5-2023

Measuring Change in Copyright Exceptions for Text and Data Mining

Michael Palmedo

Momina Imran

Miguel Alvarenga

Luca Schirru

Duc Le

MEASURING CHANGE IN COPYRIGHT EXCEPTIONS FOR TEXT AND DATA MINING

Michael Palmedo, Miguel Alvarenga, Momina Imran, Duc Le, and Luca Schirru[*]

ABSTRACT

Copyright exceptions for researchers are under debate at the World Intellectual Property Organization and within domestic governments, yet empirical research in this area is rare. In this early working paper, we aim to add to this nascent body of research. We expand PIJIP's previous review and classification of copyright exceptions in WIPO Members' laws by tracing changes in the laws over time. We find that most countries have copyright exceptions allowing some unauthorized uses for research purposes. However, most countries' exceptions restrict some mix of the users, uses, or types of works that are allowed. High-income countries tend to be more permissive of researcher's unauthorized uses, than countries in other income groups, and their laws have grown slightly more permissive over the past two decades. Former British colonies with a history of fair dealing tend to be more permissive than other countries, but they are becoming less permissive on average as they amended their laws.

[*] Michael Palmedo is a Post Doctoral Research Fellow at American University's Program on Information Justice and Intellectual Property (PIJIP). Luca Schirru and Duc Le are recent LL.M graduates from American University. Miguel Alvarenga and Momina Imran are current LLM students at American University.

CONTENTS

I. INTRODUCTION

Copyright exceptions are under debate at the World Intellectual Property Organization (WIPO), including copyright exceptions for researchers that impact their ability to engage in text and data mining.[1] Domestic governments are reviewing copyright laws to assess whether emerging technologies will change how we create and consume works, and their decisions may also impact researchers.[2] Some governments are already considering legislation or passing laws to amend their exceptions.[3]

Despite calls for "evidence-based" policymaking in intellectual property,[4] there is only a small body of work regarding the impact of

[1] Proposal by the African Group for a Draft work Program on Exceptions and Limitations, WIPO Standing Committee on Copyright and Related Rights. SCCR/43/8, March 17, 2023; Revised Pilot Project on Text and Data Mining to Support Research and Innovation in Universities and other Research Oriented Institutions in Africa. WIPO Committee on Development and Intellectual Property. CDIP/30/9 REV, April 28, 2023.

[2] Copyright Office Launches New Artificial Intelligence Initiative. U.S. Copyright Office press release. Issue No. 1004 - March 16, 2023 (the review includes questions about the training of AI on copyrighted materials, which is similar to the issues involved with text and data mining copyrighted materials).

[3] Jonathan Band. "New Nigerian Copyright Act Creates Open Fair Dealing exception." Infojustice. March 24, 2023. https://infojustice.org/archives/45182

[4] World Intellectual Property Organization, Guidelines to using evidence from research to support policymaking. 2019. https://www.wipo.int/publications/en/details.jsp?id=4460

copyright exceptions on *researchers*.[5] Flynn et al. categorize over 200 of the world's current copyright laws according to the degree to which they provide exceptions to copyright exclusivity for research uses.[6] Papers by Handke et al.[7] and Palmedo[8] contain data on of changing copyright exceptions over time in sets of 42 and 23 countries, respectively. However, there has not yet been a global review of copyright exceptions for researchers in most of the world's countries over time.

This early working paper introduces a new dataset of copyright exceptions for researchers in 165 countries over 21 years. Our review identifies the years in which relevant provisions of copyright laws were revised or replaced, analyzes current and older versions of the law, and provides coded data on the law for each country in each year.

Examining our sample of countries, we find that the *average* level of protection of users' rights to reproduce and share copyrighted works for research purposes remained fairly constant from 2000 through 2021. However, wealthier countries tended to have copyright exceptions that allowed researchers more freedom to reproduce and share works without authorization for research purposes. Former British colonies (with a history of fair dealing) also tended to have more robust copyright exceptions for researchers.

The paper proceeds as follows. Section two describes our review and the availability of older laws. Section three provides a brief explanation of our coding scheme. Sections four and five show where there is variation in our dataset and where it is lacking. Section six concludes.

II. AVAILABILITY OF LAWS

Our review of nations' laws relies heavily on the WIPO Lex database.[9] This is the most complete collection of intellectual property laws available, containing most countries' current laws. In some instances, current laws were not available in WIPO Lex for WIPO Member states, but we found the laws in other locations, such as Hein Online or national intellectual property offices. Augmenting our collection of laws with these secondary sources

[5] There is a slightly wider, but still narrow empirical literature on copyright exceptions more generally. Earlier PIJIP studies have analyzed changes to copyright exceptions over time in a set of 22-26 countries. Ghafele and Gibert analyze the impact of fair use in Singapore on industries that complement the creative industries.

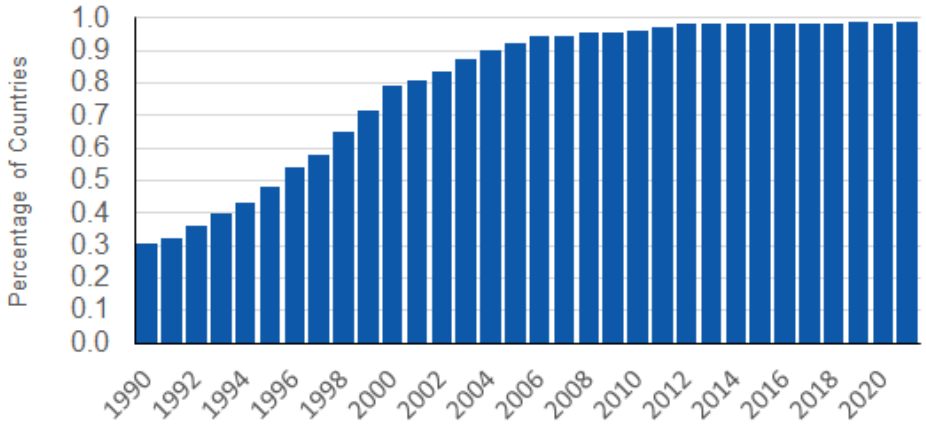
[6] Flynn, Sean; Schirru, Luca; Palmedo, Michael; and Izquierdo, Andrés. "Research Exceptions in Comparative Copyright." (2022) PIJIP/TLS Research Paper Series no. 75. https://digitalcommons.wcl.american.edu/research/75

[7] Handke, C., Guibault, L., & Vallbé, J.-J. (2021). Copyright's impact on data mining in academic research. Managerial and Decision Economics, 42(8), 1999– 2016. https://doi.org/10.1002/mde.3354

[8] Palmedo, Michael the Impact of Copyright Exceptions for Researchers on Scholarly Output. Efil Journal of Economic Research, vol. 2(6), 2019, pp. 114-139.

[9] https://www.wipo.int/wipolex/

Fig. 1: Data availability by year



yielded a set of copyright provisions from laws from 209 countries.

WIPO Lex also contains many older laws that preceded the current ones, but the inclusion of older ones is inconsistent from one country to the next, and it becomes quite uneven as one goes further back in time.

To build our dataset, we started with the current law, found the copyright exception most relevant to researchers, and coded it according to the coding scheme described below. Next, we looked at the most recent time the law had been amended (or replaced) and checked to see if the relevant provision had changed. We repeated this as far back as we could go until we could identify what the law on the books was in the year 1990. We created a large table containing the text of each country's relevant copyright exception as far back as 1990, where possible, or as far back as we could otherwise. Our table is included as Appendix 1 of this working paper

We did not find the legal texts for most countries all the way back to 1990. Figure 1 shows the percentage of countries we obtained the laws in each year from 1990 through 2021.

In 1990, we only have the copyright law that was in place for 30% of the countries in our set. The percentage increases as years pass, and by 2000 we have the copyright law that was in place for 79% of the countries in the set. We use this (admittedly arbitrary) cutoff point, and restrict our analysis to the 165 countries for which we can trace the evolution of copyright exceptions for researchers from 2000 to 2021. The list of countries in our subset is included as Appendix 2.

III. CLASSIFICATION SCHEME

This section gives an overview of the coding scheme used to evaluate countries' copyright exceptions for researchers. It is described in greater

Table 1. Summary of Coding Scheme

Color on map in Flynn et. al.	Coded score	What can researchers do with works, without authorization?
Green	5	Researchers can reproduce and share full works of any kind for research purposes
Blue	4	Researchers can reproduce full works of any kind for research purposes. But no sharing.
Light Blue	3	Researchers can reproduce full works, but only through a personal/private use exception.
Purple	2	Only institutions – such as libraries or educational establishments – can use the research exception.
Orange	1	Limits on the types of works that can be reproduced
Red	0	Only quotations or reproductions of short extracts.

detail in Flynn et al.'s working paper.[10] The scheme, presented as a system of color coding used in maps in earlier working papers and presentations, is presented in this paper as a six-point scale running from 0 to 5. Countries with the most restrictive laws (red in the color scheme) are coded 0, and countries with the most permissive laws (green in the color scheme) have a value of 5. The following subsections describe the meaning of each point along our scale, and table 1 provides a summary.

A. Most open to reproduction and sharing: 5 (Green)

These countries have laws that allow researchers to reproduce and share works with each other for research purposes, without authorization. This applies to all users, including individuals and institutions. It applies to all types of works, whether they are written works, artistic works, or other types. It also applies to works regardless of whether or not they have been published.

We do not consider whether the exception is limited to noncommercial uses, so a country can prohibit commercial uses under the exception and still be coded 5.[11]

Laws may express the ability to "share" reproductions with other

[10] Flynn, Sean; Schirru, Luca; Palmedo, Michael; and Izquierdo, Andrés. "Research Exceptions in Comparative Copyright." (2022) PIJIP/TLS Research Paper Series no. 75. https://digitalcommons.wcl.american.edu/research/75

[11] We consider noncommercial uses and not-for-profit uses to be the same type of uses in our review, though we note that some scholars have pointed out subtle differences between the two.

researchers in different ways. This includes the right to "communicate" or "make available."

Most fair use and fair dealing laws receive a coding of 5. However, some fair dealing laws restrict the types of works that can be reproduced and shared without authorization under the provision.[12] There are other types of closed lists that can result in a country being coded 5, as long as they allow users to reproduce and share works for research purposes.

B. Reproduction for research: 4 (Dark blue)

These countries' laws allow the unauthorized reproduction of works for research purposes, but do not permit researchers to share the reproductions with each other. The laws grant exceptions directly for the purpose of research, rather than broader fair uses or dealings, or for a more restrictive personal/private use exception. Since our coding scheme does not consider whether commercial uses are permitted, a country may be coded 4 even if it restricts uses under the exception to noncommercial use.

C. Private reproduction for individual use: 3 (Light blue)

This classification applies to countries that permit research uses through a personal or private use exception. These are non-commercial uses, and they do not permit sharing of copied works with other researchers.

D. Restrictions on Users / Institutional Right: 2 (Purple)

This classification applies to laws that restrict the types of users that can make unauthorized reproductions of full works for research purposes. These laws typically restrict the ability to make copies to institutions or their employees. Examples of such institutions are libraries, universities, and national research laboratories.

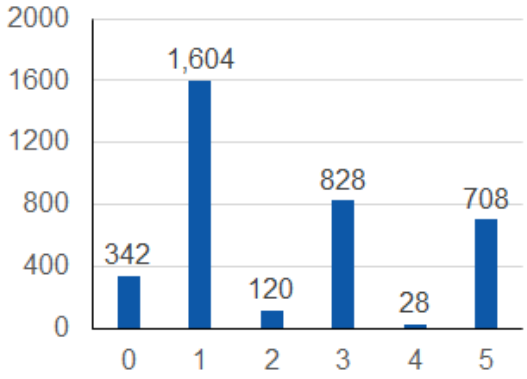
E. Restrictions on types of works: 1 (Orange)

This classification indicates that a country's copyright exception for researchers does not apply to all types of works, or carves out particular types of works. One common example is laws that grant users the right to make unauthorized reproductions for research purposes – but not the right to copy a full book (which we believe is essential for text and data mining).

Sometimes, a law will allow the reproduction of literary works, but not artistic, dramatic, or other types. Under the coding scheme for this paper, we consider any law that allows the reproduction of literary and artistic works, and has no other restriction on the type of works, to be unrestricted with respect to the types of works covered. However, a law that only applies to one or the other has a significant restriction on the type of works covered by the exception, and is coded 1.

[12] For example, South Africa's fair dealing clause applies to literary and musical works, but not to artistic works.

Fig. 2: Data Summary and Descriptive Statistics



Mean	2.27
Median	1.00
St. Dev.	1.68
Skewness	0.49
n	5,280

F. TDM Restricted: 0 (Red)

The most restrictive classification is reserved for countries that do not allow the unauthorized reproduction of any works. Most of these allow quotation of a one work in another, but quotation alone is not sufficient for modern research methods that rely on text and data mining. Some of these laws place specific quantitative limits on how much of a work – either a percentage or a number of pages – can be reproduced under a research exception. We find these limits to be severe, and code laws with a 0.

IV. DATA OVERVIEW

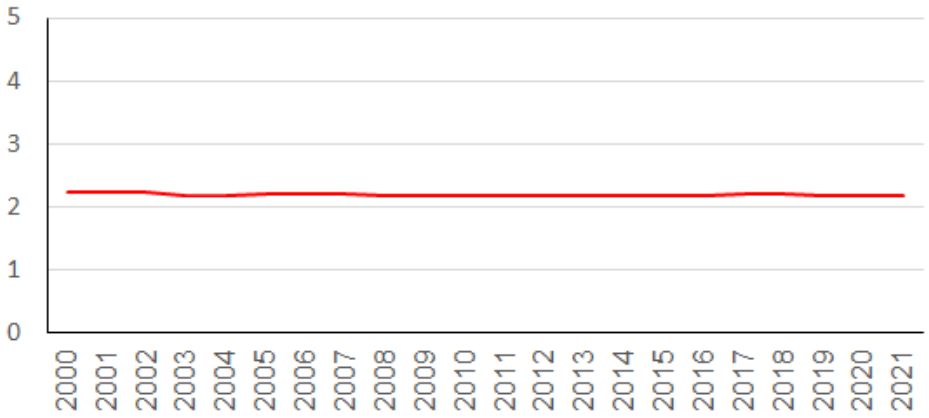
The coded data creates an ordinal dataset showing the strength of each country's copyright exception for researchers each year from 2000 through 2021. Each observation is the score for a country in a given year. If one organizes the countries alphabetically, the first observation is the coded law for Albania in 2000, the second observation is the coded law for Albania in 2001, and so on. The last observation is the coded law for Zimbabwe in 2021. In all, there are 5,280 country-year observations

The most common classification in our dataset is 1; many laws allow unauthorized reproductions for research purposes, without any right to share reproductions with other researchers, but restrict the types of works that can be reproduced. As shown in Figure 2, there are 1,604 country-year observations scored 1, the most in the dataset. There are very few observations coded either 2 or 4.

Most countries have copyright exceptions for researchers that allow some reproduction, but have significant restrictions on the exception. This can be seen in the mean score of 2.27, as well as the raw numbers in the bar graph.

Although there are many instances of laws changing between 2000 and 2021 – most countries' laws were amended at least once, many were amended more often – the average score for the dataset as a whole did not change much over time. Some countries' laws became more permissive of unauthorized

Fig. 3: Mean Score by Year



reproductions, and some countries' laws became less permissive. Figure 3 shows the average score for countries each year from 2000 to 2021. The average is remarkably constant.

V. VARIATION IN THE DATA

Though there is little variation in the aggregate score over time, one can look at the data in various ways to find variation in the dataset. This section demonstrates areas where there is variation between observations.

A. Income groups

Most variation in the data is across countries rather than over time. Wealthier countries, on average, are more likely to have copyright exceptions allowing greater unauthorized uses for research purposes than other countries. This is consistent with earlier reviews of copyright exceptions by PIJIP[13] and others[14]

Figures 4(a) and 4(b) show the average score each year for countries disaggregated by their World Bank income classification in 2000.[15] The

[13] Michael Palmedo, "A Novel Dataset Measuring Change in Copyright Exceptions," *Review of Economic Research of Copyright Issues.* Vol. 19(1), 2022, pp. 52-75; Sean Flynn and Michael Palmedo. "The User Rights Database: Measuring the Impact of Opening Copyright Exceptions," Joint PIJIP/TLS Research Paper Series no. 2018-01.

[14] Ruth L. Okediji, The Limits of International Copyright Exceptions for Developing Countries, 21 Vanderbilt Journal of Entertainment and Technology Law 689 (2020); Greg Walz-Chojnacki (interviewing Tomas A. Lipinski) "Copyright law and the implications for developing nations," UWM Report, February 22, 2017. https://uwm.edu/news/copyright-law-and-the-implications-for-developing-nations-tomas-lipinski/

[15] The World Bank classifies countries based on their income using Gross Nation Income (GNI) per capita in U.S. dollars. The cutoffs are updated each year. Historical classifications are here: https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups In 2000, the income classifications were determined using the following levels of GNI per capita: High Income countries had GNI per capita greater than $9,265; Upper-Middle Income countries had GNI per capita between $2,996 and $9,265;

Fig 4(a): Mean Score by Year and Income Classification

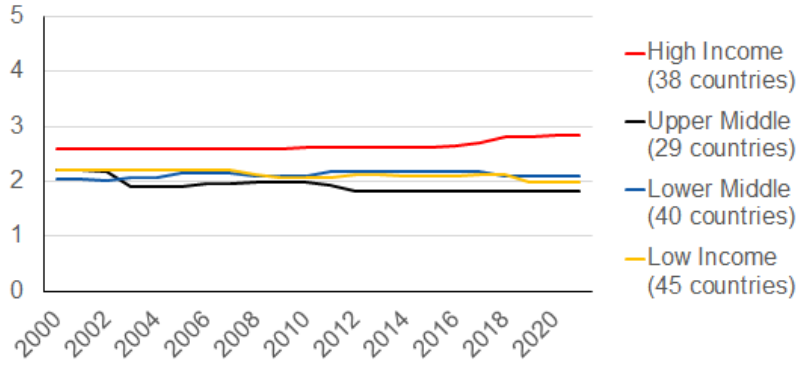
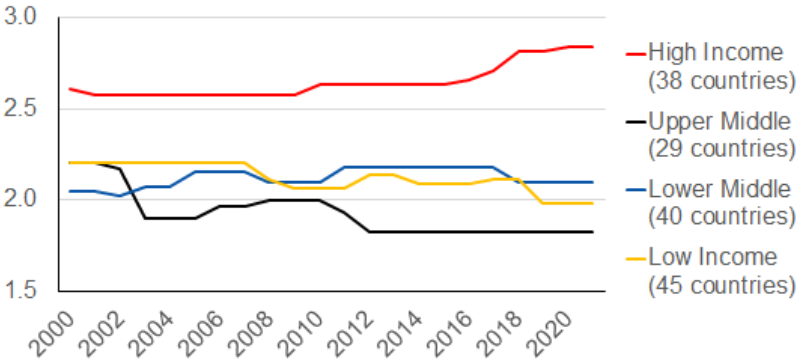


Fig 4(b): Mean Score by Year and Income Classification



figures demonstrate that countries classified as High Income had consistently higher scores (on average) than other countries over the entire time period reviewed.

Figure 4(a) shows the data with the vertical axis running its full range from 0 to 5. It emphasizes that countries in each development classification still tend to be "in the middle" on average in our scheme.

Figure 4(b) zooms into a 1.5 range on the vertical axis to better illustrate the relative changes over time. Between 2000 and 2021, there was a slight increase in the score for High Income countries. Some, such as Germany, passed new laws specifically for text and data mining.[16] Others, such as Japan introduced other types of changes that made exceptions more permissive of unauthorized reproductions for research purposes.[17] Figure 4(b) also shows a

---

Lower-Middle Income countries had GNI per capita between $756 and $2,995; and Low Income countries had GNI per capita less than or equal to $755.

[16] Copyright Act, as amended up to Act of September 1, 2017) https://wipolex.wipo.int/en/text/474263

[17] Act No. 48 of May 6, 1970, as amended up to April 28, 2020. https://www.wipo.int/wipolex/en/legislation/details/21342

Fig. 5(a) Former British Colonies v. Other Countries

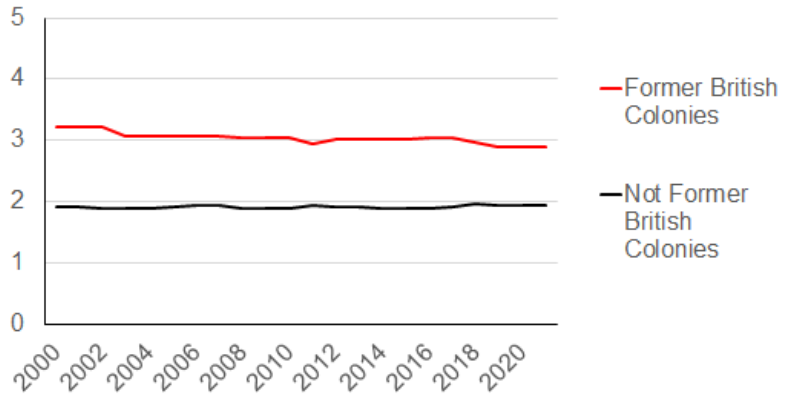
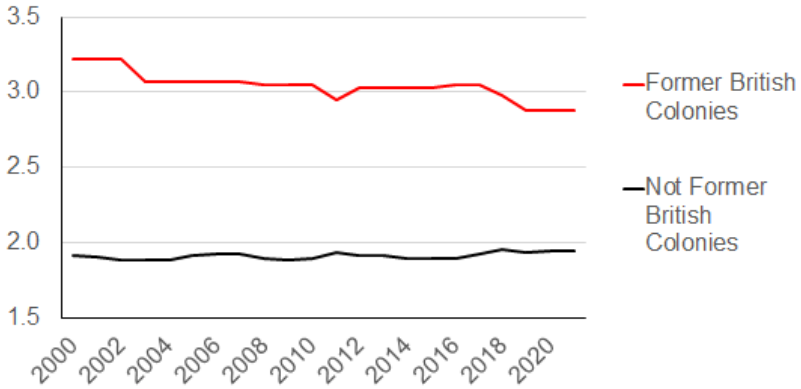


Fig. 5(b): Former British Colonies v. Other Countries



slight decrease for Upper-Middle Income countries, which were more likely to add restrictions to their laws. The average score for the subsets of Lower-Middle Income and Low Income countries changed less over this period.

B. The subset of Former British colonies

We see a notable change in the score in one group of countries: former British colonies. This group of countries tends to have laws based on the old colonial laws that included various versions of the British fair dealing clause. Our scoring system considers fair dealing to be generally permissive of the unauthorized reproduction and sharing of works, so the countries have a higher average score than others. This is illustrated in figure 5(a), which shows the difference using the full potential range of score values on the vertical axis.

As former British colonies amended their copyright laws, they often added restrictions to their exceptions for researchers. For example, before 2011, Grenada's copyright law allowed "fair dealing with a protected work, production, performance or edition for purposes of research or private

Table 2: Nations with change in their scores

Countries with Increasing Scores

Country	Change
Mauritius	4
Samoa	3
Yemen	3
Germany	2
Japan	2
Republic of Korea	2
Sri Lanka	2
Austria	2
Czech Republic	2
Netherlands	2
Ecuador	2
Switzerland	1
Albania	1
Kuwait	1
Luxembourg	1
Oman	1
Sao Tome and Principe	1

Countries with Decreasing Scores

Country	Change
Eswatini (Swaziland)	-1
Croatia	-1
Bulgaria	-1
Angola	-2
Kiribati	-2
Mali	-2
Rwanda	-2
Senegal	-2
Spain	-2
Tonga	-2
Vietnam	-2
Panama	-3
Antigua and Barbuda	-4
Dominica	-4
Grenada	-4
Myanmar	-4
Seychelles	-4

study."[18] The law placed no restrictions on this right. In 2011, it enacted a new act that instead allowed reproductions for personal use, subject to a series of restrictions on the types of works allowed.[19] Figure 5(b) illustrates this decline in the average score for former British colonies over the period. Their score falls from 3.2 to 2.8, while the score for non-British colonies remains flat.

C. The countries where we see change

While the average score of the 165 countries has held steady at around 2.27, we have established that there is *some* variation in the data. Table 2 shows which countries had increasing scores, and which had decreasing scores. Mauritius and Samoa had the largest increases, while Antigua and Barbuda, Dominica, Grenada, Myanmar, and Seychelles all had the largest

[18] Copyright Act (Cap. 67) https://wipolex.wipo.int/en/text/151135
[19] Copyright Act (Cap. 67, Act No. 21 of 2011) https://wipolex.wipo.int/en/text/378289

decreases.

VI. CONCLUSION

In this working paper, we have described our review of copyright exceptions for researchers and how these copyright exceptions have changed over time. We have explained our six-point coding scheme and trends in the data in a set of 165 countries' copyright laws.

Between 2000 and 2021, copyright exceptions for researchers tended to allow some unauthorized reproductions for research purposes, but with restrictions on the types of works that could be reproduced and/or restrictions on the ability of researchers to share their reproductions with each other. This is indicated by our mean score of 2.27, and by the small amount of deviation from the mean over this time period. The average score for High Income countries has been consistently higher, indicating less interference in researchers' rights to reproduce and share without authorization, and it has grown relative to the average scores of other income groups. Former British colonies also tend to have a higher score, though the average score of the subset of British colonies fell over the period observed.

Appendix 1 of this paper shows the text of copyright exceptions for each country. One can see the types of language commonly used in laws that grant varying levels of rights to researchers to make and share unauthorized reproductions of works for research purposes. It may be useful as a guide for drafting legislation to make copyright exceptions more accommodating for researchers who benefit from greater access to unauthorized reproductions.