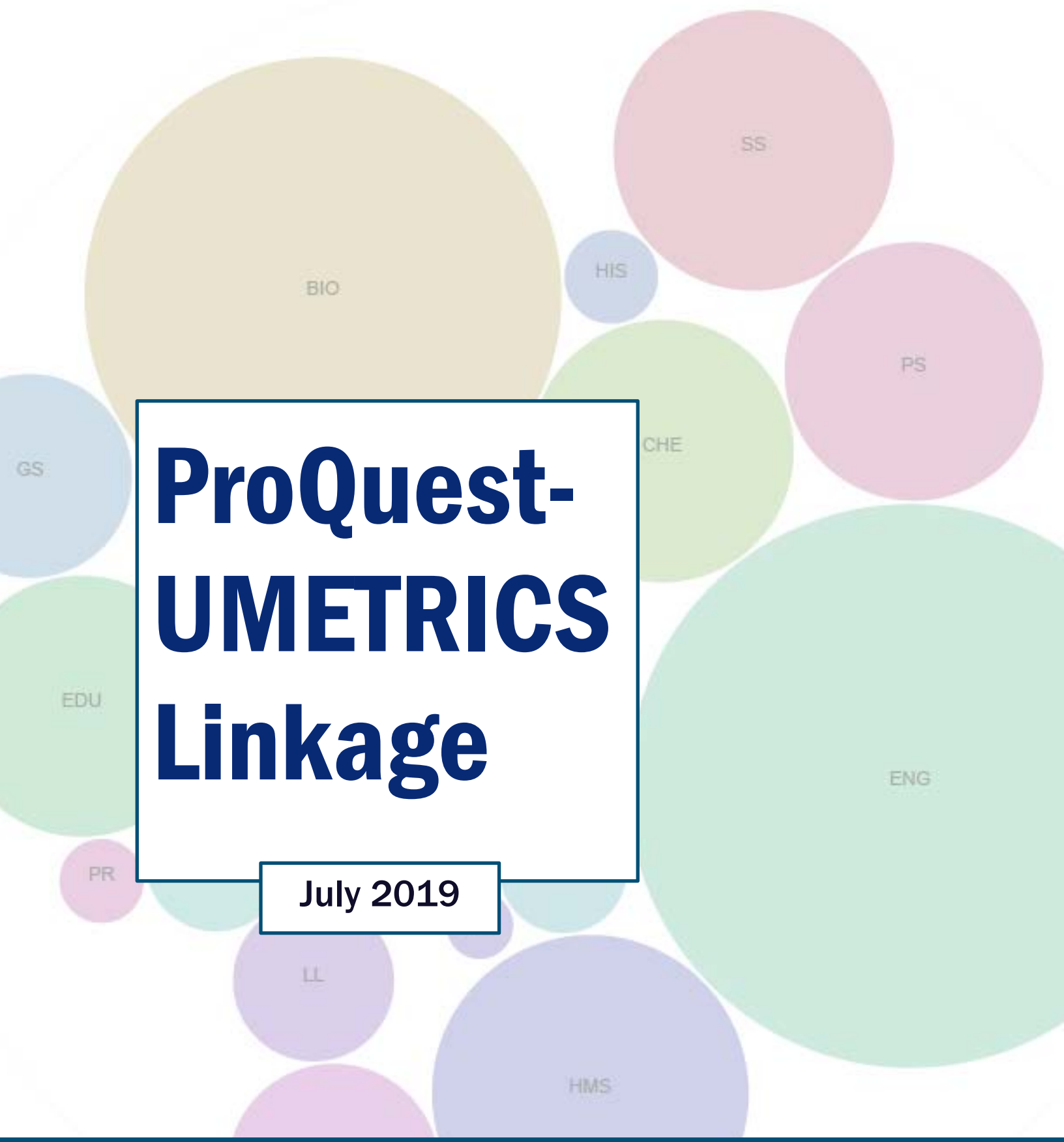


ProQuest-UMETRICS Linkage

July 2019

IRIS INSTITUTE FOR RESEARCH ON INNOVATION & SCIENCE

Citation

VanEseltine, M., Nicholls, N., Ku, R. L., & Owen-Smith, J. (2019). ProQuest-UMETRICS Linkage Documentation. Institute for Research on Innovation and Science (IRIS). https://doi.org/10.21987/T556-XZ77

Contents

Tables and Figures

ProQuest-UMETRICS Crosswalk

File Details

File Name: link_proquest_xwalk
Date Created: 7/15/2019
Record Counts: 53,284
Field/Column Counts: 14

File Summary

This file includes match results of the crosswalk between UMETRICS employee names, employee transaction records, and ProQuest publication (thesis / dissertation) data with a focus on publication 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, publication IDs originally from the ProQuest database are not made available.

Data Fields

Fields appearing across files	Fields unique to this file
institution_id	sequential_num
emp_number	link_score
	degree_year
	degree
	degree_type
	subjects
	cleaned_subjects
	number_of_subjects
	first_fineagg
	first_coarseagg
	common_fineagg
	common_coarseagg

2019 Release Notes

This supplementary release in the summer of 2019 produces the new results from linking UMETRICS employee transaction records to ProQuest dissertation data with a focus on dissertation subjects. 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. We are releasing one file that contains the ProQuest subject classification associated with theses authored by matched UMETRICS individuals (i.e., individuals who were recorded as being paid by research awards while at an IRIS university). The file includes 52,474 employees (uniquely counted by an IRIS-generated employee ID) linked to 53,284 thesis and dissertation publication records from ProQuest.[1] Not all records are a one-to-one match.

This year we made major changes in the process of linking individual names that appear in the UMETRICS and ProQuest dissertation datasets. Once we completed the initial linkage to individual names, using the crosswalk, we then applied the subject classification method that was developed by Dr. Bruce Weinberg at The Ohio State University. This year's subject classification includes only limited revision.

This year we included a few new fields in the ProQuest-UMETRICS crosswalk file, including:

1) **degree** (e.g., PhD, MA, MD, etc.) recorded originally in ProQuest data;
2) **degree type** flag that identifies a doctoral degree, not-doctoral degree, or degree type unknown;
3) **number of subjects** assigned to a dissertation;
4) **cleaned and standardized subject areas**.

The degree_type flag helps us identify the distribution of degree types. Of 53,284 publications, about 86% are doctoral degrees and 14% are master's degrees.

[1] 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, the ProQuest raw data are not included in this supplementary release file.

The number of matched records has dramatically increased in this release partly due to increased member universities but mostly due to the application of better matching methodology. Data changes are shown in Table 1.

Table 1. Data Changes from Previous Releases

Data Element	2017 Release	2018 Release	2019 Release	Change Rate from Last Release
Number of Universities (that provide IRIS with individual names for matching)	16	22	28	+27%
Number of matched thesis / dissertation records in ProQuest	13,660	28,725	53,284	+85%

Linkage and Subject Classification

Linking Methodology and Process

The ProQuest-UMETRICS linkage matches employees to the dissertations and theses at their employing institutions. An exact match between full employee names and thesis records does capture a large proportion of links, but it is not sufficient due to name variations and non-unique names. We probabilistically match the UMETRICS personally identifying information (PII) of employees to the PII contained in the ProQuest Dissertations and Theses Global (PQDT Global) database. Through a supervised machine learning process, we identify high-probability matches to be selected for the final linkage.

Preprocessing

Name Parsing

The ProQuest database stores full names, and UMETRICS captures first, middle, and last names separately. In order to produce the most consistent separation of fields, the UMETRICS fields are concatenated and both databases are run through the same name parsing process to separate first, middle, last, suffixes, and nicknames.

ProQuest Dissertations and Theses Global

Duplicate and multiple records in PQDT Global are handled in stages to arrive at a unique author to link to UMETRICS. First, 250,346 raw records are deduplicated to select a single canonical record for each of the 244,023 publication numbers. In order to estimate one-to-one record linkage across the databases, an ad-hoc author id was created. Using the Python package *dedupe*[2], the 244,023 unique publications were condensed into one author per row with combined thesis title and subject information, with a final *n* of 242,316.

[2] Available at: https://dedupe.io/.

UMETRICS

Some UMETRICS employees have multiple versions of names recorded. The row within each employee id that provides the longest name record (i.e., the most letters in the complete concatenated name) is used as the canonical UMETRICS record for this linkage (prior to parsing). The final count was 1,019,665 employees as input for the matching process.

Learning

We also use *dedupe* to link individuals across datasets using supervised pair learning. The table below shows the fields cleaned and selected for linkage. The active learning process determines the most predictive matching predicates to use in final match estimates.

Table 2. Dataset Fields Compared for Linkage

UMETRICS (n = 1,019,665)	**ProQuest (n = 242,316)**
Full name	Full name
First name	First name
Middle name	Middle name
Last name	Last name
Name suffix(es)	Name suffix(es)
Nickname	Nickname
Year of earliest observed employment	Year of earliest observed thesis
Year of most recent employment	Year of most recent thesis
Paying grant award titles and funders	Thesis titles and subject areas

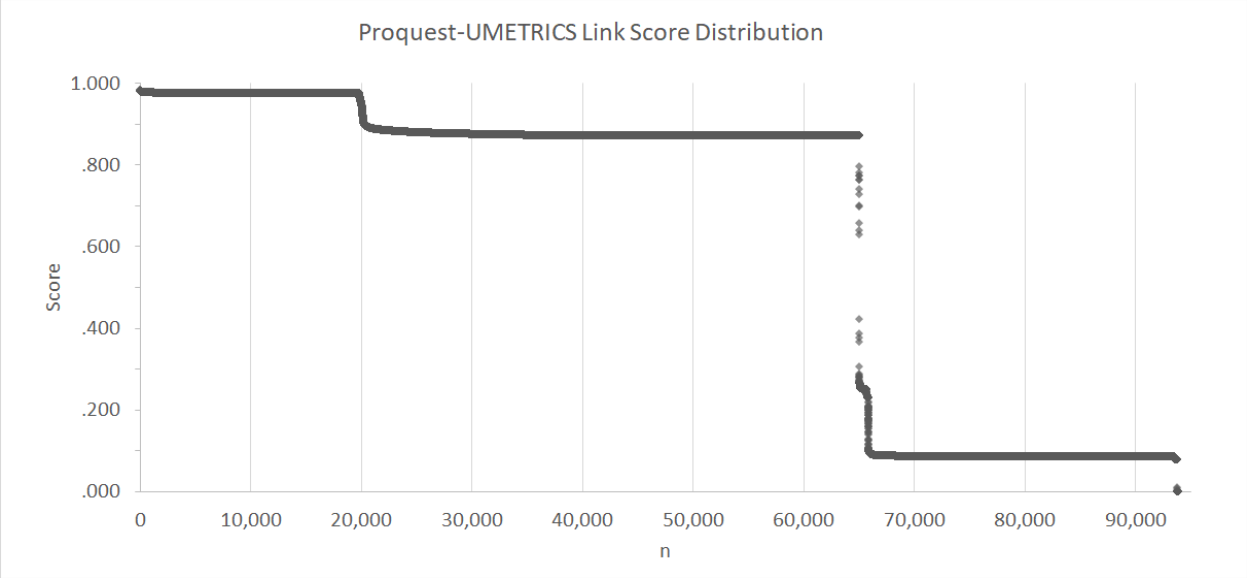
The supervised learning involved assessing 48 pairs of records selected from the global inputs, 30 of which were identified as matches and 18 of which were identified as distinct. Following

supervised learning, the final matching process was run separately within each UMETRICS institution. No other hard rules or blocking criteria were used.

The *dedupe* package implements a long list of possible string and numeric comparators from which to select the most predictive predicates, which can compare fields and combinations of fields using a range of different algorithms. For details, please see the *dedupe* documentation at https://dedupe.io/documentation/.

Results

Figure 1. ProQuest-UMETRICS Link Score Distribution



This figure displays potential links and scores generated by *dedupe* for the final linkage results. The dropoff between 60,000 and 70,000 indicates a strong break between the high-probability (0.87 and higher) and low-probability (0.29 and lower) matches, which suggests a good quality linkage. The model was able to identify many people very likely to be the same with relatively few unclear cases. Cleric review of surrounding records and samples of the full distribution supported this break, and we selected a final cutoff of 0.5. Employee-author pairs with link scores above 0.5 were matched, and the ad-hoc ProQuest author identifiers were crosswalked back to publication numbers.

The final count matches 52,474 UMETRICS employees to 53,284 ProQuest publication numbers.

Subject Classification

ProQuest provides subject area tags for dissertations, representing one or more fields of study covered by the dissertation. For the 53,284 entries linked from the ProQuest raw data to UMETRICS employee records, we extracted the subjects for each dissertation. Again, note that the number of publications (53,284 entries) in this file is slightly larger than uniquely counted 52,474 employees in this crosswalk file because some appear in ProQuest data more than once for multiple degrees and their different degree-related publications.

To extract said subjects from ProQuest raw data, a lookup table for standardizing duplicate subject names produced by Dr. Bruce Weinberg's team at The Ohio State University was utilized, which used ProQuest's 2015-2016 academic subject list as reference.[3] See Table 3 for an illustration of the steps undertaken.

Preprocessing

Step 1. Pre-Processing ProQuest Data

For dissertations tagged under multiple subject areas, ProQuest provides a string of all subjects covered by the dissertation with subjects separated by a period (“.”). Each subject is extracted from the provided string, and set to lowercase for standardization.

Step 2. Subject name standardization and updating reference files

The ProQuest raw data sometimes lists the same subject in different ways (e.g., “biology, microbiology” and “microbiology”). Due to additional subjects in the ProQuest Subject Category list for the 2018-2019 academic year,[4] an additional step was taken to identify which of the dissertations contained subjects not included in the lookup table. Subjects not included in the lookup table were added and categorized. For subjects not in the ProQuest list, the broadest subject category similar to the subject name in the ProQuest list was applied.

[3] See: http://corpweb.proquest.com/assets/etd/umi_subjectcategoriesguide.pdf
[4] See: https://media2.proquest.com/documents/subject-categories-academic.pdf

Table 3. Subject Processing Steps

	"subjects" field	Extracted subjects	Set to lowercase	Name standardization	Fine aggregate fields assigned	Coarse aggregate fields assigned
Example 1	GLBT Studies.	GLBT Studies	glbt studies	lgbtq studies	Area, Ethnic, and Gender Studies	Social and Behavioral Sciences
Example 2	Political Science, General. Mass Communications.	Political Science, General	political science, general	political science, general	Social Sciences	Social and Behavioral Sciences
		Mass Communications	mass communications	mass communication	Communication and Information Sciences	Social and Behavioral Sciences
Example 3	Chemical Engineering. Engineering, Biomedical. Agricultural Engineering.	Chemical Engineering	chemical engineering	engineering, chemical	Engineering	Mathematical and Physical Sciences
		Engineering, Biomedical	engineering, biomedical	engineering, biomedical	Engineering	Mathematical and Physical Sciences
		Agricultural Engineering	agricultural engineering	engineering, agricultural	Agriculture	Natural Sciences

Step 3. Assigning Aggregate-level Subjects

Each subject was assigned a fine aggregate field (referred to as "FineAgg"), representing one of 21 subject areas taken from the ProQuest Subject Category List for the 2018-2019 academic year. Note that the "Mathematical and Physical Sciences" subject area in the ProQuest Subject Category list was further subdivided into the "Mathematics", "Physical Sciences", and "Chemistry" categories, resulting in 23 distinct categories for the FineAgg classification.

Additionally, a coarse aggregate field (referred to as "CoarseAgg"), representing 13 subject fields was generated from the FineAgg classification, as shown in Table 4.

Table 4. Fine and Coarse ProQuest Subject Aggregations

FineAgg	CoarseAgg
Agriculture	Agriculture
Fine and Performing Arts	Arts and Humanities
History	
Language and Literature	
Philosophy and Religion	
Behavioral Sciences	Behavioral
Biological Sciences	Biology
Chemistry	Chemistry
Ecosystem Sciences	Earth
Environmental Sciences	
GeoSciences	
Education	Education
Architecture	Engineering
Engineering	
Health and Medical Sciences	Health Medical
Mathematics	Math
Communication and Information Sciences	Other
Law and Legal Studies	
Physical Sciences	Physical
Area, Ethnic, and Gender Studies	Social
Business	
Interdisciplinary	
Social Sciences	

Step 4. Identify first and common aggregate fields

Lastly, we generated two fields to represent the coarse and fine aggregate field corresponding to the first subject area assigned to a dissertation.

Additionally, there are two fields that represent the coarse and fine aggregate field corresponding to the most common coarse and fine aggregate field from the list of subjects for a specific dissertation.

Results

Tables 5 and 6 show some characteristics of publications whose authors are matched employees in UMETRICS employee transactions. Table 5 shows the subject distribution by number of assigned subjects per publication ranging between 1 and 9; Table 6 shows most-covered subject areas using the 23 most common fine subject fields.

Table 5. Subject Distribution (Number of subjects listed per paper)

Number of ProQuest Subjects Listed	Number of Dissertations
1	19462
2	16946
3	14139
4	2163
5	440
6	90
7	29
8	13
9	2

Table 6. Most-covered ProQuest Subject Areas Based on Most Common Fine Aggregate Field

ProQuest Subject Area (common_fineagg)	Number of Dissertations
Engineering	13640
Biological Sciences	9776
Health and Medical Sciences	4303
Social Sciences	3369
Chemistry	2956
Education	2922
Physical Sciences	2867
Behavioral Sciences	2268
Mathematics	1925
GeoSciences	1780
Agriculture	1697
Language and Literature	1111
Environmental Sciences	859
Fine and Performing Arts	694
Communication and Information Sciences	662
Area, Ethnic, and Gender Studies	639
Ecosystem Sciences	475
Business	393
History	375
Philosophy and Religion	301
Interdisciplinary	190
Architecture	73
Law and Legal Studies	9

Table 7. Common Fine and First Fine Subjects for Matched PQ Dissertations

FIRST FINE SUBJECT \ COMMON FINE SUBJECT	AGR	ARC	AGE	BS	BIO	BUS	CHE	CIS	ECO	EDU	ENG	ENV	FPA	GS	HMS	HIS	ID	LL	LLS	MAT	PR	PS	SS	Total
AGR	**1622**	0	5	1	101	1	3	0	0	6	11	15	0	3	7	1	2	0	0	0	0	0	8	1786
ARC	1	**72**	1	0	0	0	0	0	0	0	5	0	0	0	0	2	1	0	0	0	0	0	4	86
AEG	3	0	**359**	3	0	0	0	1	0	3	1	3	3	0	1	1	1	6	0	0	3	0	12	400
BS	0	0	21	**2116**	12	0	0	2	0	47	2	1	0	0	26	0	0	1	0	0	0	0	21	2249
BIO	31	0	1	46	**9052**	0	24	0	0	0	43	6	0	5	128	0	0	0	0	4	0	4	3	9347
BUS	1	0	1	2	0	**378**	0	2	0	3	7	0	0	0	0	0	0	0	0	0	0	0	10	404
CHE	1	0	0	0	61	0	**2880**	0	0	7	30	2	0	8	10	0	1	0	0	0	0	22	0	3022
CIS	0	0	17	10	1	2	0	**634**	0	13	3	0	0	0	7	1	2	2	0	0	1	0	18	711
ECO	11	0	0	0	45	0	0	0	**475**	0	0	13	0	11	2	0	0	0	0	0	0	0	0	557
EDU	0	0	33	17	0	1	0	2	0	**2731**	0	1	0	1	11	2	1	6	1	0	1	0	14	2822
ENG	2	0	1	7	159	0	31	6	0	16	**13404**	19	0	35	58	0	6	0	0	11	0	113	10	13878
ENV	12	0	0	1	9	0	3	0	0	1	6	**741**	0	29	21	0	5	1	0	0	1	0	10	840
FPA	0	0	18	1	0	0	0	1	0	2	4	0	**683**	0	0	5	1	7	0	0	3	0	6	731
GS	2	0	0	0	17	0	5	0	0	2	7	24	0	**1663**	0	0	0	0	0	2	0	6	0	1728
HMS	5	0	19	28	263	2	5	4	0	9	15	7	0	1	**3977**	0	0	1	0	1	0	0	17	4354
HIS	0	0	14	0	0	1	0	0	0	4	0	0	1	0	2	**346**	0	0	0	0	1	0	15	384
ID	1	0	1	0	1	1	1	1	0	2	7	2	0	2	7	0	**166**	0	0	0	0	0	7	199
LL	0	0	47	10	0	0	0	1	0	30	0	0	5	0	0	3	1	**1080**	0	0	6	0	4	1187
LLS	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	**8**	0	0	0	3	14
MAT	0	0	0	0	22	0	0	0	0	1	6	1	0	3	3	0	0	0	0	**1906**	0	5	1	1948
PR	0	0	9	2	0	0	0	0	0	3	0	0	0	0	0	4	0	3	0	0	**283**	1	1	306
PS	0	0	0	1	22	0	4	0	0	5	83	2	0	9	7	0	1	1	0	0	0	**2716**	0	2851
SS	5	1	91	23	11	7	0	7	0	37	6	21	2	10	36	10	2	3	0	1	2	0	**3205**	3480
Total	1697	73	639	2268	9776	393	2956	662	475	2922	13640	859	694	1780	4303	375	190	1111	9	1925	301	2867	3369	**53284**

Table 8. List of Subject Acronyms

Acronyms Used in Table 7	Subject
AGR	Agriculture
ARC	Architecture
AEG	Area, Ethnic, and Gender studies
BS	Behavioral Sciences
BIO	Biological Sciences
BUS	Business
CHE	Chemistry
CIS	Communication and Information sciences
ECO	Ecosystem Sciences
EDU	Education
ENG	Engineering
ENV	Environmental Sciences
FPA	Fine and Performing arts
GS	GeoSciences
HMS	Health and Medical Sciences
HIS	History
ID	Interdisciplinary
LL	Language and Literature
LLS	Law and Legal Studies
MAT	Mathematics
PR	Philosophy and Religion
PS	Physical Sciences
SS	Social Sciences

Data Dictionary

Table 9. 'link_proquest_xwalk' Data Fields

Field Name	Column Name	Data Type	Set Length	Max Length	Field Definition
Dissertation Sequential Number	sequential_num	int	8	8	IRIS-generated sequential number assigned to individual ProQuest (thesis / dissertation) publication records selected in this crosswalk table; this number is not an original publication number or dissertation ID generated by ProQuest
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
IRIS Employee Number	emp_number	varchar	100	32	IRIS-generated unique identifier assigned to all personnel being paid by awards; this is the same employee ID found in the Employee transaction file released as part of the UMETRICS Core Collection.
Link Score	link_score	decimal	13	5	Score generated for linkage result
Degree Year	degree_year	int	4	4	The year, in the form of the 4-digit year (e.g., "2010"), when each thesis / dissertation was submitted and accepted for a degree; as recorded in ProQuest
Degree	degree	varchar	50	12	Degree awarded (e.g., PhD, MA, MD) as recorded in ProQuest
Degree Type	degree_type	int	4	4	IRIS-generated flag indicating whether a degree awarded is a doctoral degree or not; 1=doctoral degree, 0=not doctoral degree (i.e., master's

					degree), 99=unknown or unclear
Subjects	subjects	varchar	500	219	Subject(s) covered by a thesis / dissertation, as tagged by ProQuest
Cleaned Subjects	cleaned_subjects	varchar	500	233	Subject(s) as cleaned by IRIS to standardize names, case, etc.
Number of Subjects	number_of_subjects	int	4	4	Number of subjects tagged by ProQuest for a thesis / dissertation
First Fine Aggregated Subject	first_fineagg	varchar	200	38	Representing the first subject area assigned by ProQuest to a thesis / dissertation, as categorized into one of 23 distinct categories from the ProQuest Subject Category list for 2018-2019
First Coarse Aggregated Subject	first_coarseagg	varchar	200	34	Representing the first subject area assigned by ProQuest to a thesis / dissertation, as categorized into one of 13 broader categories generated from the fine aggregated subject classification (e.g., the FineAgg categories of “Area, Ethnic, and Gender Studies” and “Social Sciences” are both assigned to the CoarseAgg category of “Social and Behavioral Sciences”).
Common Fine Aggregated Subject	common_fineagg	varchar	200	38	The most common fine aggregated subject from the list of all subjects assigned for a specific thesis / dissertation
Common Coarse Aggregated Subject	common_coarseagg	varchar	200	34	The most common coarse aggregated subject from the list of all subjects assigned for a specific thesis / dissertation