Received: 15 December 2020 | Revised: 10 March 2021 | Accepted: 12 April 2021

DOI: 10.1002/mde.3354

RESEARCH ARTICLE

WILEY

Copyright's impact on data mining in academic research

Christian Handke[1,2] | Lucie Guibault[3] | Joan-Josep Vallbé[4]

[1]ESHCC, Erasmus University Rotterdam, Rotterdam, The Netherlands

[2]IViR, University of Amsterdam, Amsterdam, The Netherlands

[3]Schulich School of Law, Dalhousie University, Halifax, Nova Scotia, Canada

[4]GREL, University of Barcelona, Barcelona, Spain

Correspondence
Christian Handke, ESHCC, Erasmus University Rotterdam, Rotterdam, The Netherlands.
Email: handke@eshcc.eur.nl

With the proliferation of digital data, data mining (DM)—in the sense of the discovery of valuable structures in large sets of data—is expected to increase the productivity of many types of research. This paper discusses how copyright affects DM by academic researchers. In some territories, academic DM is lawful if researchers have lawful access to input works. In other territories such as the European Union, lawful DM additionally requires specific consent by rights holders. Based on bibliometric data and quasi-experimental research designs, we show that where academic DM requires specific rights holder consent: (1) DM publications make up a significantly lower share of total research output, and (2) stronger rule of law is associated with less DM research. To our knowledge, this study is the first to empirically document an adverse effect of intellectual property (IP) on innovation under particular circumstances. There is strong evidence that copyright exceptions or limitations promote the adoption of DM research.

1 | INTRODUCTION

This paper discusses the effect of copyright on data mining (DM) by academic researchers. Hand et al. (2001) broadly define DM as "the discovery of interesting, unexpected or valuable structures in large datasets."[1]

With the proliferation of digital data, DM is widely expected to increase the productivity of many types of research activities and to become a main driver of economic growth (Einav & Levin, 2014; OECD, 2014, 2015; Varian, 2014). For an overview of DM applications in various aspects of the economy, see Dean (2014). That DM already has commercial value and contributes to economic growth is easily illustrated. DM plays an important role in eliciting value from data, and for instance, according to a recent report for the European Commission, the "overall impact of the data market on the economy as a whole" in 27 European Union (EU) Member States was €325 billion in 2019, up 7% since 2018, and accounting for 2.6% of GDP (Cattaneo et al., 2020, p. 13). What is more, among the 10 most valuable companies in 2015 according to Fortune 500 (Gandel, 2016), at least two were founded quite recently as suppliers of "free" online services—Alphabet (formerly Google) ranked second and Facebook ranked fifth—and initially relied on the collection and analysis of user data for generating rapidly growing revenues.[2] Academic research is another area in which DM is expected to foster value creation, and as we will show, DM has been the topic of an increasing share of total academic research output over the last two decades.

Copyright relates to a trade-off regarding DM. Effective copyright protection should increase the supply of potential DM input works but can also increase the costs of using existing data for those not holding relevant copyrights. DM by means of digital information and communication technology (ICT) technically requires the reproduction of input works and may thus fall under copyright, even if only aspects of individual input works are relevant for a DM project. We analyze bibliometric data to establish how various copyright policies affect the application of DM in academic research. We show that in countries in which DM for academic research requires the express consent of rights holders, DM-related articles make up a significantly smaller share of total research output.

How copyright is applied to DM will continue to affect many academic researchers in coming decades. The evidence presented in this paper relates to a policy debate in particular in the EU. Under current EU legislation, DM requires prior authorization of rights

holders even if the potential user has lawful access to the research articles and databases in question (Directive 2001/29/EC, 2001, art. 3 and 5). The situation will change with the implementation, by June 7, 2021, of Article 3 of Directive (EU) 2019/790 which expressly allows text mining and DM to take place for the purposes of scientific research carried out by research organizations and cultural heritage institutions. The United States have a more permissive copyright policy regarding DM and recent rulings seem to confirm greater scope for DM without express consent by rights holders.[3] Other countries like the United Kingdom and several Asian countries have recently introduced relatively permissive legislation, the application of which will probably be defined further in the courts. As yet, the situation is uncertain for many academic researchers and other stakeholders.

2 | THEORY

DM is a novel technology to conduct research. According to standard economic theory, researchers will conduct DM as long as expected returns exceed the opportunity costs of the best alternative allocation of researchers' resources. The uptake of DM should be affected by demand conditions, the price and characteristics of inputs (including suitable data) as well as of related goods and services, the conditions of production, competition, and government policies and regulations including relevant aspects of intellectual property (IP).

However, incentive schemes for academic research often diverge from typical markets (Dasgupta & David, 1994) so that an application of production theory is not entirely straightforward. In particular, there is no conventional demand formation and thus only incomplete market coordination. Academic research has public good attributes, and in many territories, it is largely financed through public means. Academic researchers' returns depend less on sales of research output but come in the form of research funding and long-term employment with prestigious universities or research institutes. These types of returns hinge on peer recognition for which the publication record is central.

We assume that researchers seek to maximize the (quality-adjusted) number of articles they publish by employing the most efficient technologies available to them. As with any new technology, there may be uncertainty, and DM uptake per country may be affected by the specific characteristics of domestic researchers and research organizations. Nevertheless, in the aggregate choices of researchers between various technologies should provide the best available indication of the optimal allocation of resources under specific circumstances within countries. This paper documents the effect of copyright law on this choice.

Like other types of IP, economists often address copyright as a means to mitigate market failure in the private provision of goods with public good attributes (Arrow, 1962; Novos & Waldman, 1984; Samuelson, 1954). The explicit aim is to promote the supply of valuable copyright works by endowing those investing in the development of relevant works with temporary market power. Effective copyright protection has ambiguous effects on the supply of new creative works: on the one hand, it increases returns to rights holders; on the other, stronger copyright protection increases the total cost of input works to potential DM researchers due to higher prices and greater transaction costs compared with a situation where data are available without an explicit, additional license from the rights holders (Landes & Posner, 1989). From a welfare economic perspective, copyright thus fights fire with fire: it mitigates one source of market failure (underprovision of public goods) with another (market power and underutilization of public goods).

Our empirical work is based on several related assumptions. First, DM is often conducted by researchers, who are not the rights holders of all adequate data.[4] Second, DM by academic researchers increases in the quantity and quality of supply of suitable data. Third, academic DM decreases in the costs of accessing relevant data. Fourth, effective copyright protection affects the supply of suitable data and/or the full economic costs of accessing data and conducting DM. We thus hypothesize that variations in relevant copyright policy between countries will affect the amount of DM by researchers residing in those countries. Because copyright has ambivalent effects on follow-up use of protected works, the direction of copyright's effect on DM is unclear at the outset.

3 | DATA

Tables 1 and 2 give an overview of variables used in this paper.[5] One important measure of research output is the number of academic journal articles published. We collected data from Thomson Reuter's Web of Science (WoS), using the entire WoS Core Collection Database including the so-called Science Citation Index Expanded, Social Science Citation Index and Art & Humanities Citation Index.

To generate the variable "DM Output," we extracted the number of all published research articles on DM from 42 large economies.[6] The Boolean searches on the WoS database were defined by three simultaneous restrictions: (1) "data mining" entered in inverted commas in the field "Topic"; (2) a country name according to the format used on WoS in the field author's "Address," which relates to the country of residence of the first or main author; and (3) a year of publication in the field "Year Published." Search results were further restricted by ticking the option "Articles" in the user interface of WoS, so that results only contain academic journal articles rather than conference proceedings, book reviews, and the like. For each country and year, we recorded the number of different items in the WoS database that fulfill these search criteria. Our panel includes the 15 largest EU Member States, as well as the 27 largest other economies based on national GDP in 2013 according to the World Bank. The data covers the years 1992 to 2014. WoS includes articles published since 1975. It contains no articles on DM published before 1992. We thus have 966 country-year observations. In the data analysis, some countries had to be excluded because they could not be classified in terms of relevant copyright provisions. The articles featured in the search results contain DM applications and related conceptual and

TABLE 1 List of variables used in data analysis

Short variable name	Description (all data per country and year)	Source	Additional information
DM Output	Number of DM-related research articles	Own collection on Web of Science	
Research Output	Total number of research articles	Own collection on Web of Science	
DM Share (‰)	Dependent variable; quotient of "DM Output" and "Research Output" times 1000	Own calculations; derived variable	
Copyright (a) Consent Required (b) Probably Required (c) Probably not Required (d) Not Required Categories (b) to (d) are combined in some cases	Ordinal variable and main predictor; categorization of countries according to whether specific consent of rights holders is required for academic DM research to be consistent with copyright law	Own classification	Categories (b) to (d) merged in some regressions so that we get a binary distinction between "Consent Required" and "Not Definitely Required"; for full documentation see Table 3 and the Supporting Information
Switch	Dummy variable for seven countries that changed from "Probably Required" (Code 0) to "Probably not Required" (Code 1)	Own classification; derived from "Copyright"	See Table A2
Rule of Law	Index for law abidance; normalized to a mean of 0 and standard deviation of 1; scaled between −2 and 2	World Bank (2015a, 2015b)	No data before 1996 nor for 2014[a]
Population	Number of inhabitants	World Bank (2015a, 2015b)	
Broadband (%)	Share of households with broadband internet subscription	World Bank (2015a, 2015b)	360 missing observations; no data between 1992 and 1997 and very incomplete until 1999; no data available for Taiwan
GDP/capita ($1000)	Gross domestic product per inhabitant for country	World Bank (2015a, 2015b)	52 missing observations; no data for 2014, Argentina (2007–2014), Taiwan (2011–2014)
EU	Dummy for members states of the European Union (EU) or of the European Economic Area (EEA) or countries preparing for EU accession during the year in question		See Table 3

[a]Until 2003, we only have values for alternate years. To avoid loss of data and given the generally low variation of this indicator, the scores for 1997, 1999, and 2001 for each country were estimated computing the arithmetic mean of the rule of law score in the previous and posterior year.

TABLE 2 Descriptives for variables used in data analysis

Variable	*N*	Mean	SD	Min	Max
DM Output	966	19.090	46.409	0	396
Research Output	966	24,640	45,453	65	368,469
DM Share (‰)	966	0.620[a]	0.781	0.000	7.937
Rule of Law (scaled between −2 and 2)	725	0.704	0.998	−1.790	2.000[b]
Population ('000)	924	107,595	249,045	2013	1,357,380
Broadband (%)	606	13.191	12.488	0.000[c]	42.562
GDP/capita ($1000)	914	22,249	16,960	0.413	67,805

[a]This is the unweighted average of averages per country and year. Across the entire panel, the average "DM Share" is 0.77‰.
[b]The maximum "Rule of Law" score in our panel was 1.9996 for Denmark in 2007.
[c]The minimum for "Broadband" is 0.0003 for Nigeria in 2005.

methodological work. Among the countries covered, searches on WoS brought up 18,441 DM-related articles between 1993 and 2014.

We also collected data on the total number of research articles published for the same set of countries and years to generate the variable "Research Output." Search parameters were the same as reported above, except that no "Topic" was specified. This brought up 23,802,650 articles for the entire panel. Over the 22 years covered, 0.77‰ of all articles had DM as a topic.

In our empirical analysis, for each country and year, we used the ratio of "DM Output" and "Research Output" as the dependent variable, multiplied by 1000 to avoid dealing with very small fractional numbers. This variable is referred to as "DM Share." We thus mitigate one of the major problems in using bibliometric data to assess countries' relative performance: varying degrees of coverage over different countries, languages, or academic disciplines. It is well documented that WoS (and other major research databases) cover English-language publications most comprehensively. The output of countries where many researchers publish in other languages is thus easily underestimated (Mongeon & Paul-Hus, 2016; Van Raan et al., 2011). By contrast, with our dependent variable "DM Share," we can produce valid comparisons between groups of countries as long as the relatively weak assumption holds that the probability of DM-related articles featuring in WoS compared with the probability of other articles does not systematically and strongly deviate over time between different copyright categories.

Yearly scores for "DM Output" and "DM Share" have increased substantially since 1992. See Figures 1 and 2 for an illustration and Appendix A for an overview of the data by country.

We classify countries according to the type of copyright law that applies to DM, similar to an approach pioneered by Ginarte and Park (1997). See the Supporting Information for a detailed discussion. We use two aspects of the copyright system: (1) whether copyright exceptions or limitations are in place that could apply to DM by academic researchers who have lawful access to potential input works and (2) whether there is relevant case law specifying the applicability of existing exceptions and limitations. Table 3 gives an overview of the four country categorizations from 1992 to 2014 according to DM-related copyright law. Table 4 presents descriptive data regarding "DM Share" in the copyright categories.

There is often a discrepancy between IP law and social practice, because IP is hard to enforce.[7] We therefore incorporate a "Rule of Law" indicator as reported by the Worldwide Governance Indicators (WGI) project (World Bank, 2015a, 2015b) and documented in Kaufmann et al. (1999) and Kaufmann et al. (2010). This indicator is defined as "the extent to which agents have confidence in and abide by the rules of society" (World Bank, 2015b), including the quality of contract enforcement and property rights. We use it as a proxy for the level of enforcement of quasi-property rights such as copyright.[8] We further use GDP per capita, population size, and broadband penetration as control variables. The raw data are available in a replication data set.

To prepare our econometric analysis, Figure 3 displays differences between the average "DM Share" of 23 countries in the "Consent Required" copyright category and 14 countries with more permissive copyright legislation (mostly "Probably Required" and "Probably not Required"; nonclassifiable countries excluded) represented by the

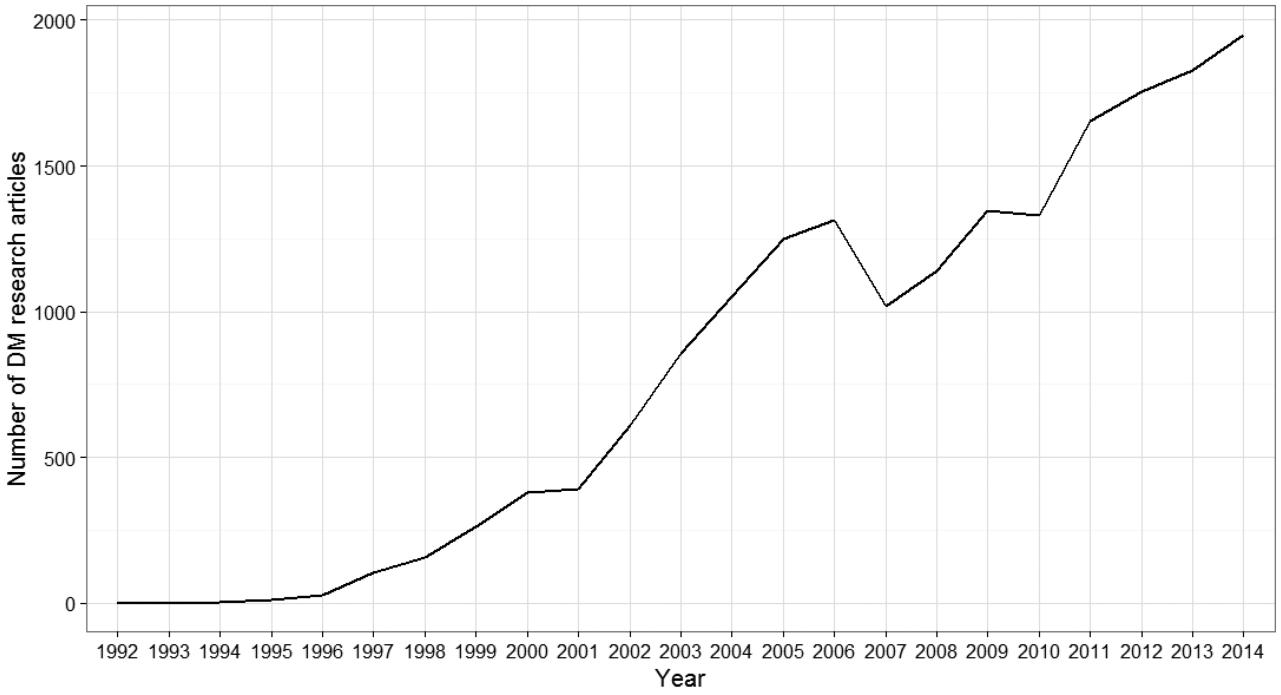

FIGURE 1 The absolute number of data mining (DM) research articles published per year (42 countries; 1992 to 2014).
Sources: Own calculations based on search results on the WoS database

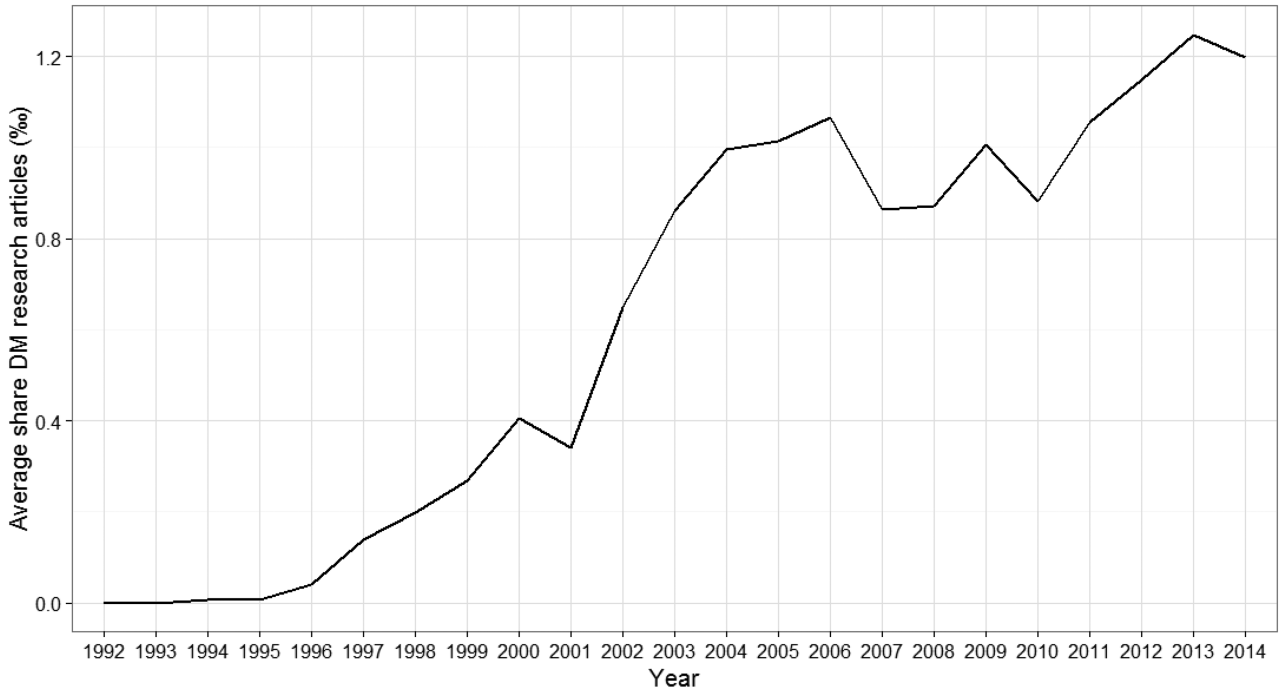

FIGURE 2 The average share of data mining (DM) research articles in the total number of research articles published ("DM Share") per year and country in ‰ (42 countries; 1992 to 2014).
Sources: Own calculations based on search results on the WoS database

zero line. We further distinguish countries from the "Consent Required" category by their "Rule of Law" scores. Copyright law should have a stronger effect in countries with stronger enforcement or a greater cultural propensity to adhere to legal norms. The black line represents 12 countries with "Rule of Law" scores greater than 1.2 during any year for which data is available. The gray line represents 11 countries with "Rule of Law" scores lower than 1.2. Between 1996, when DM publications gradually became more numerous, and 2014, all countries from the "Consent Required" category display relatively low "DM Share" (−37% with "Rule of Law" > 1.2 and −28% with "Rule of Law" < 1.2). Since 2005, "Consent Required" countries with low "Rule of Law" seem to be catching up with countries from other copyright categories, and in 2014, their average "DM Share" was 8% below that in countries in different copyright categories. "Consent Required" countries with higher "Rule of Law" do not exhibit any consistent trend towards catching up. In 2014, the "DM Share" in "Consent Required" countries with a high rule of law was 37% lower than in countries in other copyright categories. This descriptive analysis provides some indication that DM by academic researchers is sensitive to observed variations in copyright law and that this effect is moderated by the rule of law within countries. We address these issues more systematically in the econometric analysis in Section 5.

4 | RESEARCH DESIGN

We adopt quasi-experimental research designs, with "DM Share" as dependent variable, the copyright category "Consent Required" as control group, and other copyright categories as treatments. There is no verifiable random assignment of treatments across our panel.[9] We thus construct several complementary quasi-experiments, each with its own strengths and weaknesses as a means to test for the effect of copyright on DM research (Meyer, 1995; Shadish et al., 2002). To mitigate challenges to validity, we also make use of control variables, multilevel models, interactions between independent variables, and difference-in-difference (DID) models exploiting the panel structure of our data. The specific quasi-experimental setups and their relative merits are discussed in Section 5.

5 | DATA ANALYSIS

5.1 | Multilevel regressions with the full copyright categorization

In a first quasi-experimental design, we use all four copyright categories with "Consent Required" as reference category/control group. There are virtually no pretest observations, as only one territory switched from "Consent Required" to any other copyright category (England in 2014). Observations were excluded when a territory was nonclassifiable for any year. In the time period covered, eight countries switched between other copyright categories: six countries switched at various times from "Probably Required" to "Probably not Required"; Japan switched from "Probably Required" to "Not Required" in 2010; England (listed separately from other parts of the United Kingdom) switched from "Consent Required" to "Not

TABLE 3 Categorization of countries according to whether DM requires express consent by rights holders to be legal

Copyright category	Relevant copyright exception or limitation?	Relevant case law?	Applies to: *(for the years 1992 to 2014)*
"Consent Required" DM requires specific consent of rights holders	There is a closed list of exceptions and limitations (what is not explicitly allowed is infringing on copyright); no relevant exception for DM by academic researchers.	None	- All EU/EEA Member States, except for the United Kingdom since 2014. - Countries preparing for EU accession: Austria, Finland, and Sweden (accession in 1995) as well as Poland since 1996 (accession in 2004). - All Latin American countries covered - Switzerland - Turkey
"Probably Required" DM probably requires express consent, but there has been no ruling against DM researchers	There is a fair dealing defense that could potentially cover DM.	There is no relevant case law specifying whether DM qualifies as an act of fair dealing.	- Australia - Canada before 2012 - China 2007–2011 - India - Israel before 2008 - Japan before 2010 - Korea before 2011 - Malaysia - Nigeria - Singapore before 2005 - South Africa - Taiwan before 2003 - Thailand
"Probably not Required" Lawful access is probably sufficient, but there has been no ruling in favor or DM researcher	There is a fair use defense that could be used to justify DM without express consent.	There is no relevant case law specifying whether DM qualifies as an act of fair use.	- Canada since 2012 - China since 2012 - Israel since 2008 - Korea since 2012 - Singapore since 2005 - Taiwan since 2003 - USA
"Not Required" Lawful access entails the right to apply DM	There is either a relevant copyright exception that applies explicitly to DM by academic researchers …	… and/or relevant case law has established that DM by academic researchers is an act of fair use.	- Japan since 2010 - England since 2014
Not classifiable			- China before 2007 - Indonesia - Iran - Poland before 1996 - Russia Saudi Arabia United Arab Emirates

TABLE 4 Descriptive statistics regarding the DM share in copyright categories

Copyright category	Average DM share in ‰	Standard deviation	Number of observations
Consent Required	0.56	0.56	546
Probably Required	0.59	0.66	222
Probably not Required	1.78	1.43	57
Not Required	0.60	0.17	6

Note: Switching countries included; see Appendix A for a per country overview. Source: Own calculations based on search results on the WoS database.

Required" in 2014 (see Table 3). Thus, bias due to self-selection and simultaneity is a concern in this first setup, but it does capture relatively much variation in copyright law.

Table 5 reports pooled ordinary least squares regressions and multilevel regressions with random country effects (varying intercepts by country). The number of observations is reduced in models with

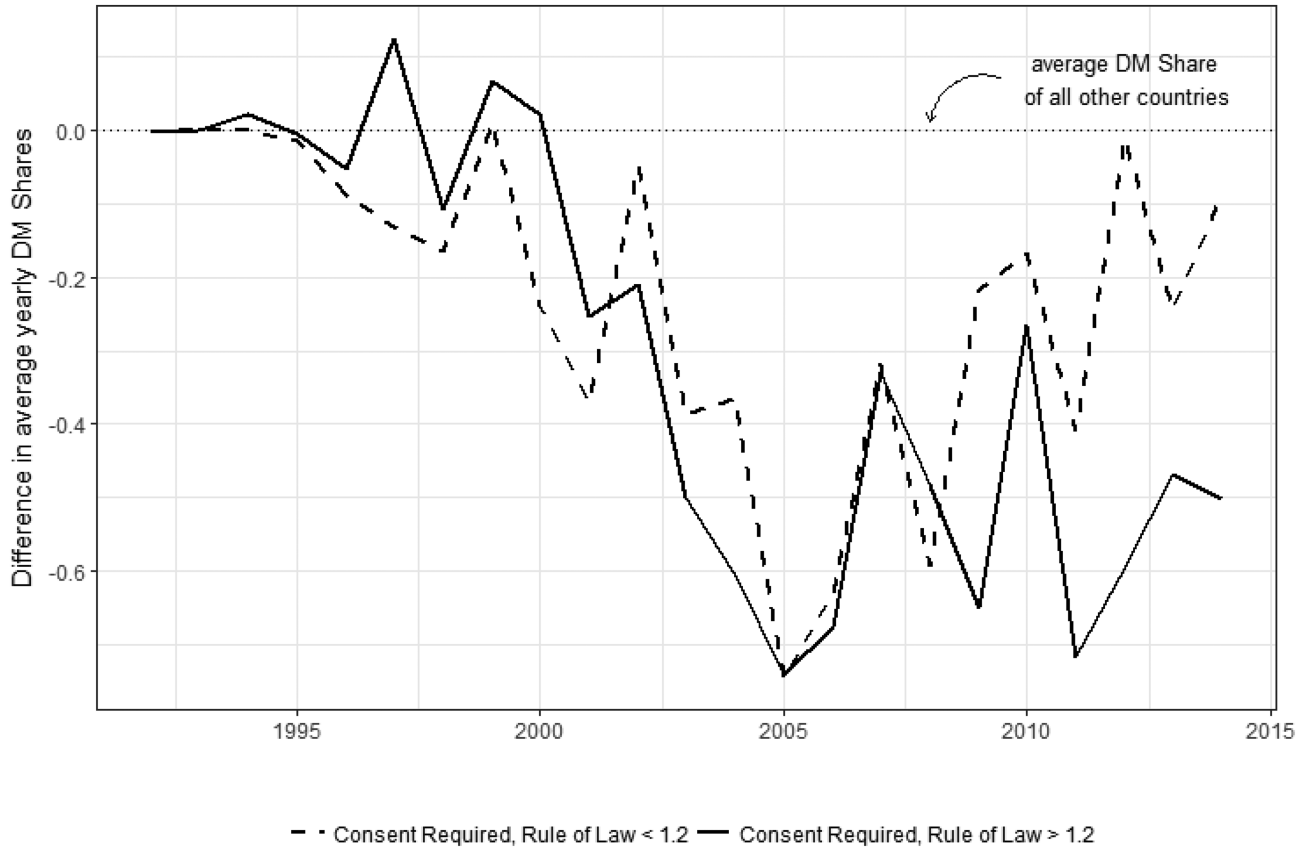

FIGURE 3 The difference in "DM Share" (‰) between "Consent Required" countries and countries in all other copyright categories, subject to "Rule of Law" (unclassifiable countries excluded)

control variables, because of missing data for instance on "Rule of Law" and "Broadband."[10] As expected, the "Not Required" category rarely yields significant coefficients: it contains merely six observations and we report it only for completeness.

"Probably not Required" yields significant positive coefficients in Models 1a, 1b, 2a ($p < .01$), and 2b ($p < .05$). This suggests that a more permissive copyright framework is associated with more DM research. "Probably Required" only yields significant coefficients at the .05 level in Models 1a and 2a, without random effects. Coefficients for "Probably not Required" are consistently larger than for "Probably Required." Results are in line with our ordinal categorization: there is a stronger and more reliably significant coefficient for the category that differs more from the reference category "Consent Required." These results suggest that DM share is lower in countries in the "Consent Required" category than in countries with more permissive DM-related copyright.

In Models 2a and 2b, the log-transformed total "Research Output" has a positive and significant coefficient. Countries with a high share of DM articles in total research output also tend to have larger total research output. There is no indication that DM would reduce incentives for other types of research within the same country.[11] EU membership captured by the variable "EU15" has no significant effect. Apparently results hold throughout the "Consent Required" category.

5.2 | Multilevel regressions with a binary copyright categorization

In a second setup, we use a binary distinction between "Consent Required" countries (control group) and countries in all other copyright categories, referred to as "Not Definitely Required," as a single treatment group. As in the first setup, there are no useful pretest observations, but there is a closer approximation of random assignment. The distinctive feature of countries in the "Consent Required" category is that there is a closed list of copyright exceptions or limitations that does not contain any provisions for DM. Thus, in these countries, DM without specific consent by rights holders is definitely in breach of copyright. All other countries have at least an open list of copyright exceptions and limitations that could apply to DM, see Table 3. This difference of closed and open lists of exceptions and limitations between national copyright systems was established decades before the concept of DM emerged and with a view to very different issues. Furthermore, with the exception of a single observation (England in 2014), there was no switching in the binary categorization over the 23 years covered. In this setup, we can thus virtually exclude bias due to self-selection or simultaneity.[12] Because of the permanence of copyright status, lagged effects can hardly bias results either. Furthermore, the dependent variable "DM Share" has a score of 0 for all countries any time before the period investigated, so that

TABLE 5 Regressions with of "DM Share" as dependent variable and the full copyright categorization

	Pooled (1a)	Multilevel (1b)	Pooled (2a)	Multilevel (2b)
Copyright				
Not Required *(only 6 observations)*	−0.504** (0.234)	−0.339 (0.224)	−0.072 (0.286)	0.113 (0.302)
Probably not Required	0.959*** (0.080)	0.944*** (0.146)	0.665*** (0.114)	0.539** (0.210)
Probably Required	0.100** (0.045)	0.089 (0.124)	0.160** (0.073)	0.264 (0.172)
GDP/capita			−0.025*** (0.003)	−0.017** (0.007)
Population (log)			−0.280*** (0.042)	−0.285*** (0.080)
Rule of Law			0.168** (0.072)	−0.045 (0.129)
Research output (log)			0.176*** (0.044)	0.190** (0.074)
Broadband			−0.004 (0.003)	0.006 (0.004)
EU			0.072 (0.071)	0.154 (0.176)
Year	0.059*** (0.003)	0.059*** (0.002)	0.047*** (0.008)	0.026** (0.011)
Constant	−117.407*** (6.101)	−117.067*** (4.947)	−89.994*** (17.079)	−47.134** (21.789)
Observations	831	831	459	459
R^2	0.427		0.332	
Marginal R^2		0.419		0.244
Conditional R^2		0.642		0.569
Adjusted R^2	0.424		0.318	
Log likelihood		−569.733		−276.863
Akaike Inf. Crit.		1153.467		579.726
Bayesian Inf. Crit.		1,86.525		633.404
Residual standard error	0.567 ($df = 826$)		0.476 ($df = 448$)	
F statistic	153.97*** ($df = 4; 826$)		22.31*** ($df = 10; 448$)	

*$p < .1$.
**$p < .05$.
***$p < .01$.

there is no concern with prior trends. The major challenge to validity in this setup are omitted variable bias and the crude categorization of copyright into two types only, which does not fully capture all relevant variations in the treatment on which information is available.

In Table 6 coefficients for "Not Definitely Required" are consistently positive. Without random effects in Models 1a and 2a, "Not Definitely Required" yields significant positive coefficients ($p < .01$). With random effects and thus better control for constant, unobserved country differences, Model 1b yields no significant effect of "Not Definitely Required." Model 2b with further controls but fewer observations yields a weak significant coefficient for the same dummy variable ($p < .1$). These results suggest that there is a weak positive effect on DM Share when academic researchers are not definitely obliged to acquire specific consent of rights holders to conduct lawful DM. Results are not as conclusive as in Table 1. This may be due to the greater number of observations for the "Probably Required" category (222)—which differs less from "Consent Required" and has no consistent effect according to results displayed in Table 5—than for the "Probably not Required" category (57) and "Not Required" (6), which differ more from "Consent Required." There could also be less bias due to self-selection and simultaneity in this quasi-experimental setup than in the results presented in Table 5. However, incorporating interactions between copyright categories and "Rule of Law" leads to a different result, as discussed in the following section.

5.3 | Multilevel regressions with interaction terms between copyright categories and "Rule of Law"

Among our control variables, of particular interest is "Rule of Law" as a proxy for the enforcement of and cultural propensity to adhere to legal norms. Greater rule of law should make copyright law more effective. To test for this, Table 7 includes a multiplicative interaction between copyright categories and the rule of law indicator. In these models, the coefficients of the variables that constitute the interaction (the categories of copyright regulation and "Rule of Law") are no longer to be interpreted as unconditional marginal effects.[13] The main coefficients of interest in these models are those for the interaction terms, which illustrate any moderating effect of "Rule of Law" on the association between "DM Share" and copyright categories.

Table 7 presents results for the full copyright categorization and for the binary copyright categorization with "Consent Required" as the reference category. All models yield significant, positive coefficients for multiplicative interaction terms (but not for all copyright categories in Models 1a and 1b). With interaction terms, the coefficients for

TABLE 6 Regressions with of "DM Share" as dependent variable and the binary copyright categorization between "Consent Required" and "Not Definitely Required"

	Pooled (1a)	Multilevel (1b)	Pooled (2a)	Multilevel (2b)
Not Definitely Required	0.259*** (0.044)	0.205 (0.132)	0.234*** (0.072)	0.300* (0.177)
GDP/capita ($1000)			−0.023*** (0.003)	−0.015** (0.007)
Population (log)			−0.302*** (0.042)	−0.275*** (0.081)
Research output (log)			0.214*** (0.044)	0.190** (0.075)
Rule of Law			0.137* (0.073)	−0.054 (0.131)
Broadband			−0.005 (0.003)	0.005 (0.004)
EU			0.034 (0.072)	0.128 (0.182)
Year	0.064*** (0.003)	0.064*** (0.002)	0.049*** (0.009)	0.027** (0.011)
Constant	−126.905*** (6.326)	−127.044*** (4.955)	−93.908*** (17.405)	−50.564** (21.623)
Observations	831	831	459	459
R^2	0.349		0.296	
Marginal R^2		0.338		0.218
Conditional R^2		0.605		0.572
Adjusted R^2	0.347		0.284	
Log likelihood		−608.473		−277.504
Akaike Inf. Crit.		1226.946		577.007
Bayesian Inf. Crit.		1250.559		622.427
Residual standard error	0.603 ($df = 828$)		0.487 ($df = 450$)	
F statistic	221.818*** ($df = 2; 828$)		23.679*** ($df = 8; 450$)	

*$p < .1$.
**$p < .05$.
***$p < .01$.

copyright categories are hardly significant. This suggests that where "Rule of Law" is 0, and thus lower than in most countries in our panel, copyright protection has no effect on "DM Share." However, as "Rule of Law" increases, countries in the "Consent Required" reference category exhibit a lower "DM Share." Overall, the results in Table 7 suggest that "Rule of Law" moderates the effect of restrictive copyright law. In particular the combination of strong copyright law and strong rule of law reduces academic researchers' DM performance.

Figure 4 provides further illustration. Based on Model 2b in Table 7, it plots the marginal effects (due to interaction) of "Rule of Law" on the coefficient for "DM Share" for countries in the "Not Definitely Required" copyright category with "Consent Required" as reference category. There is no significant difference for countries with low levels of "Rule of Law." With "Rule of Law" scores of about 0.6 and higher (just above the scores of Italy and Malaysia for much of the time period covered), countries in the "Not Definitely Required" category exhibit significantly higher "DM Share" than "Consent Required" countries.

5.4 | The effects of switching between copyright categories

In a fourth quasi-experimental setting, we document the effects of several switches (treatments) from "Probably Required" to "Probably not Required." Only this type of switch has occurred frequently enough for us to meaningfully address its consequences; see Table A2 for a list of the switching countries and some of their characteristics.[14]

Table 8 reports the results for DID regressions with two dummies: "Switch-Yes" marking all full calendar years after a switch in copyright category and "Switcher-Yes" marking all countries that underwent the relevant switch at some point in time. We use two nonequivalent control groups to check whether results are consistent. First, in Models 1a and 1b, we use all 13 countries that were initially in the "Probably Required" copyright category (except for Japan, who underwent a different type of switch).[15] Second, in Models 2a and 2b, we use all 37 countries within our panel that be classified into copyright categories (excluding Japan and England, who underwent other switches).[16] With these two panels, we can isolate the effects of switching on "DM Share" controlling for (1) prior trends in countries that switched from "Probably Required" to "Probably not Required" over the period investigated; (2) pretreatment and posttreatment changes in nonswitching countries from the "Probably Required" category; and (3) changes in all countries for which data are available. In contrast to the other experimental setups reported on in Tables 5–7, here, there are useful pretest observations. DID is a relatively effective means to mitigate challenges due to endogeneity.

The main independent variable of interest is "Switch-Yes," which yields significant and positive coefficients in all models ($p < .01$). (This also holds where Japan and England, who underwent other switches, are included.) Switches from "Probably Required" to "Probably not

TABLE 7 Regressions with "DM Share" as dependent variable and interactions between copyright and "Rule of Law"

	Full categorization		Binary categorization	
	Pooled (1a)	Multilevel (1b)	Pooled (2a)	Multilevel (2b)
Copyright				
Not Definitely Required			−0.032 (0.080)	0.075 (0.179)
Not Required *(only six observations)*	9.460 (25.358)	9.565 (21.474)		
Probably not Required	−0.865* (0.495)	0.198 (0.506)		
Probably Required	−0.021 (0.080)	0.073 (0.176)		
GDP/capita ($1000)	−0.022*** (0.003)	−0.016** (0.007)	−0.020*** (0.003)	−0.014** (0.006)
Population (log)	−0.213*** (0.045)	−0.213*** (0.083)	−0.192*** (0.044)	−0.203** (0.081)
Research output (log)	0.102** (0.047)	0.122 (0.077)	0.093** (0.046)	0.121 (0.076)
Rule of Law	0.100 (0.070)	−0.102 (0.128)	0.091 (0.070)	−0.121 (0.129)
Broadband	−0.005 (0.003)	0.004 (0.004)	−0.006** (0.003)	0.003 (0.004)
EU	0.250*** (0.076)	0.344* (0.181)	0.261*** (0.077)	0.349* (0.185)
Year	0.057*** (0.008)	0.033*** (0.011)	0.061*** (0.008)	0.034*** (0.011)
Interactions				
Not Definitely Required * Rule of Law			0.434*** (0.066)	0.413*** (0.144)
Not Required * Rule of Law	−7.179 (19.286)	−7.013 (16.332)		
Probably not Required * Rule of Law	1.160*** (0.328)	0.421 (0.357)		
Probably Required * Rule of Law	0.351*** (0.070)	0.369** (0.149)		
Constant	−110.159*** (17.088)	−63.037*** (22.109)	−118.150*** (17.053)	−65.183*** (21.534)
Observations	459	459	459	459
R^2	0.385		0.358	
Adjusted R^2	0.367		0.345	
Log likelihood		−271.070		−274.642
Akaike Inf. Crit.		574.140		573.284
Bayesian Inf. Crit.		640.204		622.832
Residual Std. Error	0.458 ($df = 445$)		0.466 ($df = 449$)	
F statistic	21.458*** ($df = 13$)		27.776*** ($df = 9$)	

*$p < .1$.
**$p < .05$.
***$p < .01$.

Required" are associated with greater growth in "DM Share." Furthermore, "DM Share" is consistently higher for countries, who underwent this specific switch ("Switcher-Yes"; $p < .01$). This gives some indication of reversed causality: countries with higher "DM Share" have been more likely to switch to a copyright category with less obligation or researchers to attain specific rights holder consent for DM. Nevertheless, switching has a significant positive effect with this control. Figure 5 illustrates the effect of switching with all controls and 95% confidence intervals. Overall, there is strong evidence that the share of DM research in total research output increases, where researchers do not need to acquire specific consent by rights holders.

6 | DISCUSSION

This paper documents an inverse association between copyright strength and DM uptake: countries in which academic researchers must acquire the express consent of rights holders to conduct lawful DM exhibit a lower share of DM research output in their total research output. That result transpires reasonably consistently across a number of complementary quasi-experiments. This implies that an application of copyright exceptions or limitations that establish the right to mine for academic researchers—if they have lawful access to input works and irrespective of explicit rights holder consent—boosts DM research. In this section, we discuss four potential challenges to the validity of this interpretation of our results.

6.1 | Measurement validity

We employ a plain method to identify relevant articles, and no more definitive measure of the number of DM publications is available for comparison. We further discuss the dependent variable and measurement validity in the Supporting Information. Among other things, there,

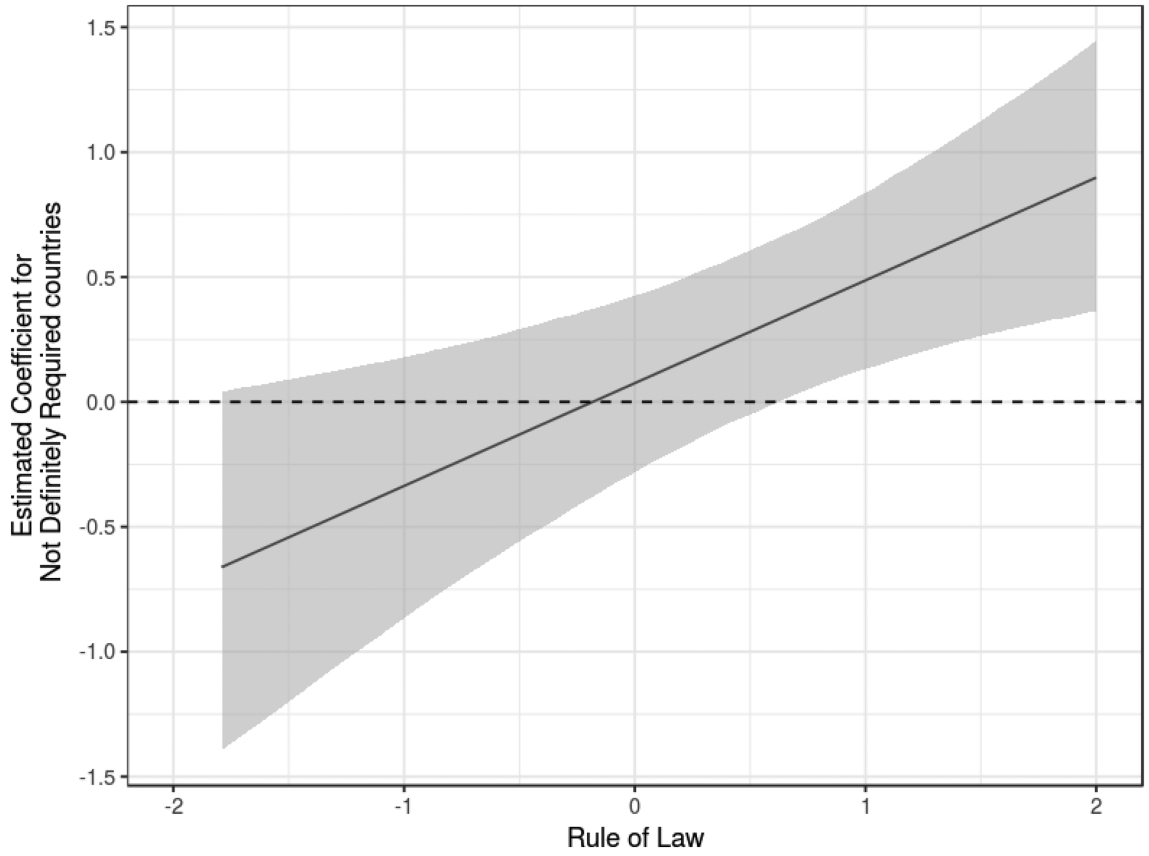

FIGURE 4 Marginal effect of "Rule of Law" on the coefficient for the "Not Definitely Required" copyright category with "DM Share" as dependent variable (including 95% confidence intervals)

TABLE 8 Difference-in-difference (DID) regressions with "DM Share" as dependent variable regarding switches from "Probably Required" to "Probably not Required"

	All countries initially in the "Probably Required" category (except Japan)		All countries (except England and Japan)	
	(1a)	(1b)	(2a)	(2b)
Switch-Yes	1.434*** (0.155)	0.454*** (0.161)	1.235*** (0.105)	0.410*** (0.143)
Switcher-Yes	0.328*** (0.092)	0.400*** (0.116)	0.313*** (0.054)	0.328*** (0.072)
GDP/capita		−0.012 (0.009)		−0.018*** (0.003)
Population (log)		0.025 (0.071)		−0.207*** (0.042)
Rule of Law		0.670*** (0.155)		0.180*** (0.064)
Research output (log)		−0.018 (0.082)		0.179*** (0.043)
Broadband		−0.012* (0.007)		−0.011*** (0.003)
Constant	−0.174 (0.210)	−0.459 (0.883)	−0.056 (0.091)	2.321*** (0.519)
Observations	261	135	785	431
Year dummies	Yes	Yes	Yes	Yes
R^2	0.599	0.579	0.547	0.403
Adjusted R^2	0.558	0.496	0.532	0.371
Residual standard error	0.677 ($df = 236$)	0.469 ($df = 112$)	0.521 ($df = 760$)	0.466 ($df = 408$)
F statistic	14.693*** ($df = 24; 236$)	6.988*** ($df = 22; 112$)	38.187*** ($df = 24; 760$)	12.533*** ($df = 22; 408$)

*$p < .1$.
**$p < .05$.
***$p < .01$.

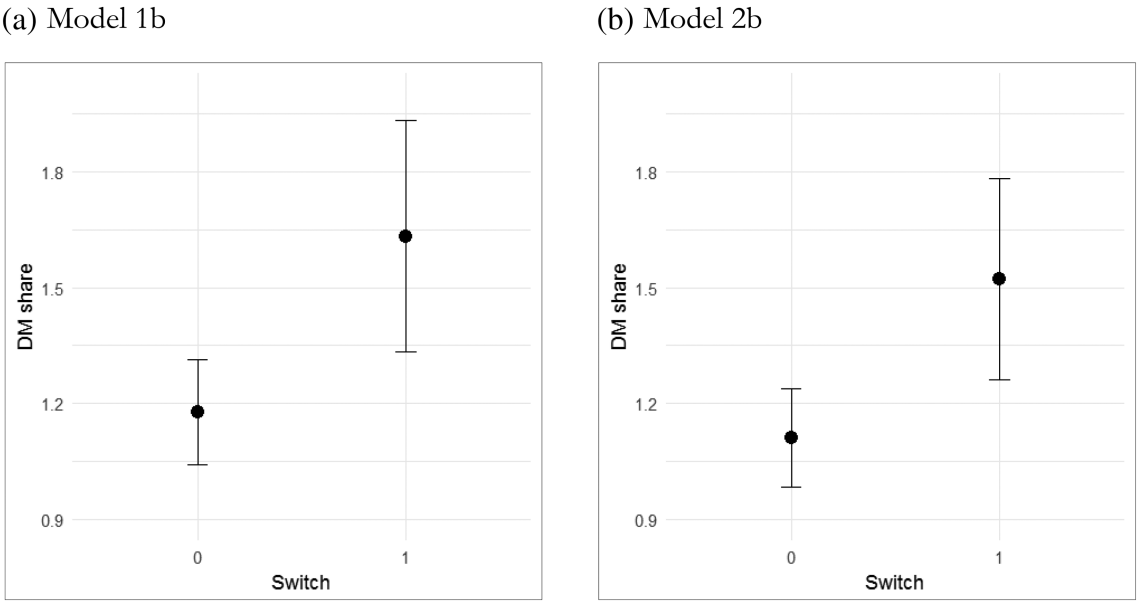

FIGURE 5 Effects of switching according to the difference-in-difference (DID) analyses in Table 8 and with 95% confidence intervals

we document that "data mining" is the most popular and central term used for the research practice addressed in this paper and that there is no indication that other (combinations of) search terms would have improved the validity of our results in terms of measuring the trend in the number of relevant articles per country and copyright category.

We have encountered three specific criticisms regarding measurement validity. First, articles to do with the definitive research methods of DM will not be identified in our data collection if they do not prominently feature the expression "data mining." However, for our regression results to validly reflect any association between DM and copyright, no perfect absolute measure of the number of articles concerned with DM practices is required. For our purpose it is sufficient that omission or exclusion error in our variable "DM Output" is reasonably constant at least across copyright categories, because we do control for constant country differences with "Research Output" and varying country intercepts. To be sure, a formal proof that this holds is not feasible.

Second, researchers in countries with restrictive copyright could have an incentive not to prominently signal that they conducted potentially copyright infringing data collection practices by including the term "data mining" in the title, abstract, or key words. However, we can at least control for any constant propensity of "hiding" by including varying country intercepts.[17]

Third, with our identification method, we can classify many articles at low cost, but we cannot distinguish between applications of DM and conceptual or methodological papers. Copyrights regarding potential input works can directly affect incentives to apply DM applications. Copyright should have less of an effect on researchers' conceptual or methodological work on DM. If so, our data underestimates the effect of copyright on applied DM research. The inverse association between copyright strength and DM output would be more pronounced where only DM applications are concerned. Therefore, this concern constitutes no major challenge to the validity of our results.

6.2 | Endogeneity and omitted variable bias

As discussed in Sections 4 and 5, with our data, no single quasi-experimental design can provide entirely conclusive results, and our strategy is to construct several, complementary quasi-experiments to mitigate that challenge. Our main results are clearly significant where we use relatively refined copyright categorizations (Table 5) and where we check for the consequences of relevant changes in copyright law (switches), including controls for self-selection and simultaneity (Table 8). Our results are weaker ($p < .1$ with all controls) where we exclude self-selection and simultaneity by classifying countries into a binary and virtually permanent distinction into "Consent Required" and "Not Definitely Required" countries according to relevant copyright law (Table 6). However, where we include the interaction between "Rule of Law" and copyright, we attain the reasonably clear result that the combination of the "Consent Required" copyright category with high "Rule of Law" scores—as observed in most EU Member States—leads to lower DM activity by academic researchers (Table 7).

With country panels, omitted variable bias is hard to exclude. There are already challenges in the application of economic theory to specify determinants of research output and the adoption of new research technologies, because incentives for publishing academic articles are not shaped in conventional markets. What is more, many available indicators do not perfectly correspond to one specific theoretical determinant. Nevertheless, we do have good controls for the most outstanding factors determining DM uptake by academic researchers. With our dependent variable "DM Share" (the quotient

of the number of DM articles and the total number of articles published by authors from a country), we do not only have an effective control for the resources available for domestic research and the productivity of domestic researchers (in terms of articles produced relative to research resources).[18] With this derived dependent variable and varying country intercepts, we also have some control for constant, unobserved country differences in incentives for DM uptake, for instance, due to different compositions of research activities within countries that could affect the efficient scale and scope of DM.[19] Furthermore, broadband penetration should be correlated with the costs and quality of ICT and the propensity and skills of residents to use digital ICT. Competition between researchers may be positively correlated with our control variable "Population" assuming that there is a disutility of researchers to relocate to another country. The availability of relevant input works irrespective of copyright should also be positively correlated with country size, assuming that researchers are more likely to use data on domestic phenomena.

No satisfactory indicators are available on some potential determinants of "DM Share," for instance the costs of tradable DM inputs such as specialized ICT hardware and software and to some extent even labor. For these, we can only control for constant country differences, which is more effective for determinants that change slowly over time within countries. Our panel mostly consists of large, diversified economies, which makes substantial and sudden changes within reasonably populated copyright categories less probable. Furthermore, extensive integration of many of the economies studied here make it improbable that prices of inputs would diverge very substantially over time between the treatment group(s) and the control group.

Other specific determinants for which we have no controls are the following. First, changes in the share of various academic disciplines in countries' research activities or academic cultures to do with technological innovation could affect "DM Share" but are unlikely to trend rapidly over time. Second, tastes and preferences of researchers regarding innovative research methods could be somewhat controlled for by data on the demographics of academic researchers. To the best of our knowledge, there is no such data available for a sufficient number of countries. Third, there are no data available on how DM would affect working conditions of researchers, except for the effect of greater productivity on career prospects and legal risks associated with copyright captured by our main predictor. Perhaps the most worrying omission is that no suitable data are available on targeted funding of academic DM within specific territories.

There is potential missing data bias, as data on all controls are only available for 55% of yearly observations from countries that could be classified according to their DM-related copyright law. However, due to the high degree of statistical significance and power of our main results, the probability is high that the main results hold in spite of any remaining omitted variable bias. Although the coefficients of determination in our regressions are in a respectable range, it is noteworthy that these are deflated, because we incorporate our main control variable "Research Output" in our dependent variable "DM Share."

6.3 | Cross-country effects

According to the literature on patents, even where there is no positive effect of IP on domestic innovation, IP may still increase "technological transfer"—the influx of new ideas from other countries (Branstetter et al., 2006; Hall & Harhoff, 2012; Helpman, 1993; Jarvorcik, 2004). However, pure information goods suitable for DM are less excludable than patents and the underlying technologies. Then, strong domestic copyright protection may inhibit transfer and use of input works into countries, whereas valuable data will be accessible in territories with less copyright protection. High protection countries may get the worst of both worlds: extensive unauthorized use of domestically produced data abroad and high costs of conducting DM domestically.

In talks with DM practitioners, we were even told that it is common practice to deliberately locate DM activities in territories with weak de facto copyright protection and to seek out suitable partners from such territories in international DM cooperations. Therefore, it is not clear whether a strong DM performance of some countries is self-sufficient or whether it is due to strategic decisions by researchers and/or free riding on data produced in other territories. To investigate this further requires a content analysis of DM-related research output. In particular, future research should establish to what extent input works for DM research come from countries with more or less restrictive copyright regulation. This is beyond the scope of this paper.

6.4 | Generalizability

No database on research output is comprehensive in the sense that it would cover all valuable research output. Greater coverage is not even necessarily better. Publications in top journals are typically regarded as many times more valuable than publications in lower ranking journals.[20] There is no widely accepted metric of value that would allow for valid weighting of publications across all disciplines.[21] WoS is the most selective of the major research databases and provides the standard assessment of impact factors of journals (the Science Citation Index).[22] We rely on their inclusion criteria to cover a reasonably stable share of the most valuable total research output and DM research output per copyright category.

For the purpose of measuring countries' research output, WoS has no superior alternative (Burnham, 2006). Scopus is the only reputable alternative database covering virtually all academic disciplines (Falagas et al., 2008).[23] Scopus initially did not fully cover publications prior to 1996, however (Archambault et al., 2009). For publications after 1996, the coverage of Scopus and WoS overlap to a great extent, and Archambault et al. (2009) document that between 1996 and 2007, Scopus produces virtually the same country ranks and very similar absolute publication counts at the aggregate and for a number of different time periods or academic disciplines. Nevertheless, as WoS and Scopus are continuously adjusted, future replication of our study using Scopus or other databases may still be useful.

There is a sizable literature on the extent of bias in WoS (as well as Scopus) due to uneven coverage across countries, languages, and disciplines. However, it is common practice for prestigious academic journals publishing in any language to include English abstracts. Thus, language bias should be weaker in our data than if we had assessed full articles with search terms in English. Furthermore, the evidence is that any bias in WoS has been rather stable over time for any period systematically investigated; see Mongeon and Paul-Hus (2016) for a recent summary of the literature. Therefore, varying country intercepts should provide a reasonable control for the combined effect of biased coverage in WoS and stable country characteristics.

Our results may be less valid regarding the humanities and some social sciences. On the one hand, in these disciplines, book publications often have a relatively greater weight in determining individual researchers' career prospects, and these are not covered in our data. On the other hand, journals in these disciplines are relatively less likely to be included by WoS (or Scopus) (Mongeon & Paul-Hus, 2016). Another aspect of this is that qualitative empirical research is not covered well by our data. The data in this article covers (quantitative) DM and not qualitative methods of text mining.

7 | CONCLUSIONS

DM is the topic of an increasing number of academic journal articles. Copyright protection of data in EU Member States is relatively strong. We document that this is associated with less DM research output by academic researchers.

DM research often draws on many input works to which others hold copyrights. In virtually all EU Member States, as well as a couple of other countries, there are no relevant exceptions or limitations to copyright, so that DM requires express consent of rights holders. With this regulation, academic DM research has fallen behind developments in other territories. The benefits of allowing DM for all users, who have lawful access to data, seem to be greater than any adverse effects of weaker copyright protection on the creation of new input works for DM. Our results suggest that there has been market failure regarding the licensing of data for academic DM. Copyright does not appear to attain its ostensible goal of fostering innovation in this particular context. To our knowledge, this study is the first to empirically document an adverse net effect of IP on innovation, in the sense that there is strong evidence for stricter copyright hindering the wide adoption of novel ways to build on copyright works and generate derivative works.

As new technologies mature, early leadership can give rise to stable advantages, so that the stakes during formative years are high. For as long as DM continues to offer productivity increases in academic research, researchers in the EU and other territories with similar copyright law risk become less competitive because of greater copyright restrictions for this novel type of research.

Our results do provide a better evidence base for policy than has been available so far. The results of several, complementary quasi-experiments presented in this paper are reasonably consistent. Nevertheless, there is clearly scope for further research. For instance, the identification of DM output in this paper is efficient but plain. Furthermore, cross-country effects require further attention, as data produced in one territory may be analyzed elsewhere.

ACKNOWLEDGMENTS

No external funding was used in the preparation of this paper. The authors are grateful to Merel Goedknegt and Saskia Woutersen-Windhouwer for effective research support. This paper also benefitted from comments by the participants of the research seminar on Cultural Economics at ESHCC, Erasmus University Rotterdam, the Ligue des Bibliothèques Européennes de Recherche (LIBER) Conference 2015, the International Conference on Electronic Publishing (ELPub) Conference 2015, and the European Policy for Intellectual Property (EPIP) Conference 2015. Earlier versions of this paper with less comprehensive data analysis received the Finalist Best Paper Award at the EPIP Conference 2015 and the LIBER Innovation Award 2015. This preliminary version is available on SSRN and has been published in the LIBER Conference 2015 proceedings.

CONFLICT OF INTEREST

None of the authors have conflicts of interest.

ORCID

Christian Handke https://orcid.org/0000-0002-3976-0272
Lucie Guibault https://orcid.org/0000-0002-9530-9312
Joan-Josep Vallbé https://orcid.org/0000-0001-8327-7774

ENDNOTES

[1] In economics and econometrics, DM traditionally has another, negative connotation regarding a type of malpractice in applied statistical data analysis (e.g., Sullivan et al., 2001; Feelders, 2002; Rockey & Temple, 2016). This is not the common use of the term and not what this paper is about. In a random sample of 250 DM-related articles from our corpus, none used the term exclusively in this sense; see the Supporting Information, Section C.1.1. Furthermore, DM is typically associated with data analysis of structured, quantitative, or nominal data, rather than text mining that concerns processing of unstructured/qualitative data and may be a preparatory step for DM in text and data mining (TDM) research projects.

[2] Google was launched in 1997 and Facebook in 2004.

[3] United States Court of Appeals for the Second Circuit, 13-4829-cv (Google Books vs. Authors' Guild) (16.10.2015)
United States Court of Appeals for the Second Circuit, June 10, 2014 (*Authors' Guild of America* vs. *Hathitrust*), No. 12-4547-cv., 755 F.3d 87, 91 (2d Cir. 2014).

[4] Popular definitions of academic DM explicitly state that data mining concerns the use of secondary data, collected by others than the researchers (Hand et al., 2001), and the combination and joint analysis of separately assembled data sets is a main aspect of DM.

[5] All data and documentation of our data analysis are available on a GitHub depository at: https://github.com/pepvallbe/Copyrights-Impact-on-Data-Mining-in-Academic-Research.

[6] Filippov (2014) and Filippov and Hofheinz (2016) contain descriptive analyses of similar data compiled on the databases Google Scholar and ScienceDirect.

[7] For a general discussion of law enforcement and social norms, see Acemoglu and Jackson (2017).

[8] The WGI project relies on perceptions-based data from surveys of firms and households, complemented by "expert assessments produced by various [other] organizations" (Kaufmann et al., 2010, p. 18; see page 29 for an overview). This is justified, among other things, as "perceptions matter because agents base their actions on their perceptions, impression, and views" (Kaufmann et al., 2010, p. 18). Over the years, the WGI project has implemented and published the results of various tests, which so far have not revealed any major validity problems (Kaufmann et al., 2010, p. 19ff.).

[9] What is more, the number of units of analysis (countries) under investigation is modest, as there are few relevant observations in smaller economies.

[10] There is further missing data on "Broadband" and "GDP/capita," see Table 1.

[11] The addition of "Research Output" in the model could raise multicollinearity issues with "GDP/capita" and "Population." However, the correlation between these variables and total research output is low in our data (.37 and .31, respectively) so that collinearity is unlikely.

[12] Simultaneity or reverse causation could bias results if an observed or expected relatively strong performance regarding domestic DM research (or lobbying by stakeholders) would have affected changes in national copyright legislation over the period studied.

[13] For instance, the coefficient of the "Probably not Required" constitutive term (0.865) represents the effect of this type of copyright regulation only when rule of law is zero (about the level of Turkey and India in many years). In our panel, there are only four observations in the data with rule of law between −0.01 and 0.01 (there are no exact zero matches), which are South Africa in 1996, Argentina in 1997, and Brazil in 2010 and 2011.

[14] There are six countries that switched from "Probably Required" to "Probably not Required." Due to missing data on "Rule of Law," Taiwan is not included in models with control variables.

[15] Japan underwent a different switch from "Probably Required" to "Not Required." We exclude Japan in the analysis. The sign and significance levels of the "Switch-Yes" and "Switcher-Yes" dummies are the same with Japan included.

[16] The sign and significance levels of the "Switch-Yes" and "Switcher-Yes" dummies are the same with Japan and England included.

[17] "Hiding" is costly for researchers where it affects the quantity or quality of attention their publications receive. What is more, avoiding the term "data mining" is a limited means to mitigate legal risks.

[18] This control should be superior to any measure of overall investments in academic research, which in any case is not available for a sufficient number of countries. The widely available measure of R&D expenditure includes industrial research and may only be weakly correlated with expenditure on academic research.

[19] Mongeon and Paul-Hus (2016) document that compared with the most comprehensive list of scholarly/academic periodicals (Ulrich's periodical directory with over 60,000 journals), WoS and Scopus cover a greater share of the journals on natural sciences, engineering, and biomedical research than of those on the social sciences and humanities. It is unclear whether this is justified by consistent and adequate quality criteria.

[20] WoS only covers journals that have continuously operated for several years, employ effective peer reviewing, and have been reasonably frequently cited in other high-quality research publications. By using WoS, we thus focus on those publications that have been deemed by authors, editors, and reviewers to be original and valuable enough to merit publication in reasonably prestigious periodicals. We regard that to be an advantage over alternative databases with lower quality thresholds.

[21] Citation counts vary widely across disciplines and entail the problem that many very prestigious journals are only read/comprehensible for a small number of experts in a specific research area.

[22] Researchers in many academic disciplines have strong incentives to publish their highest quality works in journals covered by WoS and preferably in the most reputable journals, with reputation hinging to a large extent on this impact factor.

[23] As discussed in the compendium, Google Scholar was not suitable as it produced invalid results.

REFERENCES

Acemoglu, D., & Jackson, M. O. (2017). Social norms and the enforcement of law. *Journal of the European Economic Association*, *15*(2), 245–295.

Archambault, É., Campbell, D., Gingras, Y., & Larivière, V. (2009). Comparing bibliometric statistics obtained from the Web of Science and Scopus. *Journal of the American Society for Information Science and Technology*, *60*(7), 1320–1326.

Arrow, K. J. (1962). Economic welfare and the allocation of resources for invention. In National Bureau of Economic Research. (Ed.), *The rate and direction of inventive activity* (pp. 609–625). Princeton University Press.

Branstetter, L. G., Fisman, R., & Foley, C. F. (2006). Do stronger intellectual property rights increase international technology transfer? Empirical evidence from U.S. firm-level panel data. *The Quarterly Journal of Economics*, *121*(1), 321–349.

Burnham, J. F. (2006). Scopus database: A review. *Biomedical Digital Libraries*, *3*(1), 1.

Cattaneo, G., Micheletti, G., Glennon, M., La Croce, C. & Mitta, C. 2020. The European Data Market Monitoring Tool. Report of the European Commission (DG Communications Networks, Content and Technology). Luxembourg: Publications Office of the European Union.

Dasgupta, P., & David, P. A. (1994). Toward a new economics of science. *Research Policy*, *23*(5), 487–521.

Dean, J. (2014). *Big data, data mining, and machine learning: Value creation for business leaders and practitioners*. John Wiley & Sons.

Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society, OJ L 167, 22.6.2001, 2001. p. 10–19.

Eicher, T., & Garcia-Penalosa, C. (2007). Endogenous strength of intellectual property rights: Implications for economic development and growth. *European Economic Review*, *52*, 237–258.

Einav, L., & Levin, J. (2014). Economics in the age of big data. *Science*, *346* (6210), 1243089.

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., & Pappas, G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: Strengths and weaknesses. *The FASEB Journal*, *22*(2), 338–342.

Feelders, A. (2002). Data mining in economic science. In J. Meij (Ed.), *Dealing with the data flood: Mining data, text and multimedia* (Vol. 65) (pp. 166–175). International Specialized Book Service.

Filippov, S. 2014. Mapping text and data mining in academic and research communities in Europe. Brussels: Lisbon Council. Online: https://lisboncouncil.net/publication/publication/109-mapping-text-and-data-mining-in-academic-and-research-communities-in-europe.html

Filippov, S., & Hofheinz, P. 2016. Text and data mining for research and innovation—What Europe must do next. Brussels: Lisbon Council. Online: file:///C:/Users/handk/AppData/Local/Temp/LISBON_COUNCIL_Text_and_Data_Mining_for_Research_and_Innovation.pdf

Gandel, S. 2016. These are the most valuable companies in the Fortune 500. *Fortune*, February 4. Online: http://fortune.com/2016/02/04/most-valuable-companies-fortune-500-apple/

Ginarte, J. C., & Park, W. G. (1997). Determinants of patent rights: A cross-national study. *Research Policy*, *26*(13), 283–301.
Hall, B. H., & Harhoff, D. (2012). Recent research on the economics of patents. *Annual Review of Economics*, *4*, 541–565.
Hand, D. J., Mannilla, H., & Smyth, P. (2001). *Principles of data mining*. MIT Press.
Helpman, E. (1993). Innovation, imitation, and intellectual property rights. *Econometrica*, *61*(6), 1247–1280.
Jarvorcik, B. S. (2004). The composition of foreign direct investment and protection of intellectual property: Evidence from transition economies. *European Economic Review*, *48*(1), 39–62.
Kaufmann, D., Kraay, A. & Mastruzzi, M. (2010). The Worldwide Governance Indicators Project. Technical report, World Bank, Washington DC.
Kaufmann, D., Kraay, A. & Zoido, P. (1999). Governance matters. World Bank Policy Research Working Paper No. 2196, World Bank, Washington DC.
Landes, W. M., & Posner, R. A. (1989). An economic analysis of copyright law. *The Journal of Legal Studies*, *18*(2), 325–363.
Lorenczik, C., & Newiak, M. (2012). Imitation and innovation driven development under imperfect intellectual property rights. *European Economic Review*, *56*, 1361–1375.
Meyer, B. D. (1995). Natural and quasi-experiments in economics. *Journal of Business & Economic Statistics*, *13*(2), 151–161.
Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: A comparative analysis. *Scientometrics*, *106*(1), 213–228.
Novos, I. E., & Waldman, M. (1984). The effects of increased copyright protection: An analytic approach. *Journal of Political Economy*, *92*(2), 236–246.
OECD. (2014). *Measuring the digital economy: A new perspective*. OECD Publishing.
OECD. (2015). *Data-driven innovation: Big data for growth and well-being*. OECD Publishing.
Rockey, J., & Temple, J. (2016). Growth econometrics for agnostics and true believers. *European Economic Review*, *81*, 86–102.
Samuelson, P. A. (1954). The pure theory of public expenditure. *Review of Economics and Statistics*, *36*(4), 387–389.
Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Wadsworth Cengage Learning.
Sullivan, R., Timmerman, A., & White, H. (2001). Dangers of data mining: The case of calendar effects in stock returns. *Journal of Econometrics*, *105*(1), 249–286.
Van Raan, A. F., Van Leeuwen, T. N., & Visser, M. S. (2011). Severe language effect in university rankings: Particularly Germany and France are wronged in citation-based rankings. *Scientometrics*, *88*(2), 495–498.
Varian, H. R. (2014). Big data: New tricks for econometrics. *The Journal of Economic Perspectives*, *28*(2), 3–27.
World Bank. 2015a. World development indicators. World Bank, Washington DC (data). Available at: http://data.worldbank.org/data-catalog/world-development-indicators
World Bank. 2015b. Worldwide Governance Indicators (WGI) Project. World Bank, Washington DC (data). Available at: http://info.worldbank.org/governance/wgi/index.aspx#home

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: 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

APPENDIX A: AN OVERVIEW OF COUNTRY-LEVEL "DM Share" DATA, 1992–2014

For all 42 countries and the entire time period covered (1992 and 2014), the average "DM Share" score is 0.77‰. In the dozen years between 2003 and 2014, this score was at 1.04‰.

Table A1 reports indicators for countries that remained within the same copyright category throughout. Compared with the global averages, the 15 largest EU Member States exhibit below average "DM Share" (0.66‰). Within the EU, Greece (1.42), Portugal (1.15), and Spain (1.02) have relatively high DM shares, whereas the values for the most populous countries Germany (0.53) and France (0.50) are relatively low. Other countries from the "Consent Required" category exhibit even lower average "DM Share" than EU Member States (0.57). The average scores of countries from the "Probably Required" category (0.84) and for the United States ("Probably not Required"; 0.81) are about average.

Tables A2 reports indicators for countries that switched between copyright categories. Some large Asian economies are included here and exhibit relatively high "DM Share" throughout, in particular Taiwan (3.36), Singapore (1.95), and China (1.17) but not Japan (0.52). (Especially Taiwan exhibits extremely high scores for "DM Share" over later years. Due to missing data for this country on the variables "Rule of Law" and "Broadband," Taiwan is not included in any model with control variables.) It is a common assertion that developing countries have an interest in lower levels of intellectual property (IP) protection (Lorenczik & Newiak, 2012) and that this affects de jure or de facto variations between countries in IP protection (Eicher & Garcia-Penalosa, 2007). However, according to our data, highly developed countries also exhibit greater adoption of novel research methods with lower levels of copyright protection. The majority but not all countries that did switch had high "DM Share" compared with countries from the same initial copyright category or the global average. (All switches were into copyright categories with fewer obligations for academic researchers to clear rights specifically for data mining, and six out of eight switches occurred from "Probably Required" to "Probably not Required.") This provides some indication that self-selection or simultaneity is a concern in particular when comparing the "Probably Required" category to the "Probably not Required" category.

Finally, the five countries that could not be classified according to our copyright categories also exhibit low "DM Share" scores on average (0.44), albeit with considerable variance.

TABLE A1 Nonswitching countries

Initial copyright category[a]	Country	First DM article published in …	Number of DM articles; 1992–2014	Average "DM Share"; 1992–2014	Average "DM Share"; 2003–2014
Consent Required					
EU[b]	Germany	1996	863	0.53	0.74
	France	1996	591	0.50	0.74
	Italy	1997	605	0.70	0.96
	Spain	1998	696	1.02	1.38
	Netherlands	1995	256	0.49	0.66
	Sweden[c]	1997	135	0.36	0.49
	Poland[d]	1998	259	0.85	1.06
	Belgium	1997	271	0.97	1.30
	Denmark	2000	85	0.53	0.57
	Austria[c]	1999	112	0.71	0.73
	Finland[c]	1994	159	0.8	1.12
	Greece	2000	215	1.42	1.87
	Ireland	1997	140	1.11	1.29
	Portugal	1998	137	1.15	1.42
	Subtotal EU		4524	0.66	0.92
Other	Brazil	1999	281	0.68	0.80
	Mexico	1997	92	0.49	0.64
	Turkey	1998	243	0.87	1.01
	Switzerland	1997	159	0.43	0.59
	Norway[e]	1997	46	0.30	0.42
	Argentina	1999	35	0.30	0.42
	Colombia	1999	24	0.80	0.88
	Venezuela	2002	11	0.49	0.74
	Subtotal Other		891	0.57	0.72
Total			5415	0.65	0.88
Probably Required	Australia	1995	680	1.07	1.41
	India	1999	386	0.63	0.83
	Malaysia	2001	92	1.42	1.55
	Nigeria	2004	4	0.14	0.19
	South Africa	1997	43	0.35	0.48
	Thailand	2000	72	1.17	1.34
Total			1279	0.84	1.08
Probably not Required	USA	1992	4827	0.81	1.09

Note: Sources: Own calculations based on data collected by authors on the Web of Science database.

[a]First classifiable year from 1992.

[b]Only the 15 largest EU Member States included by GDP in 2013 (World Bank, 2015a, 2015b); England is presented separately as a switching territory.

[c]EU since 1996.

[d]Not classifiable before 1996; EU since 2004.

[e]European Economic Area (EEA) Members.

TABLE A2 Switching countries

Initial copyright category[a]	Destination copyright category (arrival)	Country	First DM article published in …	Total number of DM articles; 1992–2014	Average "DM Share"; 1992–2014 (5 years before switch)	Average DM share; 2003–2014 (after switch)	Deviation from countries in same initial copyright category; average of 5 years preceding switch	Deviation from average of all countries last year before switch
Consent Required	Not Required (2014)	England[b]	1994	913	0.62 (0.97)	0.89 (0.89)	51.3%	−5.3%
Probably Required	Probably not Required (2012)	Canada	1992	782	0.82 (0.95)	1.13 (1.14)	−7.0%	−12.9%
	Probably not Required (2012)	China[c]	1997	2063	1.17 (1.12)	1.29 (1.13)	21.4%	2.4%
	Probably not Required (2008)	Israel	1997	216	0.84 (0.90)	1.20 (1.39)	−48.5%	3.6%
	Probably not Required (2011)	South Korea	1997	583	1.10 (1.14)	1.24 (1.15)	4.2%	19.7%
	Probably not Required (2005)	Singapore	1996	246	1.95 (2.04)	2.18 (2.16)	290.0%	96.5%
	Probably not Required (2003)	Taiwan[d]	1997	1160	3.36 (0.89)	4.30 (4.30)	19.1%	32.0%
	Total			5050	1.27 (n.a.)	1.53 (n.a.)	n.a.	n.a.
Probably Required	Not Required (2010)	Japan	1996	585	0.37 (0.52)	0.52 (0.55)	−72.0%	−51.6%

Note: Sources: Own calculations based on data collected by authors on the Web of Science database.

[a]First classifiable year from 1992.

[b]Web of Science reports on England, Scotland, Wales, and Northern Ireland separately; we only report the figures for England.

[c]Unclassifiable before 2007.

[d]Taiwan is not recently listed by the World Bank, not as Republic of China, either; we collected the information on Taiwan's GDP from the IMF World Economic Outlook 2015; Taiwan provides an interesting case because it was early in switching to "Probably not Required" and exhibits high DM shares.

TABLE A3 Unclassifiable according to copyright categories (excluded from econometric analysis)

Country	First DM article published in …	Total number of DM articles; 1992–2014	Average DM share; 1992–2014	Average DM share; 2003–2014
Russia	1997	83	0.14	0.20
Indonesia	2004	7	0.46	0.53
Saudi Arabia	2003	48	0.77	0.98
United Arab Emirates	2007	3	0.71	1.09
Iran	2005	233	1.31	1.39
Total		374	0.44	0.66

Note: Sources: Own calculations based on data collected by authors on the Web of Science database.