definition
Institution ID	Num	8	Unique identifier (a four or five digit number) assigned to each IRIS member university for de-identification purpose
UMETRICS Unique Award Number	Char	100	Unique identifier specifying an award and its funding source, as defined by concatenating the 6-position funding source code (e.g., CFDA code) with an award identifier—either the federal award ID from the awarding Federal Agency (such as the federal grant number, federal contract number, or the federal loan number) or an internal award ID for non-federal awards—with 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)
UMETRICS Abbreviated Unique Award Number	Char	50	UMETRICS's unique award identifier without a CFDA number—it is generated by removing CFDA numbers from UMETRICS Unique Award Number
Agency	Char	50	Indicator of one of the three federal agencies (NIH, NSF, or USDA) whose award data are matched to UMETRICS award data
Agency Award Number	Char	50	Unique identifier assigned to each award by federal agencies; award number format varies across agencies
Match Process Step	Char	50	Indicator of which matching step generated each matched pair between UMETRICS and federal agency award record; values include: 1, 2, 3, 4, 5.1 and 5.2 as the code is written to match award numbers through six different methods
Match Rate	Num	8	The proportion of string match between federal agency and UMETRICS award numbers; the value ranges between 0 and 1

Appendix L. UMETRICS-ProQuest Crosswalk Fields

Field Name	Data Type	Max Length	Field Definition
Dissertation Sequential Number	Num	8	A sequential number assigned by IRIS to an individual dissertation selected in this crosswalk table; this number is not an original publication number or dissertation ID generated by ProQuest
Institution ID	Num	8	Unique identifier assigned to each IRIS member university for de-identification purposes; a four or five digit number
IRIS Employee Number	Char	50	Unique employee identifier (random numbers) assigned by IRIS for grant funded personnel
Degree Year	Num	8	The year, in the form of the 4-digit year (e.g., "2010"), when each dissertation was submitted and accepted for a PhD degree
Aggregated Subject	Char	64	The first of 13 aggregated subject fields assigned for each dissertation; retrieved from ProQuest data through record linkage
Fine Aggregated Subject	Char	62	The first of the 13 fine aggregated subject fields assigned for each dissertation; retrieved from ProQuest data through record linkage

Appendix M. NIH Award Details Fields

Field Name	Data Type	Max Length	Field Definition
Application ID	Char	500	A unique identifier of the project record
Activity	Char	50	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	Char	500	Administering Institute or Center - A two-character code to designate the agency, NIH Institute, or Center administering the grant
Application Type	Char	50	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	Char	50	“Y” indicates a project supported by funds appropriated through the American Recovery and Reinvestment Act of 2009
Award Notice Date	Char	50	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 Start	Num	8	The date when a project’s funding for a particular fiscal year begins
Budget End	Num	8	The date when a project’s funding for a particular fiscal year ends
CFDA Code	Char	50	Federal programs are assigned a number in the Catalog of Federal Domestic Assistance (CFDA), which is referred to as the "CFDA code." The CFDA database helps the Federal government track all programs it has domestically funded
Core Project Number	Char	50	Core project number
ED Inst Type	Char	200	Institution type

FOA Number	Char	50	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	Char	50	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	Char	1024	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	Num	8	The fiscal year appropriation from which project funds were obligated
IC Name	Char	500	Full name of the administering agency, Institute, or Center
NIH Spending CATS	Char	1024	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	Char	200	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	Char	50	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	Char	50	The country code of the grantee organization or contractor as defined in the Federal Information Processing Standard

PHR	Char	1024	Submitted as part of a grant application, this statement articulates a project's potential to improve public health
Project Start	Char	50	The start date of a project. For subprojects of a multi-project grant, this is the start date of the parent award
Project End	Char	50	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 Terms	Char	1024	These were thesaurus terms assigned by NIH CRISP indexers, only applicable to projects funded prior to the fiscal year 2008
Project Title	Char	1024	Title of the funded grant, contract, or intramural (sub)project
Serial Number	Char	500	A six-digit number assigned in serial number order within each administering organization
Study Section	Char	500	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	Char	500	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	Num	8	A unique numeric designation assigned to subprojects of a “parent” multi-project research grant
Suffix	Char	50	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	Char	50	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	Num	8	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	Num	8	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 fiscal years
Abstract Text	Char	32800	Abstract of the award

Appendix N. NSF Award Details Fields

Field Name	Data Type	Max Length	Field Definition
Award ID	Char	200	The agency assigned award number (a seven digit number)
Award Title	Char	500	Descriptive title of the project
Award Effective Date	Num	8	Effective date of the award
Award Expiration Date	Num	8	The date on which the award expires
Award Amount	Num	8	The amount obligated to date for the project
Award Instrument	Char	300	Type of Award
Award Instrument Code	Char	100	Code associated with type of award
Organization Code	Char	200	Awardee Institution code
Directorate	Char	100	Department of NSF funding the award
Division	Char	300	Division of NSF funding the award
Abstract Narration	Char	1024	Abstract of the award
ARRA Amount	Char	50	Amount of funding obligated designated as ARRA funding

Appendix O. USDA Award Details Fields

Field Name	Data Type	Max Length	Field Definition
Accession Number	Num	8	One of the two project identifiers for each funded project; Accession Numbers begin with zero (0) and consist of seven (7) digits, e.g., Accession No. (AN) should look like 0134036
Grant Number	Char	50	A Grant No. consists of a two-digit year (2000 and earlier) or a four-digit year (2001 and later) followed by a hyphen and the five-digit financial data code, and then another hyphen and a four- or five-digit sequence number (e.g., 00-38814-9538 or 2004-45066-03027).
Proposal Number	Char	50	A Proposal No. consists of a four-digit year followed by a hyphen and a five-digit sequence number (e.g., 2004-01478).
Project Number	Char	50	One of the two project identifiers for each funded project; Project No. (PN) should look like, KY01056—the project number may contain dashes
Project Title	Char	500	The title of award
Project Status	Char	100	The identification of the status of project: active, extended, new, pending, revised, or terminated
Start Date	Num	8	The start date of a project
End Date	Num	8	The end date of a project
Sponsor Institution	Char	500	The name of sponsoring institution. These include: State Agricultural Experiment Station; Forest Service/USDA; National Institute of Food and Agriculture; Other Cooperating Institutions; Economic Research Service/USDA; Cooperating Schools of Veterinary Medicine; Rural Business-Cooperative Service; Agricultural Research Service/USDA
Project Type	Char	100	This field indicates the funding mechanism for a project from a USDA and/or NIFA perspective. These include: 3D Grant; Animal Health; Cooperative Agreement; Evans-Allen; Hatch; McIntire-Stennis; NRI Competitive Grant; Other Extension Grant; Other Grant; RREA; SERD Grant; Small Business Grant; Special Grant; State; USDA Contract; USDA Grant, and; USDA In-house