THE INSTITUTE FOR RESEARCH ON INNOVATION & SCIENCE

Summary Documentation for UMETRICS 2016Q3a Dataset

March 2017 Data Release

IRIS Technical Team

3/15/2017

A primary goal of the IRIS data release is to enable the research community to access and use this dataset, subject to responsible privacy and confidentiality restrictions. We encourage researchers from all disciplines to apply for approval to access IRIS data through our virtual data enclave. For more information see http://iris.isr.umich.edu/research/data-access-faq/. 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 UMETRICS 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 UMETRICS and external datasets. This data release is also being integrated with U.S. Census Bureau data and will be available in the Federal Statistical Research Data Center (FSRDC) system for those researchers with Special Sworn Status in May 2017.

Table of Contents

Overview

Project Description

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, patenting, research activity, research impact, science of innovation, scientific productivity, and scientific workforce.

Principal Investigator(s)

Jason Owen-Smith, University of Michigan Institute for Social Research
Julia Lane (New York University)
Bruce Weinberg (Ohio State University / National Bureau of Economic Research)
Ron Jarmin (U.S. Census Bureau)
Barbara McFadden Allen (Big Ten Academic Alliance)
James Evans (University of Chicago)

Funding

This project is funded by the Ewing Marion Kauffman and Alfred P. Sloan foundations and by the support of IRIS Member universities.

Citation

The Institute for Research on Innovation & Science (IRIS) UMETRICS Initiative (Universities: Measuring the Impacts of Research on Innovation). Summary Documentation for UMETRICS 2016Q3a Dataset. Ann Arbor, MI: IRIS [distributor], 2017-03-15.

About This Release

A primary goal of the IRIS data release is to enable the research community to access and use this dataset, subject to responsible privacy and confidentiality restrictions. We encourage researchers from all disciplines to apply for approval to access IRIS data through our virtual data enclave. For more information see http://iris.isr.umich.edu/research/data-access-faq/. 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 UMETRICS 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 UMETRICS and external datasets. This data release is also being integrated with U.S. Census Bureau data and will be available in the Federal Statistical Research Data Center

(FSRDC) system for those researchers with Special Sworn Status in May 2017.

Table 1: About This Release

Version	UMETRICS 2016Q3a
Original Data Release	15 March 2017
Current Data Release	15 March 2017

Data Access Notes

Data files in this collection are available in two environments. First, reserachers who are approved to access data in compliance with the IRIS Restricted Data Use Agreement. Approved users may access IRIS data by logging in to the IRIS Virtual Data Enclave (VDE). No downloadable or otherwise publicly accessible data are available outside of the enclave. The Data Access Application Form and Data Use Agreement can be downloaded from the Research Data Access Page of our website (http://iris.isr.umich.edu/research/data-access-faq/).

Second, a copy of this dataset with additional crosswalks to restricted U.S. Census Bureau data resources will be available through the FSRDC system for those researchers with Special Sworn Status (http://www.census.gov/fsrdc).

Regarding crosswalks to Census data, prior to this current data release, the IRIS dataset was loaded to the U.S. Census Bureau's FSRDC environment. 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 RDC administrator for any questions.

Dataset Summary

The UMETRICS 2016Q3a Dataset is comprised of two collections. 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. UMETRICS 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 core files include university data on sponsored project awards, direct cost wage payments from awards to employees, purchases of goods and services from vendors, and subaward transactions to subcontractors. Additional files provide supporting information to characterize and describe IRIS member institutions, identify sub-university units responsible for particular grants, and provide additional detail on object codes included by some data providers.

In addition to core files, we are releasing crosswalk files linking UMETRICS data to external datasets at the individual and award level. In the 2016Q3a release we include match

tables that: (i) link individual UMETRICS 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. This documentation includes details about the data as well as the matching process. The data release includes code and original data files to allow replication and improvement of matching procedures by research users.

Methodology

Mode of Data Collection

- The data collection was initially done by each IRIS member university:
    - Boston 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 Hawaii
    - University of Illinois at Urbana-Champaign
    - University of Iowa
    - University of Kansas
    - University of Michigan
    - University of Missouri
    - University of Pittsburgh
    - University of Wisconsin
- IRIS PIs were not involved in primary data collection

Unit(s) of Observation

- Each expense transaction is a monthly record (except for data anomalies)

Data Contributor(s)

- IRIS member institutions (19 institutions included in UMETRICS 2016Q3a)
- 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 current version of dataset (UMETRICS 2016Q3a) for distribution for research purposes

Data Processing

As part of preparing the dataset for research use, IRIS has worked on data processing. Methods include but are not limited to:

1) Removal of institutional names and campus location information of IRIS member universities;
2) Removal of any personally identifiable information (e.g., any individual names, personal employee identification numbers, and EINs if subrecipient, vendor, or subaward recipients 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 grant funded personnel (see UMETRICS Occupation Classification Coding below); and,
6) Record linkage using a variety of algorithms.

UMETRICS Occupation Classification Coding

Dr. Bruce Weinberg (an IRIS PI) has led his research team to work on occupation classification. 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. The current coding has taken a manual approach:

1) Assign two occupations (primary and secondary);
2) Apply two-level aggregation relationship to university:
    a. The first tier considers relationships to a university (five categories): Faculty, Staff, Postdoc, Graduate Student, and Undergraduate
    b. The second tier considers job responsibilities that disaggregate "Staff" titles from the first tier (10 categories): Clinician, Staff Scientist, Research Analyst, Technician, Research Support, Technical Support, Research Administration, Research Coordinator, Instructional, and Staff Other
3) Reflect classification coding decisions to the IRIS data.

Usage Statistics

This is the first data release; no data usage statistics are available as of 15 March 2017. Before this first release, the UMETRICS dataset was accessed, used, and analyzed only by PIs and their collaborators for research purposes.

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, with the approval identifier REP00000017. Since then two continuing review applications have been reviewed and approved by the IRB. The most recent approval is valid through March 2018.

Related Publications

Kaye G. Husbands Fealing, Stanley R. Johnson, John King, and Julia I. Lane eds. (2017, forthcoming). *Pathways to Research Impact: The Case of Food Safety*, Cambridge University Press.

Catherine Buffington, Benjamin Cerf, Christina Jones, and Bruce A. Weinberg (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

Nikolas Zolas, Nathan Goldschlag, Ron Jarmin, Paula Stephan, Jason Owen- Smith, Rebecca F. Rosen, Barbara McFadden Allen, Bruce A. Weinberg, Julia I. Lane (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

Bruce A. Weinberg, Jason Owen-Smith, Rebecca F. Rosen, Lou Schwarz, Barbara McFadden Allen, Roy E. Weiss, Julia Lane (2014). Science Funding and Short-Term Economic Activity. *Science,* 344(6179): 41-43. DOI: 10.1126/science.1250055

Julia I. Lane, Jason Owen-Smith, Rebecca F. Rosen, and Bruce A. Weinberg (2014). New Linked Data on Research Investments: Scientific Workforce, Productivity, and Public Value. *Research Policy*, 44(9): 1659–1671. DOI: 10.1016/j.respol.2014.12.013

UMETRICS 2016Q3a Dataset Description

File Description

The IRIS 2016Q3a data release includes fourteen (14) files. Data profile information (file name, abbreviation, file descriptions, size, record counts) is shown in Tables 2, 3, 4, and 5 below.

Table 2: File Name and Abbreviation

Category	File Name	File Abbreviation
UMETRICS Core (7 files)	UMETRICS 2016Q3a Core Award Transaction	rawd
	UMETRICS 2016Q3a Core Employee Transaction	remp
	UMETRICS 2016Q3a Core Vendor Transaction	rven
	UMETRICS 2016Q3a Core Subaward Transaction	rsub
	UMETRICS 2016Q3a Core Sub-organization Units	rsbo
	UMETRICS 2016Q3a Core Object Code	robc
	UMETRICS 2016Q3a Core Institutional Fastfacts	riff
UMETRICS Crosswalk (7 files)	UMETRICS 2016Q3a Comprehensive Award List	rcaw
	UMETRICS 2016Q3a NIH-NSF-USDA Award Number List	ranl
	UMETRICS 2016Q3a NIH Award Details	rnih
	UMETRICS 2016Q3a NSF Award Details	rnsf
	UMETRICS 2016Q3a USDA Award Details	rsda
	UMETRICS 2016Q3a UMETRICS-Federal Agency Award Crosswalk	rawx
	UMETRICS 2016Q3a UMETRICS-ProQuest Subject-focused Crosswalk	rpqx

Table 3: File Size and Record Counts

Category	File Name	File Size (in csv)	Record Counts
UMETRICS 2016Q3a Core	Award Transaction	1,356,224 KB	4,688,886
	Employee Transaction	2,811,262 KB	11,205,243
	Vendor Transaction	3,368,592 KB	12,541,329
	Subaward Transaction	99,434 KB	303,109
	Sub-organization Units	308 KB	1,115
	Object Code	25,528 KB	17,950
	Institutional Fastfacts	57 KB	285
UMETRICS 2016Q3a Crosswalk	Comprehensive Award List	7,871 KB	176,971
	NIH-NSF-USDA Award Number List	74,646 KB	1,691,383
	UMETRICS 2016Q3a NIH Award Details	1,832,000 KB	1,477,034
	UMETRICS 2016Q3a NSF Award Details	322,599 KB	197,502
	UMETRICS 2016Q3a USDA Award Details	20,165 KB	64,043
	UMETRICS-Federal Agency Award Crosswalk	7,067 KB	76,497
	UMETRICS-ProQuest Subject-focused Crosswalk	2,185 KB	88,124

Table 4: UMETRICS 2016 Q3a Core File Description

File Name	File Description
Award Transaction	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. Each expense transaction is a monthly record. Any award that is referenced in the Employee, Vendor, or Subaward file should be present in this award file.
Employee Transaction	This file includes university payroll transactions for employees paid on any of the sponsored projects in the Award file. Each expense transaction is a monthly record. 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.
Vendor Transaction	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. These payment records should correspond to the awards that are reported in the Award file. The data could include repeated transactions with both positive and negative payment amounts. Also, data should be rolled up monthly (if there are 10 transactions for a specific vendor in one month those transactions should be summed together and reported as a single value). 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.
Subaward Transaction	This file includes University expenditure transactions of subawards and subcontracts paid on any award reported in the award file. Subaward payments are grouped by recipients. The data could include repeated transactions with both positive and negative payment amounts. Also, data should be rolled up monthly (if there are 10 transactions for a specific subawardee in one month those transactions should be summed together and reported as a single value). 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.
Sub-organization Units	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 other files.
Object Code	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. Object codes are included at the discretion of member institutions.
Institutional Fastfacts	This file contains information on nineteen (19) IRIS member universities, characterizing each institution by its regional location, student enrollment, highest degree, R&D expenditures, number of postdocs, number of PIs, etc. For de-identification purposes, this file was created by the IRIS Technical Team, retrieving information from different sources, including the NSF Higher Education R&D Survey (NSF HERD), the NSF-NIH Survey of Graduate Students & Postdoctorates in Science and Engineering, the Integrated Postsecondary Education Data System (IPEDS) Enrollment Survey via Webcaster, and the Carnegie Classification of Institutions of Higher Education website.

Table 5: UMETRICS 2016 Q3a Crosswalk File Description

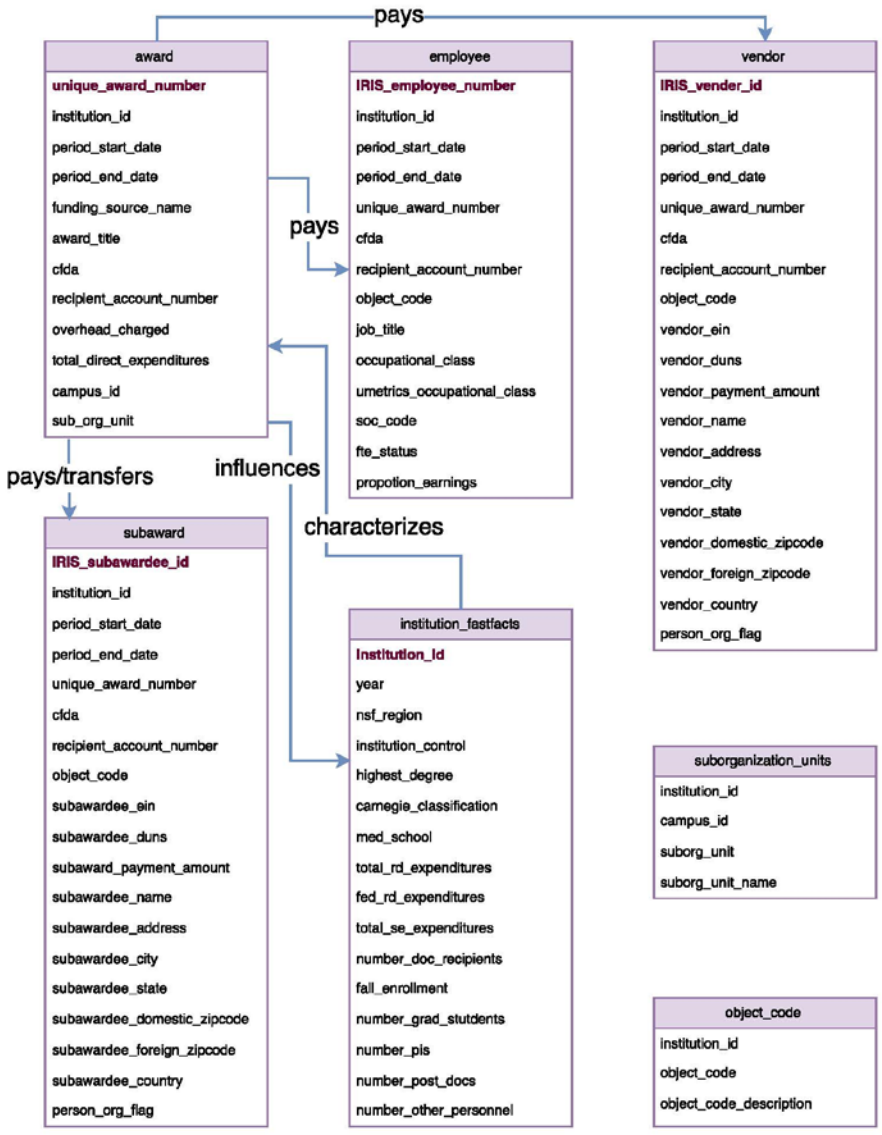
File Name	File Description
Comprehensive Award List	This file contains all awards that appear in Employee, Vendor, and Subaward Files. Ideally, each IRIS member institution should submit to IRIS an award file that includes all awards from other files, but some awards were missing from original files. Therefore, the IRIS Technical Team has compiled a comprehensive list of awards that appear in any relevant core file.
NIH-NSF-USDA Award Number List	This file was compiled for the award crosswalk. This includes only award numbers from NIH, NSF, and USDA. Award numbers were taken from NIH, NSF, and USDA Award Details Files and used for award record linkage.
NIH Award Details	This file includes all publicly available NIH award data downloaded from NIH ExPORTER. Most of the original data fields are kept except for PI and funded institution name fields. Data coverage is for years 1999-2016. In addition to the aforementioned NIH-NSF-USDA award number list, this detailed award file is used for record linkage.
NSF Award Details	This file includes all publicly available NSF award data downloaded from NSF Award Search. Most of the original data fields are kept except for PI and funded institution name fields. Data coverage is for years 1999-2016. In addition to the aforementioned NIH-NSF-USDA award number list, this detailed award file is used for record linkage.
USDA Award Details	This file includes all publicly available USDA award data downloaded from USDA CRIS. Most of the original data fields are kept except for PI and funded institution name fields. Data coverage is for years 1999-2016. In addition to the aforementioned NIH-NSF-USDA award number list, this detailed award file is used for record linkage.
UMETRICS-Federal Agency Award Crosswalk	This file includes all 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 this crosswalk is NIH-NSF-USDA Award Number file.
UMETRICS-ProQuest Crosswalk	This match table includes results of the crosswalk between UMETRICS employee data and ProQuest dissertation data with a focus on dissertation subjects. ProQuest, through active partnerships with more than 700 universities, disseminates and archives more than 90,000 new graduate works each year. These works are available through library subscription databases. IRIS has conducted a pilot study with ProQuest in order to develop programming code to parse publication data and to structure and load the data to a database for crosswalk. The unit of record in this file is a de-identified individual dissertation. Individuals (graduated PhD students) from UMETRICS 2016Q3a, 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. This file includes 88,125 dissertations from 19 IRIS member universities between 2000 and 2016. Of this total, 13,661 dissertation authors (graduate students) are linked to UMETRICS employee IDs. 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 PIs and ProQuest, dissertation (publication) IDs originated in the ProQuest database are not made available.

UMETRICS 2016Q3a Core File Details

Entity Relationship

Figure 1: ER Diagram



Data Coverage

Missing Records and Unique Record Counts

The IRIS Technical Team makes a significant effort to improve the quality of data that IRIS receives from member universities. However, there are still missing records at the time of this data release. The following tables demonstrate the number of missing records in selected data fields and the proportion of it in Award, Employee, Vendor, and Subaward transaction files. Table 11 demonstrates unique award, employee, vendor, and subaward counts.

Table 6: Missing Records in Award Transaction File

Variable/Field	Number of records	Number of missing records (blank or null)	Fraction of missing records
Unique award number	4,637,023	51,863	1.11%
Award title	4,550,084	139,470	2.97%
Recipient account number	4,688,882	4	0.00%
Funding source name	4,485,558	203,328	4.34%
Total direct expenditures	4,290,282	398,604	8.50%
CFDA	4,688,886	0	0.00%

Table 7: Missing Records in Employee Transaction File

Variable/Field	Number of records	Number of Missing Records (blank or null)	Fraction of missing records
IRIS employee number	11,204,556	687	0.01%
Unique award number	11,205,241	2	0.00%
Recipient account number	11,205,243	0	0.00%
Job title	10,883,544	321,699	2.87%
Occupational class	5,227,014	5,978,229	53.35%
Umetrics occupational class	11,205,243	0	0.00%
SOC code	5,075,629	6,129,614	54.70%
FTE status	11,141,781	63,462	0.57%
Proportion earnings	10,209,131	996,112	8.89%
Object code	9,416,687	1,788,556	15.96%

Table 8: Missing Records in Vendor Transaction File

Variable/Field	Number of records	Number of missing records (blank or null)	Fraction of missing records
IRIS vendor id	12,541,329	0	0.00%
Unique award number	12,538,049	3,280	0.03%
Recipient account number	12,538,049	3,280	0.03%
Object code	12,402,098	139,231	1.11%
Vendor EIN (before removing individual EINs)	3,421,694	9,119,635	72.72%
Vendor DUNS	3,156,276	9,385,053	74.83%
Vendor payment amount	12,541,329	0	0.00%
Vendor name (before removing individual names)	8,168,096	4,373,233	34.87%

Table 9: Missing Records in Subaward Transaction File

Variable/Field	Number of records	Number of missing records (blank or null)	Fraction of missing records
IRIS subawardee id	303,109	0	0.00%
Unique award number	302,946	163	0.05%
Subawardee name	302,789	320	0.11%
Recipient account number	302,956	153	0.05%
Object code	300,539	2,570	0.85%
Subawardee EIN	189,951	113,158	37.33%
Subawardee DUNS	98,651	204,458	67.45%
Subaward payment amount	303,109	0	0.00%

Table 10: Unique Counts in UMETRICS Q3a Core Files, All Years

Awards	Employees	Vendors*	Subawards
176,971	333,944	442,206	11,246

Note: Each unique count was calculated by following unique IDs: 1) unique award number was used for award counts; 2) IRIS-generated employee ID was used for employee counts; 3) IRIS-generated vendor ID was used for vendor counts, and; 4) IRIS-generated subawardee ID was used for subaward counts.

* Vendor count includes both organizations and individuals.