ONLINE APPENDIX I TO

REGISTERING AUTHORS: CHALLENGING COPYRIGHT'S RACE, GENDER AND AGE BLINDNESS

Robert Brauneis and Dotan Oliar

FROM THE COPYRIGHT OFFICE CATALOG TO THE ORIGINAL VALID MONOGRAPH REGISTRATION DATASETS: SOME HISTORY AND TECHNICAL DETAILS

The records in the Copyright Office Electronic Catalog are currently maintained in the Machine-Readable Cataloging (MARC) format for bibliographic records. A family of MARC formats was originally developed at the Library of Congress in the late 1960s and early 1970s under the direction of Henriette Avram.[1] The MARC format for bibliographic records was developed to store information about books, sound recordings, video recordings, and other items in library collections. The Copyright Office has adapted that format to store information about Copyright Office transactions. Before 2007, the Copyright Office used a database system called "COPICS" (for "Copyright Office Publication and Interactive Cataloging System") that had been developed in-house at the Library of Congress.[2] The COPICS system in use from 1978 until 2007 was composed of three indexes:

- COHD, "Copyright Office History Documents," an index to recorded documents such as transfers, grants of security interests, etc.
- COHM, "Copyright Office History Monographs," an index to registrations of everything but serials.
- COHS, "Copyright Office History Serials," an index to registrations of serials.[3]

In 2007, records in the old COPICS formats were converted to MARC format. In the 9/2014 Catalog, over 19 million records have a "last modified" date in June of 2007, dating the mass conversion and uploading to that month.

---

[1] *See* Henrietta D. Avram, MARC: Its History and Implications (1975).

[2] Beginning in about 1974, COPICS was used to produce the Catalog of Copyright Entries. It was revised to allow more general use as of the effective date of the Copyright Act of 1976 – January 1, 1978 – and the revised system is often referred to as "COPICS II," to distinguish it from "COPICS I," the pre-1978 system. The data now electronically available dates from January 1, 1978.

[3] The COPICS databases used a custom metadata scheme developed for copyright records. Data fields were identified by four-letter labels; for example, "TITL" identified the field for the title of the work, "CLNA" was the field for claimant names, and "APAU" was for authors' names appearing only on the application. The COPICS files first became available for searching over the Internet through a Telnet interface on April 30, 1993. *See* "Remote Access to Library of Congress Computer Files Now Available" (Press Release April 6, 1993), *available at* http://www.loc.gov/today/pr/1993/93-059.html. A web-based search was made available on the Copyright Office website in 2001. *See* "Copyright Office Announces New Search System," August 17, 2001, available at http://www.loc.gov/today/pr/2001/01-114.html (last visited June 13, 2014).

Typically, there is one Catalog record generated per Copyright Office transaction: one registration, for example, usually results in one Catalog record. There are, however, two principal exceptions to that rule. The first exception concerns recorded documents that reference more than one work (technically, more than one title or registration number) – for example, an assignment that transfers title in ten motion pictures. If a recorded document references more than one work, then one record is generated for the document, and separate records are generated for each work to which the document refers. Thus, the 8,748,658 Catalog records relating to recorded documents represent only about 522,000 documents; the other 8.2 million records represent titles of works to which one of the recorded documents refers.[4]

The second exception concerns serials, that is to say, works that are published in series, such as magazines and newspapers. Until recently, all registrations for issues of serials published during a particular year were grouped in a single registration record. Thus, the 9/2014 Catalog contains 1,165,340 registration records respecting serials, but those records represent 4,024,075 registrations of serials.[5]

II. Copyright Catalog Design: Well-Informed Human Readers of Entire Records versus Machine Readers of Particular Fields.

There are a number of difficulties that must be addressed when attempting to produce datasets suitable for quantitative analysis based on the Copyright Catalog. A Copyright Catalog registration record in its native MARC format is divided into many data fields, and each field in theory presents a specific type of information about the registration. In practice, however, the Catalog has not been designed or implemented to be machine-readable or to enable quantitative analysis. Rather, it has been designed and implemented with the goal of providing copyright information to a human reader who has a natural language understanding of English and some familiarity with copyright, and who is reading the entire record of a copyright registration. That catalog design and implementation allows for many types of informalities, variations, and inconsistencies that make field-based quantitative analysis more difficult. Those informalities, variations and inconsistencies include the following:

- **Reliance on Natural Language and Contextual Interpretation.** Records may rely on people to draw conclusions about copyright-relevant facts based on relatively complex contextual interpretation. For example, for most of the history of the Catalog, the field that lists claimants – field 249c – lists all claimants in a single string of characters in a single instance of the field. Ampersands and commas are typically used to separate the names of multiple claimants in that field, but they are also used within the names of those claimants. Thus, it

[4] 521,899 document records have a field (the 291 "Party One" field in the MARC record) that should mark those records as referring to a document rather than a title mentioned in a document; 522,342 records are missing a field (the 787 field in the MARC record, used for linking titles to the principal document record) that should mark those records as referring to titles mentioned in a document. There is a 443-record discrepancy between those two figures; we would have to look at each of those 443 records to determine what they were, and we have not done so. For more on recorded document records, including statistics on recorded documents from 1978 through 2009, see Robert Brauneis, Transforming Document Recordation at the United States Copyright Office 36-45 (2014), *available at* http://copyright.gov/technology-reports/reports/recordation-report.pdf.

[5] As noted in the text, the earlier records for serials were "containers" filled with information on all registrations made for the particular serial in question in a given year. The Copyright Office sometimes created records for serials for years during which no issues of those serials were registered. As a result, 79,239 serial records in the 9/2014 Catalog contain no registrations at all – they are empty containers.

becomes a more complicated task to understand, for example, that "John Wiley & Sons, Inc."[6] is a single claimant, but that "David Mallett & Old Road Music"[7] are two claimants. In this case and others, comparing data across two fields may help. The "index" field for the first of the two records mentioned above, field 710, lists a single claimant name, "John Wiley & Sons." The index field for the second of the two records lists two claimant names, "Mallett, David, 1951-" and "Old Road Music." Thus, by comparing information across two fields, we are able to form a better founded opinion about the number of claimants in the registration records in question. Lest you think that the index fields by themselves might be sufficient, some examples can show that they are not. For example, the index fields for another registration list two claimants, "Discovery Records" and "Warner Music Discovery, Inc."[8] Only by inspecting the claimant name field does one learn that the claimant is "Discovery Records a.k.a. Warner Music Discovery, Inc." – in other words, that there is a single claimant known by two names.[9]

- **Varying Fields.** A particular type of information does not always have to be stored in the same database field. For example, a notation that limits the claim made in a registration, and information about the birth year of the claimant, could be in the same field as the claimant's name – for example, "on music, recording; Lucas James Rodenbush, 1977-";[10] or the claim limitation and birth year information could be in their own dedicated fields.

- **Data Type Mixing.** A field may contain two or more types of information. For example, as in the example immediately above, a field may contain a limitation of the claim; the name of the claimant; and the birth year of the claimant.

- **Data Instance Combination.** A field may contain a single item of information of a particular type – for example, the name of a single claimant – or it may contain multiple instances – the names of two or more claimants in a single field. MARC records are designed to allow for the repetition of fields, and of subfields within those fields. In some records, two co-claimants may be identified in a single instance of a field or subfield – "Fiddleback Music Publishing Co, Inc., & New Start Music, Inc."[11] – while in other records, each co-claimant may be identified in a separate instance of a field or subfield, for example, "Jorge Luis Chacun Music" and "Universal Musica Unica Publishing."[12]

- **Inconsistent Instance Separation Within a Field.** Multiple instances of data in a single field may not always be separated by the same separator. A field identifying multiple claimants might use only commas to separate their names – "Screen Gems-E M I Music, Inc., Black Sheep Music, Tree Publishing Company, Inc."[13] Or it might use the word "and" – "Michael Caesar

---

[6] *See, e.g,* U.S. Copyright Registration No. TX0000080795.

[7] *See, e.g,* U.S. Copyright Registration No. PAu000001242. Similar examples abound. *See, e.g.,* U.S. Copyright Registration No. PAu000001520 ("Lambert & Potter Music Company"); U.S. Copyright Registration No. TX0000080050 ("David & Kay Scott"); U.S. Copyright Registration No. PAu000001684 ("Adamo Music & Good Vibes Music").

[8] U.S. Copyright Registration No. SR0000235753.

[9] *Id.*

[10] U.S. Copyright Registration No. SR0000235713 (field 249c).

[11] U.S. Copyright Registration No. PAu000968946.

[12] U.S. Copyright Registration No. PA0001914567 (two instances of field 249c).

[13] U.S. Copyright Registration No. PAu000969125.

Scharff and Adam Bayard Scharff"[14] Or it might use a comma to separate each claimant except for the last, which it separates with an ampersand or the word "and" – "Patricia Dodson, John Buchanan & James Borden"[15] or "Timothy Olinger, Brent and Paul Nickolaus" (and note the absence of Brent's last name).[16] Or it might use both a comma and an ampersand to separate two claimants – "Fiddleback Music Publishing Co, Inc., & New Start Music, Inc."[17]

- **Varying Terminology and Phrasing.** In many cases, catalogers have used varying terminology and phrasing to state facts that are necessary to know when analyzing records. For example, the fact that an individual claimant or author of a work is known by different names is noted in Catalog records with a variety of terms and structures. Consider the following nonexhaustive examples:
    - "Robert H. Hines a.k.a. William Kirberger."[18]
    - "Robert H. Hines (pseud., William Kirberger)."[19]
    - "Elizabeth Roberts (Lisa Roberts, pseud.)"[20]
    - "Walter B. Gunby, 1947-, whose pseud. is Radical Research;"[21]
    - "CHARLES DUANE WHEELER WHOSE PSEUDONYM IS BIFF."[22]
    - "Janet Whiteaker pen name J.F. Whiteaker."[23]
    - "DonIlyan legal name (Don Adcock, birth name; Zyjuer, alias)."[24]
    - "Barun Ananda Mukhopadhyay (Barun Mukherjee), Ph.D."[25]
    - "Gloria Fair, 1925- (Mrs. Edwin Fair)."[26]

  Similarly, the fact that a business entity is known by different names, or is affiliated with another named business entity, is noted in a variety of ways. Many registrations use "d.b.a." – "Frank S. Szymanski, Jr., d.b.a. Franklin Sane."[27] Other registrations use "a.a.d." – "International Business Machines Corporation, a.a.d.: IBM Corporation,"[28] "Beka Book Publications, an a.a.d. for Beka Book Publishing Company, a division of Pensacola Christian College."[29] Yet others use "trading as" – "J.H. Haynes & Company, Ltd., trading as Haynes Publishing."[30] Other examples include: "Richard D. Irwin, a Times Mirror Higher Education

[14] U.S. Copyright Registration No. PAu000969163.
[15] U.S. Copyright Registration No. PAu000968676.
[16] U.S. Copyright Registration No. PAu000968732.
[17] U.S. Copyright Registration No. PAu000968946.
[18] U.S. Copyright Registration No. PAu000970637.
[19] U.S. Copyright Registration No. PAu000838393.
[20] U.S. Copyright Registration No. TXu001255635.
[21] U.S. Copyright Registration No. TX0004201541.
[22] U.S. Copyright Registration No. PAu003378462.
[23] U.S. Copyright Registration No. TX0005832417.
[24] U.S. Copyright Registration No. PAu000054577.
[25] U.S. Copyright Registration No. TX0006944193.
[26] U.S. Copyright Registration No. TX0004483485.
[27] U.S. Copyright Registration No. PAu000974489.
[28] U.S. Copyright Registration No. TX0004187667.
[29] U.S. Copyright Registration No. TX0004190954.
[30] U.S. Copyright Registration No. TX0005833285.

Group, Inc. company;"[31] "Butterworth-Heinemann, a member of the Reed Elsevier group;"[32] "Visible Computer Supply Corporation, a subsidiary of Wallace Computer Services, Inc.;"[33] "West Services, Inc., operating as West a Thomson business (employer for hire);"[34] "Serendipity House (wholly owned by Lifeway Christian Resources) (employer for hire)."[35]

- **Inconsistent Formatting.** Information may be presented in different formats. For example, a date might be presented as "26May97"; "May 26, 1997"; 5/26/97"; and so on.
- **No Name Standardization.** Name authority files are an important part of modern bibliographic practice, and unique identifiers are of emerging importance in the effort to organize knowledge about individuals and organizations involved in the creation, distribution, and ownership of creative works. Name authorities attempt to standardize the names of individuals and organizations, and to collect and relate variants, so that, for example, someone looking for all of the works of a particular author can find them.[36] Unique identifiers, such as International Standard Name Identifiers[37] and Open Researcher and Contributor IDs,[38] are systems for assigning unique identifiers to individuals and organizations. The Copyright Office has never used or encouraged the use of any such systems in its registration or recordation records. Thus, for example, if one registration is filed by "International Business Machines, Inc." and another as "IBM, Inc.," it would take background knowledge of the world of business to understand that those two might be alternate names for the same claimant. In practice, there is a degree of consistency in the names of claimants and authors, but that results only from the practices of those claimants and authors, not from any efforts of the Copyright Office.

It is not the case that the Copyright Office has lacked cataloging standards; it has developed many such standards. However, the standards are certainly not exhaustive, and more importantly, many of them have been refined through changes over 35 years. In particular, the records generated since electronic copyright registration was introduced in 2007 involve much more parsing of information into discrete fields. As standards have improved, however, the Office has rarely gone back to restructure older records; they are typically left as they were originally created. In addition, as mentioned above, there has historically been very little validation of entries made by registration specialists, so if those specialists deviate from standards or follow varying practices to fill gaps in standards, inconsistencies in Catalog records will result. As we have developed the original monograph registrations dataset, we have tried to be sensitive to these issues, and either to correct for them when we can, or to identify them in documentation when they appear that they might be substantial but correction is difficult or impossible.

[31] U.S. Copyright Registration No. TX0004200965.
[32] U.S. Copyright Registration No. TX0004483482.
[33] U.S. Copyright Registration No. TX0004201529.
[34] U.S. Copyright Registration No. TX0005832448.
[35] U.S. Copyright Registration No. TX0005832894.
[36] *See, e.g.*, Library of Congress Names, http://id.loc.gov/authorities/names.html (last visited August 27, 2015).
[37] *See* INSI International Agency, http://www.isni.org/ (last visited August 27, 2015).
[38] *See* ORCID, Inc., http://orcid.org/ (last visited August 27, 2015).

There are four different fields in which information about the number of works in a registration transaction is sometimes found, but often not found:

- **The "Title" field** (245a in the MARC record) most often does not contain a count of works. For example, the title in Registration No. VA0001240079, which concerns a group of photographs by many different photographers registered by a stock photography company, is "Alaska Stock group registration for automated database: updates from 1-1-2004 through 3-31-2004." However, sometimes titles do include the number of works concerned. The title of Registration No. VAu000753656 is "071807 Liz Ordonez-Dawes not published group registration/555 images." (Note that machine reading of numerals in that title would have to be quite sophisticated to distinguish between "071807" and "555" and to conclude that the former had nothing to do with the number of works involved but the latter did.)

- **The "Alternative Title in Application" field** (243a in the MARC record) sometimes contains information about number of works when the title field does not. For example, in Registration No. VA0001697532, the "Title" (245a) is "04D1E~.zip, et al."; the "Alternative Title in Application" field (243a) is "Group registration photographs; 372 photos." However, that field is often blank or does not contain any numerical information.

- **The "Application Title" field** (242a in the MARC record) also may contain information about number of works. For example, in Registration No. VA0001308138, the "Title" (245a) is "2002 photos of Elle, Natalie, Sabrina & Zuse : no. JSH2002"; that might lead some to believe or at least wonder whether "2002" is the number of photographs submitted with the registration. However, in that registration the "Application Title" field states "Group registration/photos, approx. 166 photographs."

- **The "Contents Note" field** (505t in the MARC record) is designed to contain the titles of discrete parts of the work that is the subject of the registration. It appears to be most often used to list the titles of individual tracks on a phonogram (Compact Disc, vinyl, tape, or other medium) submitted as the deposit for a sound recording or musical work registration, but it is also used for other types of works. There is a great deal of variation in how titles are separated in a 505t field. In a single instance of a 505t field, titles can be separated in at least three ways:

  - by commas, *see, e.g.*, Registration No. PA0001721738 ("First Underground Nuclear Kitchen", "Love Communion", "We Shall Be Free", "Never Say Never".);
  - by semicolons, *see, e.g.*, Registration No. PAu003605315 (Cowboy In The City; Lifetime Right; Standing In The Shadow; Remember Every Child; Clickity Clack; Fargo; Thank You Girl; Profilin A Redneck; Wonder Dog; Rebel Writer); and
  - by numbers, *see, e.g.*, Registration No. SR0000695138 (1. My Star 2. Shine For You Tonight 3. The Light 4. She Is Lovely 5. More Than Anything . . .).

  In addition, each title can be placed in a separate instance of a repeated 505t field, *see, e.g.*, PAu003605256.

Table 5 presents some statistics concerning the 505t Contents Note field in Original Valid Registration records, by year from 1978 through 2012. The first column lists the number of records in each year in which a 505t field is found. The second column lists the percentage of all OVM records in which a 505t field is found. The third column lists the estimated number of titles listed in 505t fields, having counted for comma, semicolon, and period separators (periods having been

used after numbers that separated titles). Note, first, the dramatic difference in years 2007-2012, when the electronic registration system made it possible for applicants to provide Contents Note titles themselves. The percentage of records that include 505t fields quickly jumps to about 10%, from an average of about 1% over previous years. That suggests that no comparison can be made across those years, and also that much information about registered works in previous years is missing from the Catalog. Much of the variation in previous years is also likely due to changes in cataloguing policy. For example, the years 2002-2004 saw a precipitous decline in Contents Notes, and in 2005 and 2006 they all but disappeared. This is almost certainly due to a temporary order to "catalog from applications only" to quickly process backlogged registration applications, followed by a decision by Register of Copyrights Marybeth Peters to stop entering most bibliographic information into Electronic Catalog records.[39] The result is that the number of titles listed in Contents Notes is not a very useful aid for figuring out anything about the number of works registered; changes in cataloging practices obscure any changes in real numbers.

Table 5: The 505t Contents Note Field, 1978-2012

Year	# records	% records	Titles listed		Year	# records	% records	Titles listed
1978	3730	1.49%	44370		1996	4556	0.96%	52820
1979	4282	1.45%	49301		1997	4084	0.86%	47430
1980	3954	1.29%	44826		1998	5027	1.06%	57783
1981	3732	1.17%	42513		1999	4490	0.97%	53715
1982	4010	1.20%	44664		2000	4108	0.91%	49344
1983	3579	1.05%	38078		2001	3337	0.75%	42325
1984	3196	0.91%	35197		2002	2991	0.67%	38357
1985	3202	0.86%	35999		2003	2711	0.60%	35339
1986	2786	0.71%	29094		2004	1446	0.31%	17978
1987	2831	0.70%	29418		2005	363	0.08%	4537
1988	2933	0.70%	30463		2006	160	0.04%	1595
1989	2954	0.66%	31572		2007	11780	2.65%	102001
1990	3197	0.65%	35253		2008	31986	7.61%	290681
1991	3351	0.78%	38160		2009	37414	9.25%	395714
1992	3360	0.77%	38903		2010	40649	9.51%	422522
1993	4158	0.93%	44130		2011	44930	10.14%	658713

[39] *See* Copyright Office Makes Final Decision on Catalog Record, a report by Margaret Holley prepared for the Library of Congress Professional Guild (AFSCME 2910), available at http://www.guild2910.org/CopyrightCatalog.pdf (last visited February 15, 2016).

1994	3782	0.84%	42796		2012	44569	9.67%	905478
1995	4021	0.89%	46421		**Total**	**307659**		**3877490**

In conclusion, we have not found any consistent sources of information in registration records about the number of works covered by those records. Other techniques, such as examining samples of deposits, would have to be used to develop estimates of the prevalence and size of group registrations in different time periods. Some types of works, such as textual works and motion pictures, seem less likely to be the subject of group registrations, whereas other types of works, such as photographs, sound recordings, and musical works, are more likely to be registered in groups or one kind or another.

One last note: in some cases, even the examination of the deposit would not necessarily lead to a correct count of the works that are the subject of a registration. Some of the records of the registrations at issue in the Alaska Stock litigation[40] include the notations "Preexisting material: some prev. pub. photos." and "New Matter: new photos & compilation."[41] In other words, Alaska Stock submitted CDs with hundreds or thousands of photographs on them, and did not take the trouble to distinguish between those that were the subject of the registration, and those that were not. Some of the photographs might have even been in the public domain, but presumably Alaska Stock would contend that it was not committing fraud on the Copyright Office because it revealed that it was not claiming copyright for all of the photographs – though it did not identify those photographs for which it was not claiming copyright. There is a serious question, however, whether claimants should be receiving the benefits of registration when providing so little information about what they are actually claiming copyright in.

1. Information About Claimants

In the Catalog MARC records, the principal field that contains that information about claimants is field 249. However, as detailed above, in earlier records, a single instance of field 249a contains the names of all claimants in one character string, and it is difficult to parse that string to figure how many names of claimants it contains, because every character or word used to separate claimant names – principally, the comma, the ampersand, and the word "and" – is also used within claimant names.

For this reason, for purposes of counting and classifying claimants, we decided that fields 700 and 710 are better sources. Those fields were created for purposes of indexing records. Field 700 contains the names of individuals identified as either claimants or authors, and field 710 contains the names of corporations and other organizations identified as either claimants or authors.[42] Both fields are repeatable – a MARC record can contain as many instances of field 700

[40] *See* Alaska Stock, LLC v. Houghton Mifflin Harcourt Pub. Co., 2010 WL 3785720 (D. Ak.), *reversed and remanded*, 747 F.3d 673 (9th Cir. 2013).

[41] *See, e.g.*, U.S. Copyright Registration No. VA0001240079; U.S. Copyright Registration No. VA0001316377.

[42] There are 31,202 registrations that do not have 700 or 710 index entries. Many of these registrations are for works the author of which is anonymous. *See, e.g.*, U.S. Copyright Registration No. PAu000047685 ("Songs in Praise of Avatar Meher Baba"); U.S. Copyright Registration No. TXu001211035 ("America's eMortgageMall for commercial & business loans") U.S. Copyright Registration No. TXu001207354 ("Sacrificial lambs"); U.S. Copyright Regsitration No. VAu000567004 ("Jerome Goldstein, man of confusion"). Some are apparently for works for which no authorship information was given or recorded. *See, e.g.*, U.S. Copyright Registration No. TXu001213073 ("A journey for lovers");

or 710 as there are claimants and authors of a work. Each instance of field 700 or 710 contains one and only one individual or corporate name, in subfield 700a or 710a, and a designation that the individual or corporation named is either a claimant or an author or both, in subfield 700c or 710c.

Pre-MARC-format records did not distinguish between individual and corporate names, and thus when the Copyright Office converted those records to MARC format in 2007, a simple method was needed to sort existing index entries into the categories of individual and corporate. Individual names had been (and continue to be) entered last name first, with a comma between the last and first names; corporate names were and are entered without any change in order, and in principle without any commas. It was therefore decided to sort the names into individual and corporate categories by searching for the presence of a comma in the name. If a comma was present, the name was classified as an individual name; if a comma was not present, the name was classified as a corporate name. This method was known to be imperfect; for example, individuals who had single names, and especially single-name pseudonyms, were classified as corporate entities. We believe, however, that the number of misclassifications is relatively small – much less than one percent – and is therefore adequate for most statistical purposes.

The principal problem with counting claimants using fields 700 and 710 is that claimants will be overcounted unless adjustments are made. If a claimant is represented in field 249 as having more than one name – a legal name and a pseudonym, or an official incorporated name and a "doing business as" name, for example – then catalogers often, but not always, created create index entries in separate instances of field 700 or 710 for each of the names. Similarly, if a claimant is represented as being affiliated with another named entity in field 249 – in one company is a division or subsidiary or member of another, for example – then there will sometimes, but not always, be entries in separate instances of field 700 or 710 for each of the names.

To provide claimant counts, the current version of the OVM dataset corrects for pseudonyms and other alternate names for claimants in two ways. First, entries concerning names of individuals in field 700 that are marked "claimant" and that contain the expression "pseud." are not counted as claimants unless there is only one 700 entry associated with that registration and that entry contains the expression "pseud." Registration records that contain information about the pseudonym of a claimant usually also contain the claimant's real name. If there are two names associated with the registration and one of them is marked "pseud.," we assume that the other one is the same individual's real name, whereas if there is only one name and it is marked "pseud.," that is a registration in which the individual's real name is not revealed. This method will overcorrect for pseudonyms when (1) there is more than one claimant associated with a registration, and (2) one of the claimants is identified only by a pseudonym, while the others are identified by their real names.

As for entries marked "pseud." in 710 fields, concerning corporate claimants, we disregarded those altogether for purposes of counting claimants. There are 57,364 of those in the

---

U.S. Copyright Registration No. TXu001212430 ("Jes jet"). However, many are registrations that have author information in 245c or 279c that was apparently not transferred correctly to a 700 or 710 entry. *See, e.g.*, U.S. Copyright Registration No. TXu001213078 ("My love, my life, my heaven on earth") (author information contained in 245c).

A relatively small number of 700 and 710 entries – 12, 580 entries in field 700a, and 1083 in field 710a – are not marked as either author or claimant. Some of those are "quasi-author" names – for example, the names of film directors such as Ingmar Bergman and Steven Soderbergh." U.S. Copyright Registration No. PA0001354468 (field 700a is "Bergman, Ingmar") U.S. Copyright Registration No. PA0001354843 (field 700a is "Soderbergh, Steven"). Others appear to be mistakes, for example, what appear to be the catalog numbers of sound recordings on which cover art appears. *See, e.g.*, U.S. Copyright Registration No. VA0000672336 (field 710 is "Reprise 9 45786-2").

dataset, and they are almost certainly pseudonyms for individual claimants and/or authors that should be in 700. They are single word pseudonyms, or multiple word pseudonyms that didn't follow a "first name, last name" pattern, and therefore didn't have a comma that would have led them to being placed in field 700 during the conversion to MARC records. For example, Field 710 in Registration No. PAu001767859 is "Sting, pseud.," the pseudonym for the musician and singer Gordon Sumner; Field 710 in Registration No. PA0001099838 is "Memphis Minnie," the pseudonym for blues guitarist, vocalist, and songwriter Lizzie Douglas Lawlers; and Field 710 in Registration No. PA0001096628 is "Snoop Dogg," the pseudonym for rapper and actor Calvin Broadus. Other entries marked "pseud." in Field 710 include the names of musical groups that are serving as pseudonyms for all of their individual members, such as "Black Sabbath"[43] or "The Killers."[44]

Second, in 700 fields (for individual claimants) we do not count names that immediately follow the abbreviation "a.k.a." In 710 fields (for corporate claimants), we do not count names that immediately follow nine different expressions that introduce alternate or related names, namely, "a.a.d."; "a.a.d.o."; "a.d.o."; "a.k.a."; "d.b.a."; "a division of"; "a subsidiary of"; and "operating as."

The problem of separating bibliographic authors from copyright authors can be briefly described as follows. Bibliographic authors are those credited on the deposit copy of a work submitted for registration. Authors for copyright purposes are those validly claimed to be authors in the registration application. There are a number of reasons why bibliographic authors might not count as copyright authors, and vice versa. Perhaps most frequently, the work may be a work made for hire, in which case the Copyright Act deems the employer or the commissioner of the work to be the author for copyright purposes, even though an employee or independent contractor created the work. In addition, the work claimed may be a derivative work or a collective work, and the deposit copy may credit the author of the underlying work or of components of the collective work, whereas the copyright author is only the person who created the derivative work or compiled the collective work. For example, there are many editions of Shakespeare plays published since 1978 for which William Shakespeare would be a bibliographic author, but only the author of the introduction to the particular edition (if any) would count as an author for copyright purposes.

As we did in identifying claimants, we start with the index entries in fields 700 and 710, because each instance of 700 or 710 contains the name of one person or organization. Unfortunately, however, individuals and organizations marked as "author" in 700 or 710 can be either bibliographic authors or authors for copyright purposes or both, so we need other information to distinguish the two. MARC subfield 245c contains information about authors of a work as credited on the deposit copy of that work. MARC field 279 contains information about authors of a work as claimed on the registration application. Before the implementation of the eCO registration system in 2007-2008, catalogers generally only created what would become the 279 field in a registration record if the copyright author(s) differed from the bibliographic author(s) in 245c. Since the implementation of eCO, registration records always contain one or

[43] *See, e.g.*, U.S. Copyright Registration No. PAu000092596.

[44] *See, e.g.*, U.S. Copyright Registration No. PA0001909269.

more instances of field 279, and one or more instances of subfield 279c, which contains the names of the authors for copyright purposes.[45]

In order to exclude those marked as authors in fields 700 and 710 who are not authors for copyright purposes, we checked for the presence of a 279 field in each record. If a 279 field was present, we searched to see whether the name or names that appeared in fields 700 or 710 also appeared in 279. If not, then we marked those names as not being copyright authors.

It should be noted that for some purposes, it would be helpful to know the names and characteristics of the individuals who created a work even though a corporate entity is the treated as the author of the work for copyright purposes. For example, the age, gender, and number of individual creators might be useful in analyzing the demographics of creators. However, although some registration records do contain this information, many registration records do not. Not only did some applicants provide this information while others didn't, but during some periods, the Copyright Office cataloged registrations from information on the application alone, without providing any information from the deposit copy. Thus, in constructing the fields in our dataset that count individual and corporate authors, we have tried to count only authorship for copyright purposes, although in the dataset concerning individual authors and claimants we have preserved information about individual creators who are not authors in the copyright sense.

[45] Registration records created in the eCO system can have multiple instances of 279c fields. However, the 279c fields in eCO-era registration record are repeated, not when there are multiple authors, but when the authorship of one component of a work is different from the authorship of another component. For example, the first instance of 279c in a record might read "words: James Smith & Angela Moore"; the second instance might read "music: James Smith, as employer for hire of Steven Auster, d.b.a. Arrangements Unlimited." This rule for splitting instances has more bibliographical significance than copyright significance, and it is a little odd that it is used for registration records in the Copyright Catalog. In most cases, the work will be a jointly authored work, and in those cases, it makes no difference that the authors contributed different elements of the work; they all have equal ownership interests in the entire work, and if they have transferred the work, equal termination interests. If the work is not a jointly authored work, then it could make some difference; but registration records contain no explicit information about whether the works in which copyright are being claimed are jointly authored works, and thus it will never be clear whether the presence of two or more instances of 279c has any copyright significance. What is clear is that under the current rule for creating multiple instances of 279c, counting instances does not help count the number of authors of the work or works in which copyright is being claimed.