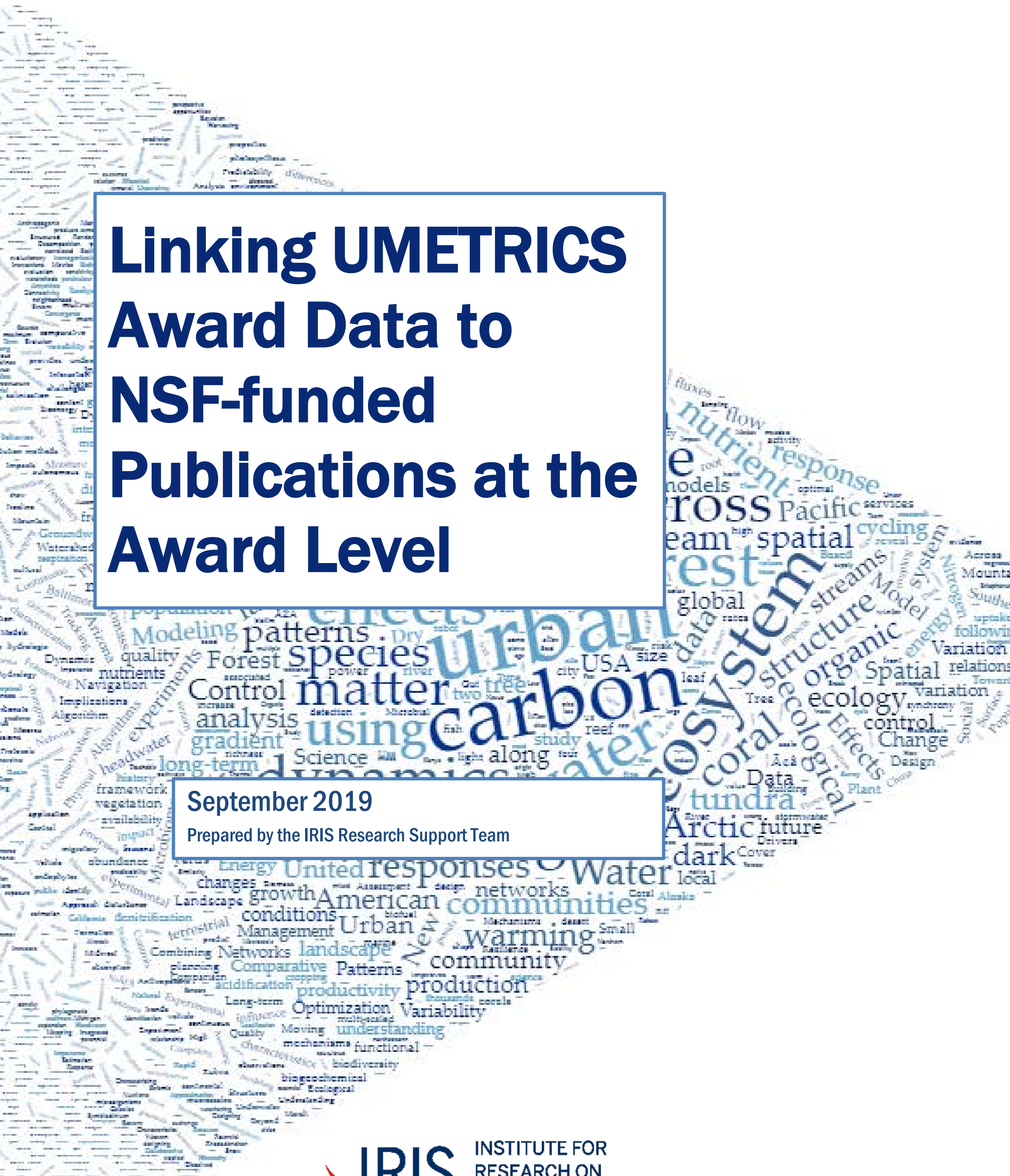

Linking UMETRICS Award Data to NSF-funded Publications at the Award Level

September 2019

Prepared by the IRIS Research Support Team

IRIS

INSTITUTE FOR
RESEARCH ON
INNOVATION & SCIENCE

Credits

Linkage work and documentation were prepared by Raphael Ku, Natsuko Nicholls, and Beth Uberseder.

Citation

The Institute for Research on Innovation & Science (IRIS). Linking UMETRICS Award Data to NSF-funded Publications at the Award Level. Ann Arbor, MI: IRIS [distributor], 2019-10-11, DOI: 10.21987/b9nd-9785

Contents

Tables

2019 Release Notes

Overview

To further support the tracking of research activities, production, and the impact of federally financed projects, this supplementary release in the fall of 2019 produces the results from linking UMETRICS award transaction records to NSF-funded research publications. The crosswalk links NSF-matched UMETRICS awards to publicly available data on NSF publications.

Similar to the NIH publication crosswalk shared in the annual release, [1] this is an award-level publication linkage, not an individual- (publication author-) level linkage. A major difference from the NIH publication linkage is that unlike NIH-funded publications, the publicly available raw data on NSF-funded publications lack a unique identifier like PMID (PubMed publication ID). In a data dictionary, we describe how we assigned two unique identifiers for NSF-funded publications for the source file and crosswalk.

In addition to the crosswalk in which we bridge between the fields of IRIS UMETRICS unique award number, NSF award ID, and IRIS-generated publication IDs, we release two source files; (1) the raw data downloaded from Federal RePORTER, and (2) a cleaned and processed source file. In preparing the latter source file to be more researcher-friendly, we made significant effort to clean, transform, and normalize the data for usability and validity. As a result, over 220,000 publication records (2008-2018) were validated (and corrected/modified/added if needed) through a DOI querying method.[2]

[1] For more details about the UMETRICS-NIH publication linkage, see the IRIS UMETRICS 2019 Annual Data Release Summary Documentation, available from https://iris.isr.umich.edu/research-data/2019datarelease-summarydoc/

[2] When handling any other publication records without a usable DOI in the raw file available from Federal RePORTER, we applied a named entity recognition parser called CERMINE, https://github.com/CeON/CERMINE, developed by the University of Warsaw, in order to determine which items in the citation string can be identified as author, title, etc. However, this method did not generate satisfying results; we thus decided not to include in the release file.

File Summary

This supplemental release includes the following three files saved under the release2019 schema in our SQL server.

Table 1. Fall 2019 Supplemental Release File Description

File name	Record Count	File Size (csv)	Description
link.nsf_pub_xwalk	60,018	5,432 KB	The UMETRICS-NSF publication crosswalk (with NSF award ID, UMETRICS unique award number, and IRIS-generated publication IDs). This crosswalk can be used to link to publication details in the source file described below and also to UMETETRICS data.
link.nsf_pub_source	221,351	87,999 KB	An IRIS-generated publication source file (2008-2018) including IRIS-generated publication IDs, DOIs, and other publication metadata (except for author name fields).
link.nsf_pub_source_raw	662,072	150,823 KB	The raw data of NSF-funded publications (2008-2018) available from Federal RePORTER, which includes only three data fields. Most of the retrievable publication information is stored in the field of 'title'.

As shown in Table 2, the crosswalk includes a list of NSF awards, 9,796 awards (verified as NSF awards through award matching) received by 29 universities (both current and former IRIS members)[3] and about 50,000 NSF-funded publications associated with these awards. If no publication record was verified, such an award was not included in this crosswalk.

The IRIS-generated extracted NSF publication dataset is comprised of 221,351 unique publication entries, representing 176,624 unique DOIs and 44,800 unique NSF award IDs. Note that the two fields in this file, DOI (as a unique identifier for a publication) and NSF award ID, are in many-to-many relationships, where one NSF-funded project can produce multiple research outputs (e.g., journal articles, conference papers, book chapters, etc.) as reported by awardees to the NSF. Similarly, one publication could be a research outcome of multiple projects funded by NSF.

[3] Although the 2019 release file includes the data from 31 universities, the linkage result in this supplementary release does not include two universities. These two universities provided insufficient award data resulting in a 0% match rate to the NSF award data. See the 2019 annual release documentation for more information.

Table 2. Unique Award and Publication Counts

		Uniquely counted NSF awards (with ≥ 1 publication)	Publication records in each file (2008-2018)
NSF Publication Data	**Raw File**	72,239	662,072 (includes duplicates and noise)
	Cleaned File	44,880	221,351 (publication records validated using DOIs)
UMETRICS-NSF Publication Linkage Crosswalk file (29 universities)		9,796	51,622

Potential Use Cases

Using this new crosswalk and source file, along with the existing IRIS release dataset, researchers are able to measure the productivity of NSF projects awarded to universities through the number of publications. As shown in Table 3, on average, one NSF award leads to 4.9 publications with a maximum of 632 papers funded by one single award. Publications are, on average, a result of 1.23 NSF-funded projects with an observed maximum of 44 projects publishing a single paper.[4] About 85% of publications were a result of one NSF funded project.

Table 3. Average Productivity (in Publications) of NSF Awards

	Percentage of NSF awards with ≥ 1 publication	Average number of publications per award	Number of awards per publication
All data	46%	4.9	1.2
IRIS data (29 universities)	49%	5.3	1.1

[4] This journal article, in fact, was written by 1,005 first author and co-authors (see article at https://arxiv.org/abs/1602.03840). It is not surprising that this particular single publication is a product of as many as 44 NSF projects over the course of a decade (2008-2018) that were awarded to one or more research investigators as PI / Co-PI from this large-scale research collaboration. Interestingly, top 10 publications that were funded by more than 25 different NSF projects all come from the same research collaboration.

Table 3 also demonstrates the aggregate data of 29 universities in comparison to the entire pool of NSF-funded publication data. Of 29 universities, one of the most “prolific” NSF projects was awarded to an IRIS university that generated more than 400 publications. In fact, this university has been awarded a dozen NSF projects that published more than 50 articles during the decade, 2008-2018. Table 4 shows the breakdown of several ranges of publication numbers per award. Finally, for comparison purposes, we should note that based on the previously released NIH-funded publication linkage work, a single NIH core project produced 14.75 publications (all data) and 16.29 publications per NIH core project awarded to universities in the IRIS dataset.

Table 4. Publication Counts per Award

Data Source	Number of NSF Awards	Number of Publications per NSF Award					
		5 or less	6-10	11-30	31-50	51-100	100+
IRIS Data	9,796	7,292	1,521	819	95	44	25
		74.4%	15.5%	8.4%	1.0%	0.4%	0.3%
All NSF recipient data	44,800	34,473	6,274	3,432	338	182	101
		76.9%	14.0%	7.7%	0.8%	0.4%	0.2%

Source File Creation and Linkage

The Federal RePORTER website provides data on publications released by the National Science Foundation (NSF), along with other federal agencies and operating divisions (https://federalreporter.nih.gov/FileDownload).[5] The CSV version retrieved contained 662,073 records, with the following fields:

- COMMON_PROJECT_NUMBER
- TITLE
- AUTHORS_LIST

For data analysis and review, the CSV was read as a DataFrame using the Python package *pandas*. Upon further investigation, it was noted that the publication dataset provided did not have any standardized method of encoding publication details, with some publications providing details not available in other entries. Some entries also contained encoding details from its source file, mainly for encoding equations and formulae. Lastly, it was found that the TITLE field may contain author data for certain entries, as seen in the examples in Table 5.

[5] Federal RePORTER provides publication records originating in awards funded by major federal agencies, including ARS, CDMRP, EPA, FS, HHS (AHRQ, CDC, FDA, NIH, and VA), and NSF. Although some agencies provide their project and/or publication data through their own website, NSF does not let users download publication records in bulk unless one attempts to do webscraping (each website page per grant lists its associated publications). We looked into possible options, but we decided to use the NSF publication data available in one csv file from Federal RePORTER for our linkage work given that more time and effort should be directed towards cleaning and verifying publication records mostly for deduplication purposes in order to release the source file that has more usability for research.

Table 5. Examination of Data Found in TITLE Field

TITLE	AUTHORS_LIST	Observations
B. and {Frey}, H., U. and {McFadden}, J. and {Carlson}, C., W. and {Angelopoulos}, V. and {Glassmeier}, K.-H. and {Sibeck}, D., G. and {Weatherwax}, A., {Coordinated observation of the dayside magnetospheric entry and exit of the THEMIS satellites with ground-based auroral imaging in Antarctica}, Journal of Geophysical Research (Space Physics), 114, 2009, 0, 10.1029/2008JA013496	{Mende}, S.	● Use of **{}** ● Names found in TITLE field ● Name format: {Last}, F. ● Name separator: **and**
Personality and Teaming in Bioscience Commercialization: Results from a Naturally-occurring Experiment, Journal of Commercial Biotechnology, 2009	York, A., McCarthy, K., and Darnold, T.	● No additional information after year ● Name format: Last, F. ● Name separator: **,**
Microbial Community Analysis of Two Field-Scale Sulfate-Reducing Bioreactors Treating Mine Drainage, Environmental Microbiology, 10, 2008, 2087	Hiibel, SR; Pereyra, LP; Inman, LY; Tischer, A; Reisman, DJ; Reardon, KF; Pruden, A	● Name format: Last, FF ● Name separator: **;**
Tracking F-region plasma depletion bands using GPS-TEC, incoherent scatter radar, and all-sky imaging at Arecibo, Earth, Planets, Space, 60, 2008, 633	Seker, I., D. J. Livneh, J. J. Makela, and J. D. Mathews	● Journal name also includes commas ● Name format: Last, F. ● Name separator: **,**
Do what the neighbors do: reopening businesses after Hurricane Katrina, Significance, 8, 2011, 160	LeSage JP, Pace RK, Lam N, Campanella R, Liu X.	● Name format: Last FF ● Name separator: **,**

Preprocessing

DOI Extraction

It was noted on a manual review of the dataset that several records contained a Digital Object Identifier (DOI) name, either as a standalone DOI identifier, or as part of a link directing to the original article. Regular expressions were used to extract all text covering the common DOI pattern (10.####/text) and all text prior to a field separator.

DOI String Cleaning

After extraction of the DOI string, the following additional steps were taken to remove extraneous text.

- Removal of common words and acronyms added after the DOI (e.g., Published …, Epub, PDF)
- Iterative removal of non-alphanumeric characters from the end of the DOI string
    - Exception: The ‘) ‘ is kept if it closes a parenthetical in the DOI.
- Removal of remaining whitespace

The resulting string after all the aforementioned steps was added to the *pandas* DataFrame. Lastly, we attempted to reduce the number of queries to be made later on by dropping all duplicate records from the DataFrame.

Methodology

Specific DOI Querying

We used doi.org’s DOI resolver,[6] which can provide metadata for publications when provided with a valid DOI URL. The cleaned DOIs were concatenated with the string ‘https://doi.org/’ to recreate a valid DOI URL. Afterwards, using the standard Python package *requests*, a content negotiated request was made using the DOI URL to the DOI resolver to return a JSON file with the metadata for the publication associated with the DOI provided. The resulting JSON file was then stored for processing and field separation.

CrossRef REST API and Record Comparison

For DOIs that failed to return metadata, the DOI extracted from the raw NSF publication data were assumed to be malformed. To extract additional information from these DOIs, the CrossRef REST API was used, which performs a search of potential matches, rather than a direct reference to a specific paper. Due to this, each query for a specific DOI was limited to the top result, returning metadata in JSON format.

[6] https://dx.doi.org/

Afterwards, the ratio of common word tokens between the title from the CrossRef result and the TITLE field from the raw publication data, as well as the longest common substring length between the CrossRef result's DOI, and the extracted DOI from the raw publication details, were derived. Only results that had a ratio above 75% for both the common title tokens and longest common DOI substring were kept in the dataset.

Processing Results

All metadata in JSON form was loaded into a separate *pandas* DataFrame. The following steps were undertaken to standardize fields, and extract more information from the JSON data structure:

1. Similar unique fields from the DOI resolver and CrossRef REST API results were merged. Notable fields merged:
    a. categories/subjects field for topics covered by the publication. Publications with multiple subjects have all subjects listed with pipes **'|'** as separators
    b. short-container-title and container-title-short fields containing an abbreviated form of the journal publishing the article.
    c. Combining editors and contributors into one cohesive field, with pipes '**|**' as separators.
2. Separating first authors and additional authors for each publication. All publications with multiple authors have all names listed with pipes '**|**' as separators.
3. Transformation of date & time strings into ISO format to be compatible with DATETIME fields in SQL.
4. Removing duplicate publications.

Missing Records

Some metadata was not retrieved and thus the source file does include some missing records, as indicated in Table 6.

Table 6. Missing Records in NSF Publication Source File

Field	Missing	% of Total
award_id	0	0.0%
created_date	301	0.1%
crossref_id	108815	49.2%
doi	0	0.0%
isbn	210087	94.9%
issn	11382	5.1%
issue	40513	18.3%
journal_name	58	0.0%
language	40744	18.4%
member	301	0.1%
page	41843	18.9%
pub_online	83870	37.9%
pub_print	162753	73.5%
publisher	6	0.0%
ref_ct	15633	7.1%
referenced_by_ct	16555	7.5%
short_journal_name	20159	9.1%
subject	145105	65.6%
title	41	0.0%
type	0	0.0%
unique_doi_id	0	0.0%
unique_pub_id	0	0.0%
url	0	0.0%
volume	14946	6.8%

UMETRICS-NSF Publication Linkage

Methodology

Using the previously released NSF-UMETRICS award crosswalk (thus 29 universities successfully matched to NSF award data), we applied an exact match method to bridge from UMETRICS award data to NSF Award ID and then to IRIS-generated NSF-funded publication IDs. Note that linking entities (NSF Award ID—Publication IDs) are in many-to-many relationships.

Findings

The NSF publication-UMETRICS award crosswalk includes 9,796 (uniquely counted) NSF award IDs. Each award ID is linked to NSF-funded publication records. A total of 9,796 (uniquely counted) NSF award IDs are matched to 51,622 publications (221,351 if uniquely counted). Since the data coverage of NSF-funded publications available from Federal RePORTER is between 2008 and 2018, we adjusted the award count by adding two additional years prior to 2008 (thus awards were selected from 2006 and 2018), considering a time lag between an awarded year and the year of publication. Summary statistics are shown in Table 7.

Table 7. NSF-funded Publication Match Results Summary Statistics

Number of Institutions	**Total number of awards validated as NSF awards through award matching**		**NSF awards with publications**	**NSF-funded publications**				
	(all years; coverage varies by university)	***(2006-2018)***		**Total**	**Average per award**	**Min**	**Max**	**SD**
29	29,211	20,091	9,796	51,622	5.27	3.75	8.51	1.26

NSF Publication Crosswalk

File Details

File Name: link_nsf_pub_xwalk
Record Counts: 60,018
Field/Column Counts: 9

File Summary

The UMETRICS-NSF publication crosswalk can be used to link to NSF publication details and also to UMETRICS data.

Data Fields

Fields appearing across files	Fields unique to this file
institution_id	begin_year
award_id	end_year
unique_award_number	
unique_pub_id	
unique_doi_id	
award_effective_date	
award_expiration_date	

Table 8. Link_nsf_pub_xwalk Data Fields

Field Name	Column Name	Data Type	Set Length	Max Length	Field Definition
Institution ID	institution_id	int	4	4	IRIS-generated unique identifier assigned to each IRIS member university for de-identification purposes. Values are four or five digit numbers
Award ID	award_id	varchar	10	7	NSF assigned award number (a seven digit number)
Unique Award Number	unique_award_number	varchar	100	43	University-generated unique identifier specifying an award and its funding source, made up of the 5-digit funding source code (e.g., CFDA number) and an award identifier. Since this crosswalk is based on the NSF-UMETRICS award match, the value in this field should start with 47.xxx (indicating the NSF's CFDA) and be in a format like “47.050 1234567”. However, some unique award numbers include the acronyms of NSF directorates or programs, because our award matching algorithm successfully captured a part of the string (7-digit NSF award ID) out of the entire string value that university-submitted as unique award number
Unique Publication ID	unique_pub_id	varchar	50	11	IRIS-generated identifier assigned to each NSF-funded publication. Each ID is a combination of NSF award ID and a serial number helpful to identify the award to which a given publication is funded; e.g., the two publications, xxxxxxx-1 and xxxxxxx-2, are both funded by the same NSF award
Unique DOI ID	unique_doi_id	int	4	4	IRIS-generated unique identifier assigned to each NSF-funded publication based on its unique DOI
Award Effective Date	award_effective_date	date	3	3	Effective date of the NSF award
Award Expiration Date	award_expiration_date	date	3	3	The date on which the NSF award expires
Begin Year	begin_year	int	4	4	Year extracted from the award_effective_date field
End Year	end_year	int	4	4	Year extracted from the award_expiration_date field

NSF Publication Details

File Details

File Name: link_nsf_pub_source
Record Counts: 221,351
Field/Column Counts: 24

File Summary

This IRIS-generated publication source file contains IRIS-generated publication IDs, DOIs, and other publication metadata (except for author name fields) for 2008-2018.

Data Fields

Fields appearing across files	**Fields unique to this file**	
award_id	doi	issue
unique_doi_id	journal_name	language
unique_pub_id	short_journal_name	member
	subject	page
	title	pub_online
	isbn	pub_print
	issn	publisher
	url	ref_ct
	crossref_id	type
	created_date	volume
	referenced_by_ct	

Table 9. Link_nsf_pub_source Data Fields

Field Name	**Column Name**	**Data Type**	**Set Length**	**Max Length**	**Field Definition**
Award ID	award_id	varchar	10	7	NSF assigned award number (a seven digit number)
Unique DOI ID	unique_doi_id	int	4	4	IRIS-generated unique identifier assigned to each NSF-funded publication based on its unique DOI.
Unique Publication ID	unique_pub_id	varchar	50	11	IRIS-generated identifier assigned to each NSF-funded publication. Each ID is a combination of NSF award ID and a serial number helpful to identify the award to which a given publication is funded; e.g., the two publications, xxxxxxx-1 and xxxxxxx-2, are both funded by the same NSF award.
DOI	doi	varchar	100	59	The Digital Object Identifier (DOI), a unique alphanumeric string beginning with '10.####/text' assigned by a registration agency to provide a persistent link to the location of a publication or other digital object
Journal Name	journal_name	varchar	1000	500	Name of the journal
Short Journal Name	short_journal_name	varchar	200	167	Abbreviated name of the journal (e.g., Am. J. Hum. Biol. for the American Journal of Human Biology)
Subject	subject	varchar	500	280	Subject fields assigned to the publication; multiple subjects are combined with '|'
Title	title	varchar	4000	3756	Title of the publication
ISBN	isbn	varchar	100	61	International Standard Book Number (ISBN), a unique numeric identifier
ISSN	issn	varchar	100	33	International Standard Serial Number (ISSN), an eight-digit serial number used to uniquely identify a serial publication
URL	url	varchar	200	141	URL of the publication from the DOI resolver
Crossref ID	crossref_id	varchar	100	78	Publication ID assigned by CrossRef
Date Created	created_date	date	3	3	Date the DOI was minted

Referenced By Count	referenced_by_ct	int	4	4	Number of times that a publication was cited
Issue	issue	varchar	100	47	The issue number in which an article is published
Language	language	varchar	50	3	Language of the publication
Member	member	int	4	4	Crossref member ID; https://www.crossref.org/reports/members-with-open-references/
Page	page	varchar	50	25	Pages for the article
Published Online	pub_online	date	3	3	Date published online
Published in Print	pub_print	date	3	3	Date published in print
Publisher	publisher	varchar	200	125	Publisher name
Reference Count	ref_ct	int	4	4	Number of references in the publication
Type	type	varchar	50	19	Type of publication, e.g., journal-article, book, book chapter, paper-conference, etc.
Volume	volume	varchar	50	9	The volume number of a published journal, or the number of a printed volume for a book or conference proceedings

NSF Publication Source File (Raw)

File Details

File Name: link_nsf_pub_source_raw
Record Counts: 662,072
Field/Column Counts: 3

File Summary

This file contains the raw data of NSF-funded publications from 2008-2018 available from Federal RePORTER. Most of the retrievable publication information is stored in the field of 'title'.

Data Fields

Fields appearing across files	Fields unique to this file
award_id	title
	authors_list

Table 10. Link_nsf_pub_source_raw Data Fields

Field Name	Column Name	Data Type	Set Length	Max Length	Field Definition
Award ID	award_id	int	4	4	NSF assigned award number (a seven digit number). In the original source file this field was named 'common_project_number'
Title	title	varchar	2000	1168	Although the raw data field is called title, this field contains various pieces of publication record and metadata, including, journal name, article title, volume, issue, publication date, etc.
Authors List	authors_list	varchar	1000	572	A list of authors