A Appendix: Nuts and Bolts

A.1 Data fields

Across all of the record types there are roughly 400 data fields and subfields, some of which – such as the distinct works included in a single copyright registrations - can repeat within a given record. Many of these fields are only applicable to a narrow subset of the records or are sparsely populated for other reasons. Table A.1 below provides an abridged accounting of data fields and the frequency with which they are populated, stratified by record type. Note that the frequency with which a data field is populated can vary over time and across types of works. For example, as a result of a change in the Copyright Office's data management platform, claimant addresses are almost never populated for pre-2007 registrations, while they are frequently available for post-2007 registrations. For a full list and description of the MARC 21 datafields that appear in the unabridged, long-form data, see https://www.copyright.gov/policy/women-in-copyright-system/LOC-Copyright-Data-as-Distributed-in-the-MARC%2021-Format.pdf. For an accounting of the variables that appear in our processed, wide-form data see the code book provided in Table A.2

A.2 Tips for using the Copyright Office data

A.2.1 Repeated data fields

The raw Copyright Office data, in their unabridged form, are not amenable to a wide-form data table format since many of the data fields are arbitrarily repeatable. For example, a musical works registration may pertain to a compilation album with dozens, or even hundreds of titles, and each of those titles may have multiple authors. The data will then contain repeated author fields for each author. It will also contain repeated entries for related fields, like author date of birth, author domicile, author citizenship, and so on. A low percentage of registrations have fields that are repeated more than 10 times, but some can have enough repeated fields to result in many thousands of columns. When these fields (usually relating to authors, claimants, and titles) are not relevant to the task at hand, it is advisable to drop them or otherwise limit the number of repeated fields before converting the data to wide-form (as we did in the tabular version of the registrations data). Otherwise, it is advisable to work with the data in long-form, to the extent possible.

A.2.2 Identifying record types and classifying types of works

Several data fields contain information about the type of record (e.g., recordation, registration, renewal, etc.) and the type of work to which the record pertains (e.g., sound recording, musical work, film, etc.), and these are sometimes intermingled. The most reliable field for identifying the type of record is the record leader, in which the record type is indicated by the two digit alphabetic code starting at the 6th character. Other fields that contain information about record

type are not comprehensive (e.g., identifying recordations, but with no distinction between registrations, renewals, cancelsations, and preregistration for the remainder of the records). In both versions of the data we have constructed, the type of record is identified primarily using information in the record leader, supplemented by other datafields when necessary. Researchers using the raw data will need will need to classify the records on their own and should note that relying on anything but the leader may result in misclassifying record types.

The most granular and reliable description of the type of work can be found in data field "917," subfield "b." For registration records, when field 917b is missing, the two-digit alphabetic code starting at the 6th character of the record leader provides information about the type of work. However, the categories of works annotated in the leader do not precisely correspond to the more granular categories in 917b. To ensure the best possible matching between these two categorization schemes, it is helpful to cross tabulate 917b with the leader code across the records in which 917b is populated. Importantly, 917b does not indicate the type of record. For example, it may reveal that the record pertains to a sound recording, but it will not distinguish between a registration, renewal, cancellation, or preregistration. Additional information on the type of work is provided in control field 001; however, this is a much less granular categorization scheme and is not particularly useful.

A.2.3 Dates

A given record may include several date fields; often, the dates across these fields are the same. When they are not the same, it is advisable to defer to the documentation for clarity on the information each field is providing. It may be the date of creation, the date of first publication, the date of anticipated publication, the date of registration, the date of preregistration, the date on which the registration took effect, the beginning or end date for a serial publication, or the date of first commercial exploitation. Additionally, the format of dates can differ across data fields, and occasionally, within a data field. The format is usually identified in the "indicator" values associated with a given entry for a data field. The control fields (CF) and data fields (DF) that contain dates relating to the genesis of a work include: CF008, DF017, DF046, DF260, DF263, DF269, DF362, DF779, and DF917. Which date is appropriate depends on one's empirical goals and the subset of records under review; nonetheless, the various subfields of DF269 tend to be reliably populated and well-documented.

Note that these complexities have been largely simplified in the cleaned tabular version of the data we constructed. In that version, date formats are standardised based on the format indicators present in the raw data and split between unambiguous descriptive headings (such as "registration date" or "publication date").

A.3 How to access the data

Three versions of the data are available: i) the raw xml data originally released by the U.S. Copyright Office; ii) a version that parses the xml data into plain text .csv files organized by type of record and type of work; and iii) a cleaned tabular version of registrations data.

The raw dataset was published by the U.S. Copyright Office in conjunction with its Women in the Copyright System Report, which made use of the data. [52] This raw version is a readout from the Library of Congress' internal database, which is maintained in the MARC 21 standard for bibliographic information.[53] It is contained within 26 .xml files.[54] Because these files do not provide the xml schema, it is advisable to treat them as plain text files and parse out the xml tags.[55] The data fields and the information therein are characterized by four attributes. The top level is a numerical identifier ranging from 001 to 991 identifying the data field or control field.[56] The next is a single alphanumeric character indicating the subfield within the data field. Finally, there are two separate "indicator" values that can provide a range of information (e.g. the source of information provided in the data field, the format of the data, etc.). Importantly, some fields are repeatable, but each repetition will have its own set of indicator values. Additionally, each record starts with a "leader" entry which provides information about the layout and contents of that record.

The Library of Congress provides a number of resources for understanding MARC 21 records. We have found two to be particularly useful. The first is copyright-specific documentation. The MARC 21 format is a general bibliographic format; however, copyright records are not always necessarily bibliographic in nature, and thus specialized documentation is required in order to interpret those records. Such documentation is provided in "Library of Congress Copyright Data as Distributed in the MARC 21 Format."[57] However, this document does not always provide the details necessary to understand how information is spread across different subfields. For that we use a second document with supplemental information: "MARC 21 Format for Bibliographic Data Field List."[58] This second document is not specific to copyright records, but nonetheless provides comprehensive details about each field and subfield. When these two documents provide conflicting information, it is advisable to defer to the former (which is copyright-specific) unless the context of the data clearly supports the latter.

The second version of the data we provide seeks to make the data more accessible while maintaining all of the information contained in the raw data. These data are organized into 31 plain text files; one each for recordations, preregistrations, cancellations, and renewals, with another 27 for registrations, divided by the type of work registered (e.g., sound recordings, motion pictures, photos, etc.).[59] Each file contains three columns; a record ID number, the field name, and the field value. The record ID uniquely identifies a record within a file and can be used to convert the data to wide form. A field name reflects the data or control field number, the two indicator values, and the subfield identifier, followed by a count indicating the repetition

[52] See https://www.copyright.gov/economic-research/usco-datasets/ (under "Raw Unparsed Uncategorized XML").

[53] See https://www.copyright.gov/policy/women-in-copyright-system/LOC-Copyright-Data-as-Distributed-in-the-MARC%2021-Format.pdf.

[54] Records in these files reflect the order in which they exist in the Library of Congress' database, and do not necessarily reflect a chronological or typological order.

[55] It is possible to impose a schema and access these files as xml files; however, this runs the risk of inadvertently misspecifing the data's hierarchical structure.

[56] Control fields range from 001 to 008 and data fields range from 017 to 991.

[57] See Seehttps://www.copyright.gov/policy/women-in-copyright-system/LOC-Copyright-Data-as-Distributed-in-the-MARC%2021-Format.pdf.

[58] See https://www.loc.gov/marc/bibliographic/ecbdlist.html.

[59] See https://www.copyright.gov/economic-research/usco-datasets/ (under "Raw Parsed CSV").

number for that particular field within a given record.[60] The field value reflects what is recorded in the raw data file; it has not been cleaned or modified in any way.

The third version of the data is, as previously discussed, abridged, cleaned, transformed to wide-form, and given descriptive variable names. These data are organized into 27 files based on type of work being registered.[61] In order to document the cleaning process and facilitate modifications to it, we provide a replication package with Stata files for first converting the raw data to the unabridged parsed data, and second converting that to the wide-form cleaned data. This replication package is available at https://github.com/watsonj-umn/USCO-data-replication.

[60] For example, df_046_sf_m_i1_1_i2_na__1 corresponds to data field 046 (local date of creation), subfield "m" (beginning valid date), a first indicator value of "1" (indicating that the type of entity for this observation is a "work") and no applicable second indicator. The final "1" reflects the fact that this is the first occurrence of this data field for this record. If there were a second entry for this field, the field name would be df_046_i1_1_i2_na_sf_m_2.

[61] See https://www.copyright.gov/economic-research/usco-datasets/ (under "Tabular").

A.4 Merging Copyright Registrations and Litigation with Compustat

Since our copyright records date back to 1978, we must account for variations in company names over time. To accomplish this, we match our copyright registration records and litigation records with company names recorded by CRSP (Center for Research in Security Prices), as Compustat does not provide a full history of company names for each entity in their database. We first link CRSP monthly data with associated Compustat entities using the CRSP/Compustat Merged Database provided by WRDS. Entities with non-relevant SIC and NAICS codes are omitted, and annotations (e.g., "(Last Known)" and origin country abbrevations) are removed from the company names. This linkage then provides a table of CRSP company names ("IssuerNm") linked to Compustat unique identifiers (gvkey) as well as CRSP identifiers (permco and permno). CRSP company names are further standardized by expanding common abbreviations. We then match this data to copyright records and litigation data using the process detailed below. The linked data and code is available at https://github.com/watsonj-umn/USCO-data-linking.

A.4.1 Rule-based Data Cleaning and Standardization

Company names from the Copyright Data are first cleaned and standardized to improve the accuracy of subsequent matching processes. Claimants are restricted to companies and other organizations appearing in the Copyright Data. Company names are preprocessed by converting them to ASCII characters using the unidecode Python library to handle accented characters and special symbols. Extra spaces and newline characters are removed with regular expressions to standardize formatting. Common acronyms and abbreviations are expanded to their full forms based on predefined mappings; for example, "IBM" is replaced with "International Business Machines". Additionally, common company suffixes such as "Inc.", "LLC", and "Ltd." are removed from the end of company names to aid in matching.

After preprocessing, each standardized company name is assigned a unique identifier by grouping identical names. This results in a cleaned and standardized dataset of company names with unique identifiers, which is ready for classification and deduplication.

A.4.2 Organization Type Classification using Few-Shot Learning

To classify organization names in the copyright registration data, we leverage few-shot learning using a pre-trained model (BAAI/bge-large-en-v1.5) that is then trained with hand-labeled company names. The training data contains 1,000 labeled names from U.S. copyright registrations, categorized into one of "business," "education," "government," "nonprofit," or "other." The training data contains company names (conm) and their corresponding types (label). All company names are converted to lowercase, and the dataset is split into training (60%), validation (20%), and test (20%) sets with stratified sampling to maintain the distribution of labels. Few-shot learning is conducted using the SetFit library, and the classifier is trained for one epoch. The classifier returns an accuracy of 96.5% on the test data. Once the model is trained,

it is used to predict labels (types) for the Copyright Data company and organization names from the previous step. These labels are assigned to the Copyright Data for use in further processing steps and downstream research. Subsequent processing is limited to organizations with predicted "business" labels.

A.4.3 Company Name Deduplication

Deduplication is performed to group similar company names and assign a representative name to each group. The standardized company names are prefixed with "The company name is " to form complete sentences suitable for pre-trained sentence embedding models. Embeddings for these sentences are generated using the BAAI/bge-large-en-v1.5 model. Single Linkage (SLINK) hierarchical clustering is then performed on the embeddings using cosine similarity as the distance metric. A threshold of 0.05 is used for clustering. Each cluster is assigned a unique cluster ID. For clusters containing multiple company names, the company with the highest registration count is selected as the representative label for the cluster. This results in a deduplicated version of the Copyright Data in which similar company names are grouped and represented by a single cluster name.

A.4.4 Initial Merge of Copyright Data with CRSP/Compustat Identifiers

The deduplicated Copyright Data is then merged with CRSP/Compustat identifiers to link company intellectual property information with financial data. To account for company name variations over time, we match claimant names to CRSP names, which are further linked to Compustat by the WRDS CRSP/Compustat Merged Database. CRSP company names are cleaned and standardized using the same methods applied to the Copyright Data. Both the Copyright Data cluster names and CRSP company names are prefixed with "The company name is " to prepare names for sentence embedding models.

Embeddings are generated for both datasets to be merged using the BAAI/bge-large-en-v1.5 pre-trained embedding model with the LinkTransformers python library to streamline the following steps. Cosine similarity scores are calculated between Copyright Data company embedding and CRSP company embeddings. For each Copyright Data company, the top 200 CRSP companies with the highest similarity scores are selected as potential matches. This yields a greedy approach that generates many potential matches including true positives and false positives, the latter are removed in the subsequent step below. Matches with a similarity score below 0.8 are excluded to avoid excess false positive matches (e.g., from overly generic company names). This process results in a merged dataset containing potential matches between the Copyright Data and CRSP companies, along with associated similarity scores.

A.4.5 Final Merge with CRSP/Compustat

To improve the accuracy of matching between the Copyright Data and CRSP company names, a more intensive binary classifier is trained on hand-labeled potential matches. The training data

contains 1,544 pairs of Copyright Data-CRSP company names and a binary label indicating whether a human researcher determined the pair to be a match.

Feature engineering is performed to create input variables for the model. String similarity metrics such as Levenshtein distance, Jaro-Winkler similarity, and fuzzy matching scores (partial ratio and token set ratio) are calculated for each pair of company names. Text length features, including the number of characters and words in each company name, are also computed. Sentence embeddings are then are generated using three models: 'BGE' (BAAI/bge-large-en-v1.5), 'EMBER' (llmrails/ember-v1), and 'LinkTransformers' (dell-research-harvard/lt-wikidata-comp-en). Cosine similarity scores are calculated for each model's embeddings between company name pairs.

The dataset is split into training and test sets (80% training/20% test) using stratified sampling. The AutoML library is used for automated hyperparameter search, model ensembling, feature selection, and further feature engineering. AUC (Area Under the ROC Curve) is used as the evaluation metric for this binary classification task. The best-performing ensemble model is selected based on the evaluation metric (AUC), and its accuracy and AUC are assessed on the test set to ensure generalization, yielding an accuracy of 91.4% and an AUC of 0.909. This classifier is then applied to the full dataset generated by the initial merge to eliminate false positive matches. With this data linking copyright claimant names to company names appearing in the CRSP database, copyright registrations are then linked by company name and year to unique company identifiers (permco).

A.4.6 Litigation Data

We use the same process above to match CRSP/Compustat identifiers with companies appearing in the federal litigation data. Some specifics of the process differ, including training data using to train the classifiers and the resulting test-set accuracy. Those differences are itemized below.

- "Rule-based Data Cleaning and Standardization": For the litigation data, "ET AL" is removed from organization names.
- "Company Type Classification Using Few-Shot Learning": The training data for the classifier covers 589 hand-labeled entities in the federal litigation data. Entities are classified into one category of "organization," "person," or "other." Classifier accuracy: 95.8%.
- "Final Merge with CRSP/Compustat": The training data for this classifier using the litigation data includes 1,140 observations (pairs). Accuracy: 91.7%, AUC: 0.903.

Table A.1: Populated data field by type of record (1978-2021)

Data Fields	Registrations	Recordations	Pre-reg	Cancellations	Renewals
Registration or Document #	100	100	100	100	100
Title	100	100	100	100	100
Type of Work	100	0	100	100	100
Registration/Recordation Date	96	100	100	0	100
Creation Date	93	0	100	3	0
Publication Date	51	0	100	1	0
Author Name	55	57	99	20	2
Birth and Death Dates of Author	41	0	0	3	3
Author Citizenship	26	0	0	1	0
Author Domicile	21	0	0	1	0
Portion of Work Attributable to Author	32	0	0	2	0
Claimant Name	99	5	100	7	100
Claimant Address	32	0	84	2	0
Employer for Hire	13	0	0	1	0
Published/Unpublished	100	0	100	100	100
Publisher Info (name, date, place)	12	0	0	0	0
Nation of 1st Publication	19	0	0	1	0
Publication Frequency	5	0	0	0	0
Scope of Registered Materials	19	0	0	1	14
Rights and Permissions	24	0	14	2	9
Series Linkage	4	0	0	0	7
Titles in Collection	4	0	0	0	0
Link to Previous Registered Versions	9	0	0	0	0
Serial Issue Number	6	0	0	0	0
ISSN or ISBN	4	0	0	0	0
Bibligraphic Notes	23	5	0	0	11
Description of Preregistered Materials	0	0	99	0	0
Cancelation Reason	0	0	0	100	0

Source: Copyright Data.

Notes: Values represent the percentage of records for which a particular data field is populated. A low percentage does not necessarily indicate that the data field is of low informational value. Some fields are highly populated within a certain context, but not applicable outside of that context, resulting in a low overall population rate. For example, claimant addresses are not recorded prior to 2007, but are almost always recorded after 2007. Similarly, "Series Linkage" is typically populated when the work to which the record pertains is part of a larger collection of work; however, most records pertain to works that are not part of a larger collection.

Table A.2: Tabular Data Code Book

Variable Name	Description
alt_title	Alternative title of the work.
author_auth_stat	Author statement - provides information about the elements of a work attributable to the associated author.
author_brth_year	Birth year of associated author.
author_citizanship	Citizenship of associated author.
author_corp_ind	Indicates whether an author is an individual or an organization (usually a corporation).
author_domicile	Domicile of associated author.
author_dth_year	Death year of associated author.
author_name	Name of first 10 authors listed. Registrations with more than 10 listed authors are truncated to only the first 10 in order to make the data easier to work with. This truncation affects less than 0.1% of records. For unabridged, long-form data see https://copyright.gov/economic-research/usco-datasets/
author_wfh	Indicates whether associated author is operating on work for hire basis, as indicated by the claimant.
biblio_note	Unstructured data field containing bibliographic notes. Can, for example, indicate that the registration is for a collection of works.
claimant_address	Address of associated claimant. Usually only the first claimant provides an address. Addresses are consistently recorded in the data starting in 2009.
claimant_brth_year	Birth year of associated claimant.
claimant_corp_ind	Indicates whether a claimant is an individual or an organization (usually a corporation).
claimant_dth_year	Death year of associated claimant.

claimant_name	Name of first 10 claimants listed. Registrations with more than 10 listed claimants are truncated to only the first 10 in order to make the data easier to work with. This truncation affects less than 0.003% of records. For unabridged, long-form data see https://copyright.gov/economic-research/usco-datasets/
claimant_trans_stat	Claimant transfer statement - notes how rights became vested in claimant.
creation_date	The year the particular version of the work was completed.
fst_commercialized	Date the work was first commercialized.
isbn	International Standard Book Number (ISBN).
issn_num	International Standard Serial Number (ISSN).
periodical_freq	Frequency of periodical.
permissions_contact	The contact information for the person or entity who can facilitate usage permissions - usually one of the claimants or an agent who can act on their behalf.
previous_reg_note	Notes about previous copyright registrations related to work.
pub_city	City in which work was published.
pub_country	The country of first publication.
pub_date	The date of publication.
publication_stat	Whether the work is published or unpublished at the time of registration.
publisher	Name of the publisher.
reg_date	The effective date of registration (the date to which certain benefits of copyright registration can be retroactively applied). "The effective date of registration is the day that the Office receives in proper form all required elements—an acceptable application, an acceptable deposit, and a nonrefundable filing fee." See, https://www.copyright.gov/circs/circ02.pdf
reg_num	Unique registration number assigned by the U.S. Copyright Office.

scope_of_rights	The scope of rights being claimed by the particular claimants. A complex work such as a film may have multiple rights associated with it, such as videography, screenplay, the musical compositions used, sound recordings of musical performances, narration, or dramatic performances. For various reasons the claimants may not have the rights to all elements of a work. This field allows claimants to identify which elements are being claimed (or disclaimed) as part of the particular registration.
series_name	Name of series.
state_of_responsiblity	Contains unstructured information submitted by claimants regarding the individuals or entities that created the works and their particular contributions to it.
title	The title of the work.
track_title	Lists up to 30 track titles included in a compilation registrations.
volume_num	Volume number.
work_type	Type of work based on MARC21 data field 917, subfield $b ("Retrieval Code"). Note, USCO classifications were modified in 1983. For full description of coding, see www.copyright.gov/policy/women-in-copyright-system/LOC-Copyright-Data-as-Distributed-in-the-MARC%2021-Format.pdf

Note: When a datafield is not applicable to a type of work it is omitted from the datatable for that type of work. For example, sound records do not have ISBNs; thus, that field does not appear in sound recording datatables.

Table A.3: Compustat-linked Registration and Litigation Data

Variable Name	**Description**
LPERMCO	CRSP PERMCO Identifier
fyear	Calendar Year
gvkey	Compustat GVKEY Identifier
LPERMNO	CRSP PERMNO Identifier
reg_count	Count of US Copyright registrations linked to LPERMCO in fyear
plt_count	Count of litigation linked to LPERMCO as plaintiff, filed in fyear
def_count	Count of suits linked to and filed against LPERMCO (as defendant) in fyear

Note: Data are indexed by *LPERMCO* and *fyear*. Because registrations and litigation actions are linked to companies by company identifiers (*LPERMCO*), multiple *gvkeys* may be linked to one *PERMCO*. In such cases, researchers should be careful to select the appropriate *gvkey*(s) for their research question