Supplementary Documentation for 2018 Data Release: UMETRICS-Patent Crosswalk and Patent Data

June 2018

IRIS Research Team

IRIS INSTITUTE FOR RESEARCH ON INNOVATION & SCIENCE

Data Access Statement

This is the publicly available supplementary documentation for 2018 Data Release. Access to the full documentation is restricted to authorized IRIS data users.

Contents

UMETRICS-Patent Crosswalk 2
2018 Release Notes 2
Discussion 3
Findings 5
Future Work 5
Patent Subset (Linkage Source) File 6
2018 Release Notes 6
USPTO PatentsView Patent Data 7
2018 Release Notes 7
Appendix A. UMETRICS-Patent Crosswalk Data Field Description 8
Appendix B. Patent Subset (Linkage Source) Data Field Description 9
Appendix C. USPTO Patent Data File and Field Descriptions 10

Tables

Table 1. Number of Individual Names Used for Linkage (and Number of Associated Patents) 4
Table 2. Linking Elements 4
Table 3. UMETRICS-Patent Linkage: Input-Output Data Summary Statistics 5

About This File

File Details

File Name: UMETRICS –Patent Crosswalk
Date Created: June 2018
Record Counts: 26,645
Field/Column Counts: 4

File Summary

This file includes match results of the crosswalk between UMETRICS employee names, employee transaction records, and USPTO patent data at the individual level with a focus on current IRIS member universities. Records are linked through IRIS-generated unique employee number and patent inventor identifier, as well as patent numbers. Due to personally identifiable information, the underlying data, i.e., UMETRICS employee names and inventor's names, are not released.

Data Fields

patent_number
inventor_id
iris_employee_number
institution_id

Data field descriptions are in Appendix A.

UMETRICS-Patent Crosswalk

2018 Release Notes

This is the first individual-level patent linkage work done by IRIS and is made available to the IRIS research community as part of the 2018 release supplement. The patent-related data and documentation are available only within the IRIS Enclave, not through FSRDC.

This particular linkage aims to associate research grant and transaction activities (as input) to the output of research with a special focus on patenting, providing another way of measuring the impact of research.

Once IRIS completed the initial linkage to individual names (personally identifiable information), a de-identified crosswalk was constructed and it was made available to other IRIS researchers to work on further linking UMETRICS data to a variety of US patent data files.

For this release, the record linkage algorithm was developed by a Research Scientist, Dr. Chia-Hsuan Yang, who worked for IRIS Co-PI Dr. Julia Lane at New York University from March 2016 to March 2018. The matching code in R is available upon request. Although her code was reviewed by PI/Co-PIs and IRIS staff, there is still room for improvement. We encourage IRIS researchers to share feedback and suggestions for improvement with us and more widely with the IRIS research community.

As part of data preparation, patent data were downloaded from PatentsView (http://www.patentsview.org/download/) in February 2018 and were linked to UMETRICS employee data. A portion of the

downloaded patent data was used to create a source file that includes patent assignee organization names associated with current IRIS universities. The crosswalk includes 23 universities—certain university data were not included in the patent source file (thus not linked to UMETRICS) due to the timing of matching (February 2018) and the availability of employee name files from the universities.

In addition to the crosswalk and source files, IRIS also provides researchers within our virtual data enclave a total of 14 relational tables that are available from PatentsView (e.g., patent application, assignee, inventor, abstract, and citation records). Although more than 50 files are available from PatentsView, these 14 files include most important elements which we hope should enable researchers to do additional linkage work. For de-identification purposes, IRIS did not download location files and also dropped personally identifiable information fields from some files. More details are discussed in the UPTO PatentsView Patent Data Section.

Discussion

Data Preparation

The data pre-processing began by downloading publicly available patent data from the USPTO PatentsVeiw website, including variables such as inventor ids, inventor names, patent numbers, and application dates. Then we filtered irrelevant data from the entire patent inventor and patent datasets in order to create a subset file of patent data which served as a linkage source file. In filtering, we first built a lookup table that contains patent assignee organization names that are relevant to IRIS member universities. A group of assignee organization names were carefully reviewed and compiled. This list contains 188 different patent assignee organization names that are mapped to 23 IRIS member universities, including official university names, acronyms, associated patent offices, associated research foundations, etc. For example:

- ABC University
- ABC University Foundation
- ABC University, Advanced Research and Technology Institute

} Mapped to ABC University (with IRIS member institution ID)

- University of XYZ
- The Regents of the University of XYZ
- University of XYZ Research Foundation

} Mapped to University of XYZ (with IRIS member institution ID)

Once data retrieval was complete, we applied data pre-processing to both input files (UMETRICS employee and patent subset files). All string were standardized by removing special characters (e.g., whitespaces, punctuations, a mix of lower and upper cases, etc.).

Source Data Overview

Tables 1 and 2 give an overview of the source data and data elements that were used for linking patent data to UMETRICS employee data.

Table 1. Number of Individual Names Used for Linkage (and Number of Associated Patents)

UMETRICS 2017Q4a Employee Name and Transaction Files	Total number of unique employees whose names are provided by IRIS member universities	428,700
Patent Subset (Linkage Source) File	Total number of unique inventor names	25,795
	Total number of unique patents	20,276

Linking Elements

Table 2. Linking Elements

UMETRICS		US PTO Patent
Employee Last Name Employee Middle Name Initial Employee First Name IRIS Member University Name	⟷	Inventor Last Name Inventor Middle Name Initial Inventor First Name Assignee Organization Name

Methodology

For blocking, we used the two data fields that are common to two datasets (university ID and an initial of employee's / inventor's last name). Within each block, other common fields (individual's first name, the initial of middle name and last name) were used to match records. In linking names, we applied one of the often-used string comparison algorithms, Jaro-Winkler, to compare each name string element, using R and its package "RecordLinkage." In particular, during the process, we used "RLBigDataLinkage" designed for large numbers of pairs in the datasets. For more details, please see the manual available from CRAN (see *Package 'RecordLinkage'*: https://cran.r-project.org/web/packages/RecordLinkage/RecordLinkage.pdf).

In evaluating string comparison results (to filter out unmatched pairs), we set a threshold of 0.95 as suggested in the aforementioned *Package 'RecordLinkage'* reference manual.

Findings

The descriptive statistics of the data in input and output files are shown in Table 3 below. There are 23 IRIS member universities in both input and output files.

Table 3. UMETRICS-Patent Linkage: Input-Output Data Summary Statistics

Input/Output	Type	Data Elements	Total Obs.	Mean	Std. Dev.	Min	Max
INPUT	UMETRICS	Unique individuals in UMETRICS 2017Q4a employee transaction file (whose names are available)	428700	18639.13	14745.14	1386	63837
	Patent (subset file)	Unique inventors	25795	1121.522	744.5553	22	3053
		Unique patents	20276	881.5652	650.5068	8	2922
OUTPUT	Employees	Unique pairs (whose names are matched)	9417	409.4348	510.1228	5	2045
	Inventors	Unique inventors in matched pairs	8015	348.4783	399.8675	5	1600
	Patents	Unique patents in matched pairs	12414	539.7391	537.5094	4	2232

Future Work

- Incorporating award source and government interest information
- Incorporating research topic and patent classification
- Improving data quality in the source file, i.e., challenges associated with assignee organization name alias

About This File

File Details

File Name: Patent Subset
Date Created: June 2018
Record Counts: 58,766
Field/Column Counts: 5

File Summary

Data Fields

institution_id
inventor_id
patent_number
app_year
app_date

Data field descriptions are in Appendix B.

Patent Subset (Linkage Source) File

2018 Release Notes

Publicly available patent data were downloaded from PatentsView (http://www.patentsview.org/download/) in February 2018 and were linked to UMETRICS employee data. A portion of the downloaded patent data was used to create a source file that includes patent assignee organization names associated with current IRIS universities. The crosswalk includes 23 universities—certain university data were not included in the patent source file (thus not linked to UMETRICS) due to the timing of matching (February 2018) and the availability of employee name files from the universities.

The data pre-processing began by downloading publicly available patent data from the USPTO PatentsVeiw website, including variables such as inventor ids, inventor names, patent numbers, and application dates. Then we filtered irrelevant data from the entire patent inventor and patent datasets in order to create a subset file of patent data which served as a linkage source file. In filtering, we first built a lookup table that contains patent assignee organization names that are relevant to IRIS member universities. A group of assignee organization names were carefully reviewed and compiled. This list contains 188 different patent assignee organization names that are mapped to 23 IRIS member universities.

Once data retrieval was complete, we applied data pre-processing to both input files (UMETRICS employee and patent subset files). All strings were standardized by removing special characters (e.g., whitespaces, punctuations, a mix of lower and upper cases, etc.).

About These Files

Files downloaded: June 2018

Number of Files: 14

List of Files:

1. Application
2. Assignee
3. Government Interest
4. Inventor
5. Main class
6. Other Reference
7. Patent
8. Patent-Assignee
9. Patent-Inventor
10. Patent-Contract Award Number
11. Patent Citation
12. USPC Current
13. WIPO
14. WIPO Field

Data file and field descriptions are in Appendix C.

Files publicly available for download at:
http://www.patentsview.org/download/

USPTO PatentsView Patent Data

2018 Release Notes

In addition to the crosswalk and source files, IRIS provides researchers within our virtual data enclave a total of 14 relational tables that are available from PatentsView (e.g., patent application, assignee, inventor, abstract, and citation records). Although more than 50 files are available from PatentsView, these 14 files include most important elements which we hope should enable researchers to do additional linkage work. For de-identification purposes, IRIS did not download location files and also dropped personally identifiable information fields from some files.

Appendix A. UMETRICS-Patent Crosswalk Data Field Description

Field Name	Data Type	Max Length	Field Definition
Patent Number	Char	50	Unique patent number
Inventor ID	Char	50	Unique inventor ID
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

Appendix B. Patent Subset (Linkage Source) Data Field Description

Field Name	Data Type	Max Length	Field Definition
Patent Number	Char	50	Unique patent number
Inventor ID	Char	50	Unique inventor ID
Institution ID	Num	8	Unique identifier (a four or five digit number) assigned to each IRIS member university for de-identification purpose
App Year	Num	8	Year of patent application filing retrieved from application date
App Date	Num	8	Date of patent application filing

Appendix C. USPTO Patent Data File and Field Descriptions

All files were downloaded by IRIS staff from the US PatentsView website in February 2018 (except for the 'otherreference' file downloaded in June 2018). File and field descriptions were also downloaded from PatentsView (http://www.patentsview.org/data/Patents_DB_dictionary_bulk_downloads.xlsx) and are copied below without modification.

1. Application

Information on the applications for granted patent.

Data Element Name	Definition	Years Present	Type
id	application id assigned by USPTO	all	varchar(36)
patent_id	patent number	all	varchar(20)
series_code	application series; "D" for some designs; (http://www.uspto.gov/web/offices/ac/ido/oeip/taf/filingyr.htm)	all	varchar(20)
number	unique application identifying number	all	varchar(64)
country	country this application was filed in	all	varchar(20)
date	date of application filing	all	date

2. Assignee

Disambiguated assignee data

Data Element Name	Definition	Years Present	Type
id	unique assignee ID generated by the disambiguation algorithm	all	varchar(36)
type	classification of assignee (2 - US Company or Corporation, 3 - Foreign Company or Corporation, 4 - US Individual, 5 - Foreign Individual, 6 - US Government, 7 - Foreign Government, 8 - Country Government, 9 - State Government (US). Note: A "1" appearing before any of these codes signifies part interest)	all	varchar(10)

Note: Three fields ('name_first', 'name_last', and organization) are dropped for de-identification purposes.

3. Government Interest

Mapping of patent numbers to raw government interest text

Data Element Name	Definition	Years Present	Type
patent_id	patent number	all	varchar(255)
gi_statement	raw government interest text	all	text

4. Inventor

Disambiguated inventor data

Data Element Name	Definition	Years Present	Type
id	unique inventor ID generated by the disambiguation algorithm	all	varchar(36)

Note: Two fields ('name_first' and 'name_last') are dropped for de-identification purposes.

5. Main class

Metadata for USPTO technology classes at patent issue date

Data Element Name	Definition	Years Present	Type
id	ID of the USPC mainclass at issue	all	varchar(20)

6. Other Reference

Citations made to non-patent documents by US patents

Data Element Name	Definition	Years Present	Type
uuid	unique id	all	varchar(36)
patent_id	patent number	all	varchar(20)
text	non-patent literature reference text	all	text
sequence	order in which this reference is cited by patent	all	int(11)

7. Patent

Data concerning granted patents

Data Element Name	Definition	Years Present	Type
id	patent this record corresponds to	all	varchar(20)
type	category of patent. Usually "Design", "reissue", etc.	all	varchar(20)
number	patent number	all	varchar(64)
country	country in which patent was granted (always US)	all	varchar(20)
date	date when patent was granted	all	date
abstract	abstract text of patent	all	text
title	title of patent	all	text
kind	WIPO document kind codes (http://www.uspto.gov/learning-and-resources/support-centers/electronic-business-center/kind-codes-included-uspto-patent)	all	varchar(10)
num_claims	number of claims	all	int
filename	name of the raw data file where patent information is parsed from	all	varchar(120)

8. Patent-Assignee

Crosswalk between patent and assignee tables

Data Element Name	Definition	Years Present	Type
patent_id	Patent number	all	varchar(20)
assignee_id	Unique assignee ID generated by the disambiguation algorithm	all	varchar(36)

9. Patent-Inventor

Crosswalk between patent and inventor tables

Data Element Name	Definition	Years Present	Type
patent_id	Patent number	all	varchar(20)
inventor_id	Unique inventor ID generated by the disambiguation algorithm	all	varchar(36)

10. Patent-Contract Award Number

Mapping of Federal contract award numbers to patent numbers

Data Element Name	Definition	Years Present	Type
patent_id	patent number	all	varchar(255)
contract_award_number	Federal contract award number	all	varchar(255)

11. Patent Citation

Citations made to US granted patents by US patents

Data Element Name	Definition	Years Present	Type
uuid	unique id	all	varchar(36)
patent_id	patent number	all	varchar(20)
citation_id	identifying number of patent to which select patent cites	all	varchar(20)
date	date select patent (patent_id) cites patent (citation_id)	all	date
name	name of cited record	all	varchar(64)
kind	WIPO document kind codes (http://www.uspto.gov/learning-and-resources/support-centers/electronic-business-center/kind-codes-included-uspto-patent)	2002 and After	varchar(10)
country	country cited patent was granted (always US)	all	varchar(10)
category	who cited the patent (examiner, applicant, other etc.)	2002 and After	varchar(20)
sequence	order in which this reference is cited by select patent	all	int

12. USPC Current

USPTO current patent classification

Data Element Name	Definition	Years Present	Type
uuid	unique id	all	varchar(36)
patent_id	patent number	all	varchar(20)
mainclass_id	uspc mainclass current	all	varchar(10)
subclass_id	uspc subclass current	all	varchar(10)
sequence	order in which uspc class appears in patent file	all	int

13. **WIPO**

WIPO technology classification of the patent

Data Element Name	Definition	Years Present	Type
patent_id	patent number	all	varchar(20)
field_id	WIPO technology field ID as derived from crosswalk http://www.wipo.int/export/sites/www/ipstats/en/statistics/patents/xls/ipc_technology.xls	all	int(10)
sequence	order in which WIPO technology field appears on patent	all	int(10)

14. **WIPO Field**

WIPO technology classification

Data Element Name	Definition	Years Present	Type
id	WIPO technology field ID as derived from crosswalk http://www.wipo.int/export/sites/www/ipstats/en/statistics/patents/xls/ipc_technology.xls	all	int(10)
sector_title	WIPO technology sector title	all	varhcar(60)
field_title	WIPO technology field title	all	varhcar(60)