IS EUROPE FALLING BEHIND IN DATA MINING?

COPYRIGHT'S IMPACT ON DATA MINING IN ACADEMIC RESEARCH

5th July 2016

Christian HANDKE[a], Lucie GUIBAULT[b] and Joan-Josep VALLBÉ[c]

[a]ESHCC, Erasmus University Rotterdam, Rotterdam and IViR, University of Amsterdam, The Netherlands; handke@eshcc.eur.nl

[b]IViR, University of Amsterdam, The Netherlands; L.Guibault@uva.nl

[c]GREL, University of Barcelona, Spain; vallbe@ub.edu

Abstract. This paper discusses how copyright affects data mining (DM) by academic researchers. Based on bibliometric data, we show that where DM for academic research requires the express consent of rights holders: (1) DM makes up a significantly lower share of total research output; and (2) stronger rule-of-law is associated with less DM research. Copyright exceptions or limitations seem effective in promoting DM research.

Keywords. Data mining; copyright, academic research; innovation.

IS EUROPE FALLING BEHIND IN DATA MINING?

COPYRIGHT'S IMPACT ON DATA MINING IN ACADEMIC RESEARCH

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 produce valuable new insights.[2] As we will show, DM has been the topic of an increasing share of total academic research output over the last two decades.

Effective copyright protection should increase the supply of potential DM input works and the costs of using them. We analyse bibliometric data to establish how various copyright policies affect the application of DM in academic research. We show that countries in which data mining for academic research requires the express consent of rights holders, data mining makes 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 European Union (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.[3] The European Commission is currently considering copyright reforms to allow for data mining without express consent of the rights holder, so that the right to read would be the right to mine.[4] The USA have generally adopted a more permissive copyright policy regarding DM and recent rulings seem to confirm great scope for DM without express consent by rights

---

1 Hand, D.J., Mannilla, H., Smyth, P., 2001. Principles of Data Mining. MIT Press, Cambridge.

2 OECD, 2014, Measuring the Digital Economy: A New Perspective. OECD Publishing, Paris. OECD, 2015, Data-Driven Innovation: Big Data for Growth and Well-Being. OECD Publishing, Paris.

3 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, p. 10–19.
Directive 1996/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases OJ L 77, 27.3.1996, p. 20–28.

4 Murray-Rust, P., 2012. Open content mining. Working Paper, Cambridge University. Online: http://www.dspace.cam.ac.uk/handle/1810/243749

holders.[5] 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.

Theory

Copyright is often addressed as a means to mitigate market failure in the private provision of goods with public good attributes.[6] 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, as it increases returns to rights holders and the costs of building on protected works.[7] From a welfare economic perspective, copyright thus fights fire with fire: it mitigates one source of market failure (under-provision of public goods) with another (market power and underutilization of public goods).

Our empirical work is based on several related assumptions. First, data mining is often conducted by researchers, who are not the rights holders of all adequate data.[8] Second, data mining by academic researchers increases in the quantity and quality of supply of suitable data. Third, data mining decreases in the costs of accessing relevant data. Fourth, effective copyright protection affects the supply of suitable data and/or the costs of ensuring access. We thus hypothesize that variations in relevant copyright policy between countries will affect the amount of data mining by researchers residing in those countries. Since copyright has ambivalent effects on follow-up use of protected works, the direction of copyright's effect on data mining is unclear at the outset.

Empirical strategy

Main factors driving the output of academic publications of any type should be (1) the means

---

[5] 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).

[6] Samuelson, P.A., 1954. The pure theory of public expenditure. Review of Economics and Statistics 36 (4); 387-389. 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. Princeton University Press, Princeton, 609-625. Novos, I.E., Waldman, M., 1984. The effects of increased copyright protection: An analytic approach. Journal of Political Economy 92 (2), 236-246.

[7] Landes, W.M., Posner, R.A., 1989. An economic analysis of copyright law. Journal of Legal Studies 18 (2), 325-363.

[8] Popular definitions of data mining 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 data mining.

available for academic research, in particular labour and capital, and (2) the productivity of researchers, as measured by the number and quality of research output relatively to the resources used. To control for the size and productivity of academic research, we use the ratio between DM-related research output and total research output per country as the dependent variable.[9]

The main independent variables of interest derive from a categorization of countries according to relevant copyright law and practice in each jurisdiction. Furthermore, there is often a gap between the provisions of IP law and social practice, since IP is hard to enforce. In our analysis, we thus also consider an indicator of the rule of law within countries.

Data

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.

We extracted the number of all published research articles on DM from 42 large economies. 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 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; (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.

The articles featured in the search results contain DM applications as well as related conceptual and methodological work. Among the countries covered, searches on WoS brought up 18,441 DM-related articles between 1993 and 2014.

[9] No measure of overall investments in academic research is 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. Neither are there data on the number of academic researchers.

We also collected data on the total number of research articles published for the same set of countries and years. 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, so that 0.7‰ of all articles had DM as a topic. In our empirical analysis, for each country and year we used the ratio of the number of DM articles and the total number of articles as the dependent variable, multiplied by 1,000 to avoid dealing with very small fractional numbers. This variable is referred to as 'DM share' below.

The absolute number of DM articles and their share in total research output has increased substantially since 1992. See Figures 1 and 2 for an illustration and Table 1 for an overview of the data. The large Asian economies exhibit relatively high DM shares, in particular Taiwan, Singapore, China and Malaysia but not Japan. On average and without weighting for country size, the 15 largest EU Member States exhibit a very similar DM share as the USA over recent years. Within the EU, Greece, Portugal and Spain have a relatively high DM share. The values for Germany, France and England are lower, also compared to the USA or major Asian economies (except for Japan).

FIGURE 1

FIGURE 2

TABLE 1

We classify countries according to the type of copyright protection that applies to DM. 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; (2) whether there is relevant case law specifying the applicability of existing exceptions and limitations. Table 2 gives an overview of country categorizations from 1992 to 2014.

TABLE 2

We also incorporate rule of law as reported by the Worldwide Governance Indicators Project (WGI).[10] 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

[10] Kaufmann, D., Kraay, A., and Zoido, P. (1999). Governance matters. World Bank Policy Research Working Paper no. 2196, World Bank, Washington DC.
Kaufmann, D., Kraay, A., and Mastruzzi, M. (2010). The Worldwide Governance Indicators Project. Technical report, World Bank, Washington DC.
World Bank, 2015. Worldwide Governance Indicators (WGI) Project. World Bank, Washington DC (data). Available at: http://info.worldbank.org/governance/wgi/index.aspx#home

copyright.[11] We also use several control variables such as the GDP per capita and country population size. All data is documented in a compendium.

Data analysis and results

Due to the panel structure of our data and the low temporal variation of copyright legal arrangements within countries, we fit a multilevel linear regression model with varying intercepts by country. The dependent variable is the share of articles on DM in the total number of articles published (DM share) per country and year. The main predictor is each country's copyright category, with 'not allowed' as the reference category. Table 4 presents the results of five different specifications of the model.

TABLE 4

Model 1 only contains the main predictor. As expected, the 'allowed' category does not yield significant coefficients: it contains only six observations and we only report it for completeness in Table 4.[12] There is a significant positive coefficient for the category 'probably allowed', which suggests that a more permissive copyright framework is associated with more DM research.

In Model 2 we control for GDP per capita, population size and the rule of law. Model 3 includes a dummy for EU membership. Model 4 also controls for the total number of research articles to test whether changes in DM share are confounded by changes in total research output. Model 5 includes interaction terms to identify the effect of rule of law for specific copyright categories. The number of observations is reduced in models with control variables, since no data are available on the 'rule of law' before 1996.

In all specifications, we find significant positive coefficients for the category 'probably allowed' ($p<.01$). For the category 'probably not allowed', results are less stable. Coefficients for the category 'allowed' are generally positive but not significant with only six observations. Coefficients for 'probably allowed' are consistently larger than for 'probably not allowed'. This is 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 'not allowed'. Overall, there is

[11] Data availability for this indicator begins in 1996 and last estimates are from 2013, but until 2003 we only have estimates for alternate years. To avoid an excessive 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. This indicator is normalized to have a mean around zero and standard deviation around 1.

[12] Furthermore, the effects of introducing permissive copyright regulations on DM share should be gradual, so that the full effect of recent changes in Japan and the UK may not have transpired.

extensive evidence that DM share is lower in countries in the 'not allowed' category - including the EU member states - than in countries with more permissive DM-related copyright.

Comparing Models 2 and 3, EU membership has virtually no effect on DM share nor on the coefficients for other variables. Apparently results hold throughout the 'not allowed' category.

In Models 4 and 5, 'total research output' has a positive and significant coefficient. Countries with a high share of data mining 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. With the control for total research output in Model 4, the coefficient for 'probably not allowed' is no longer significant, whereas the coefficient for 'probably allowed' remains positive and highly significant.[13]

In models 2 to 4, rule of law has a weakly significant, negative coefficient for the entire panel. The rule of law should make the enforcement of copyright law more effective, and thus have a stronger negative effect on DM share in countries with stronger copyright. To test for this, Model 5 includes a multiplicative interaction between copyright categories and the rule of law indicator. In this model, the coefficients of the variables that constitute the interaction (the categories of copyright regulation and rule of law in Model 5) are no longer to be interpreted as unconditional marginal effects.[14] The coefficients for the multiplicative interaction terms (copyright*rule of law) for 'probably allowed' and 'probably not allowed' are positive and significant. By contrast, the coefficient for the rule of law variable corresponding to the reference category 'not allowed' is negative and more pronounced than for the entire panel (-0.418; $p<.01$). In particular the combination of strong copyright law and strong enforcement and/or a cultural propensity to adhere to legal norms appears to reduce academic researchers' data mining performance.

Discussion

In most EU/EEA Member States, DM-related copyright protection is comparatively strong. Our results suggest that the net effect is a weaker performance of domestic academic researchers in this increasingly important type of research.

[13] The addition of 'total research output' in the model could raise multicollinearity issues with GDP per capita and population size. However, the correlation between these variables and total research output is low in our data (0.37 and 0.31 respectively) so that collinearity is unlikely.

[14] For instance, the coefficient of the 'probably allowed' constitutive term (0.880) represents the effect of this type of copyright regulation only when rule of law is zero. 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.

As is usually the case with country-level panel studies, selection bias may restrict the validity of our analysis: the status of countries' copyright arrangements may not be uncorrelated with the expected number of future DM publications. Since there is no verifiable random assignment of countries to copyright categories, our results cannot be entirely conclusive regarding the direction of causality. However, our data on DM research output starts with the very first articles explicitly mentioning DM, there are few status changes of countries over the 23 years covered, and most status changes occurred recently. This should diminish problems with selection bias, since it is improbable that policy makers foresaw the potential of DM when designing relevant copyright policy. What is more, international harmonization of copyright law in particular within the European Union restricts national policy makers in tailoring copyright to perceived domestic needs. Arguably, there is thus a strong random element in countries' status regarding DM-related copyright policy. We are not aware of any better approximation of a natural experiment, and our results do provide a better evidence-base for policy than has been available so far.

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.[15] 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 data mining practitioners, we were even told that it is common practice to deliberately locate DM activities in territories with weak copyright protection, and to seek out suitable partners from such territories in international DM co-operations. 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 that is beyond the scope of this paper.

[15] 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. 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.

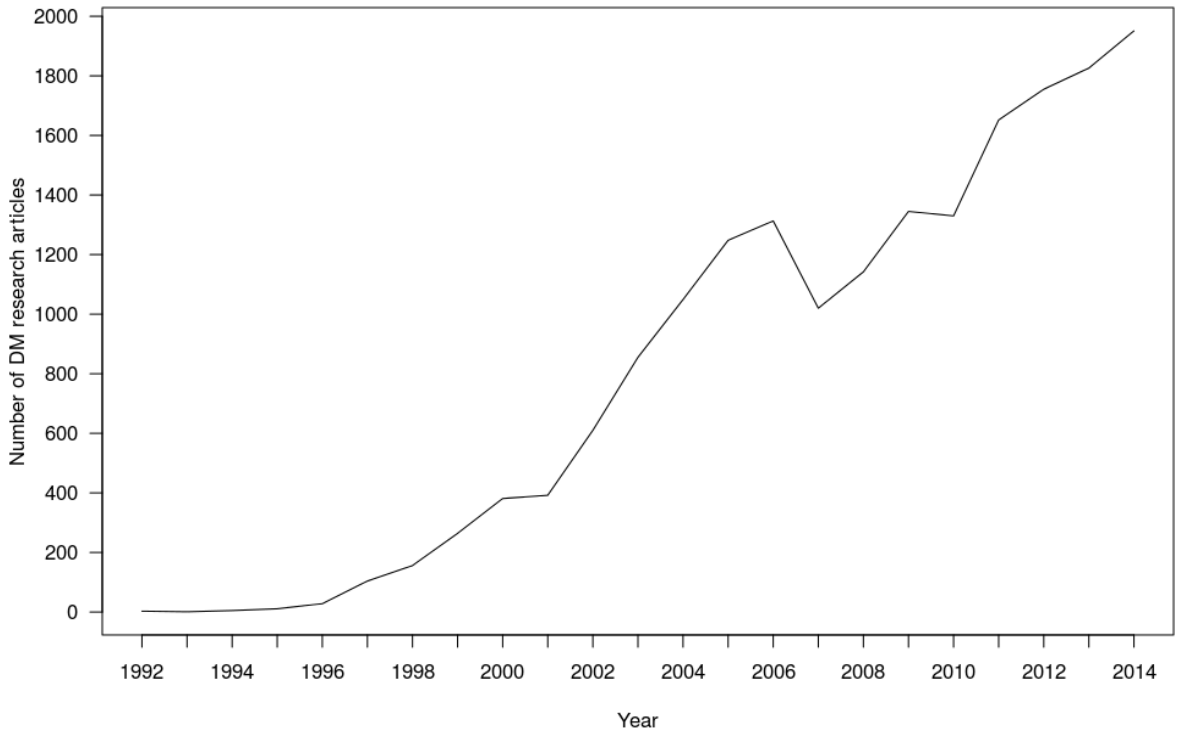
Conclusions

DM is the topic of an increasing number of academic journal articles. We show that copyright affects 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.

Our results suggest that in the case of academic research and DM, the adverse consequences of copyright protection on the creation of new information goods are greater than the benefits. DM research often draws on many input works to which others hold copyrights. Copyright exceptions or limitations could promote this type of research, at least to enable DM of input works that have been publicly financed.

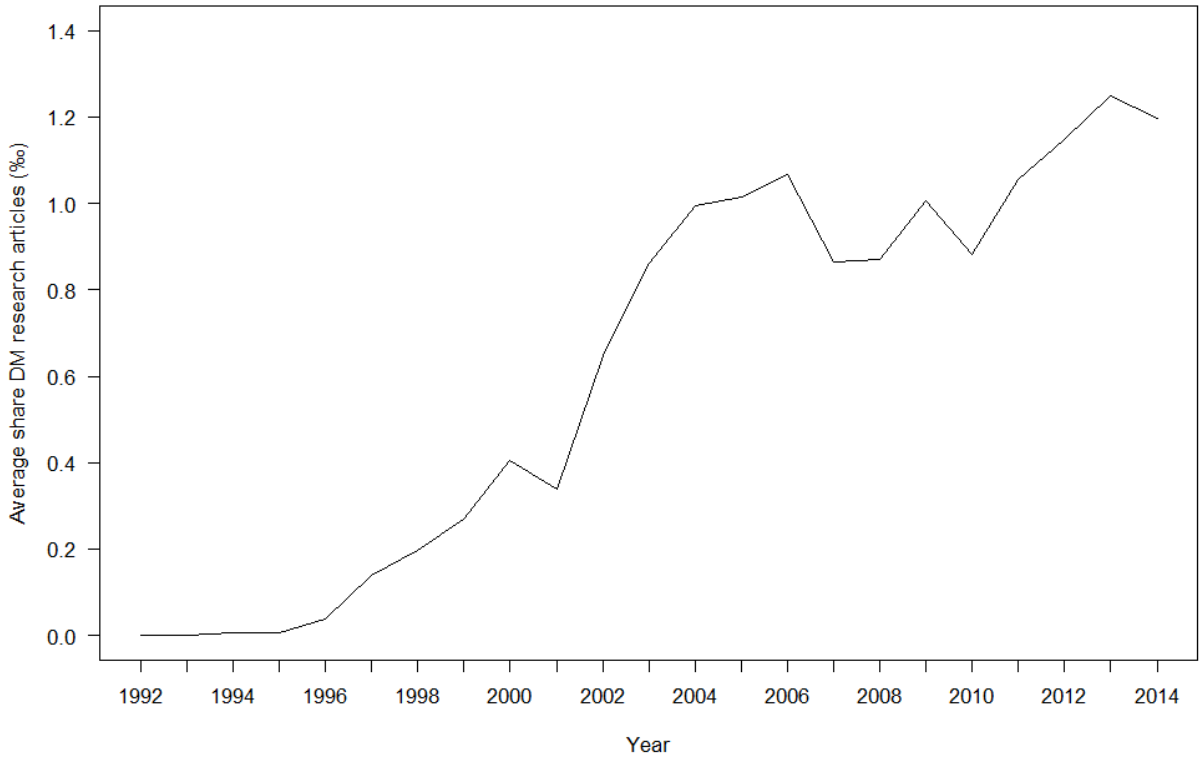
Acknowledgements. No external funding was used in the preparation of this paper. The authors are grateful to Merel Goedknegt and Saskia Woutersen-Windhouwer for effective research support. This paper also benefitted from comments by the participants of the research seminar on Cultural Economics at ESHCC, 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 feature on the conference websites and SSRN and have received the LIBER Innovation Award 2015 and the Finalist Best Paper Award at the EPIP conference 2015.

Figure 1. Absolute number of DM research articles published per year (42 countries; 1992 to 2014).



SOURCE: Own calculations based on search results on the WoS database.

Figure 2. Average share of DM research articles in the total number of research articles published per year and country in ‰ (42 countries; 1992 to 2014).



SOURCE: Own calculations based on search results on the WoS database.

Table 1. Summary data on DM research output, 1992-2014 (countries in order of GDP in 2013 according to the World Bank)

	First DM article published in …	Total number of DM articles; 1992-2014	Average DM share in ‰; 1992-2014	Average DM share in ‰; 2003-2014
EU largest 15 Member States[1]	1994	5,437	0.63	1.00
East Asia largest 7 countries[2]	1996	4,801	1.07	1.34
USA	1992	4,827	0.69	1.00
China	1997	2,063	0.84	1.27
Japan	1996	585	0.34	0.49
Germany	1996	863	0.47	0.68
France	1996	591	0.44	0.65
England [3]	1994	913	0.55	0.78
Brazil	1999	281	0.49	0.74
Italy	1997	605	0.58	0.86
Russia	1997	83	0.14	0.20
India	1999	386	0.46	0.70
Canada	1992	782	0.74	1.07
Australia	1995	680	0.92	1.31
Spain	1998	696	0.74	1.11
South Korea	1997	583	0.80	1.16
Mexico	1997	92	0.37	0.51
Indonesia	2004	7	0.30	0.46
Netherlands	1995	256	0.43	0.62
Turkey	1998	243	0.54	0.81
Saudi Arabia	2003	48	0.47	0.72
Switzerland	1997	159	0.35	0.52
Argentina	1999	35	0.22	0.33
Taiwan[4]	1997	1,160	2.41	3.64
Sweden	1997	135	0.31	0.44
Poland	1998	259	0.70	1.03
Belgium	1997	271	0.80	1.16

	First DM article published in …	Total number of DM articles; 1992-2014	Average DM share in ‰; 1992-2014	Average DM share in ‰; 2003-2014
Nigeria	2004	4	0.11	0.17
Norway	1997	46	0.24	0.32
Venezuela	2002	11	0.46	0.71
Austria	1999	112	0.50	0.74
United Arab Emirates	2007	3	0.53	0.82
Thailand	2000	72	0.71	1.09
Colombia	1999	24	0.53	0.68
Iran	2005	233	0.53	0.81
South Africa	1997	43	0.26	0.37
Denmark	2000	85	0.32	0.49
Malaysia	2001	92	0.83	1.27
Singapore	1996	246	1.59	2.15
Israel	1997	216	0.74	1.05
…				
Finland	1994	159	0.78	1.05
Greece	2000	215	1.07	1.64
Ireland	1997	140	1.00	1.31
Portugal	1998	137	0.79	1.18

SOURCE: Collected by authors on the Web of Science database.

(1) EU largest 15 are Germany, France, England (see table footnote 3 below), Italy, Spain, Netherlands, Sweden, Poland, Belgium, Austria, Denmark, Finland, Greece, Ireland and Portugal.

(2) East Asia largest 7 are China, Japan, South Korea, Taiwan Thailand, Malaysia and Singapore.

(3) Web of Science reports on England, Scotland, Wales and Northern Ireland separately. We only report the figures for England.

(4) 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. Its rank is approximate.

Table 2: 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)*
Not allowed DM requires express consent	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.	Not relevant	- All EU and EFTA member states except for the United Kingdom since 2014 - All Latin American countries covered - Russia - Switzerland - Turkey
Probably not allowed DM probably requires express consent but there has been no ruling against DM researchers	There is a fair dealing exception that could potentially apply to DM.	There is no relevant case law specifying whether DM qualifies as an act of fair dealing	- Australia - Canada before 2012 - China 2007-2012 - India - Ireland - Israel before 2008 - Japan before 2010 - Korea before 2011 - Malaysia - Nigeria - Singapore before 2005 - South Africa - Taiwan before 2003 - Thailand
Probably allowed Lawful access is probably sufficient but there has been no ruling in favor or DM researcher	There is a fair use defence 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
Allowed 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 - United Kingdom since 2014
Not classifiable			- China before 2007 - Indonesia - Iran - Saudi Arabia - United Arab Emirates

Table 3. Descriptive statistics regarding the DM share in copyright categories.

Copyright category	Average DM share in ‰	Standard deviation	Number of observations
Not allowed	0.54	0.54	551
Probably not allowed	0.60	0.66	184
Probably allowed	1.76	1.37	59
Allowed	0.70	0.18	6

Table 4. Results of multilevel regression models with DM share as dependent variable and with varying intercept (random effects).

	Model 1	Model 2	Model 3	Model 4	Model 5
(Intercept)	0.542*** (0.075)	0.709 (0.860)	0.980 (0.826)	3.800*** (0.913)	3.800*** (0.924)
Copyright [Ref. *Not allowed*]					
Allowed (6 observations)	0.301 (0.280)	0.342 (0.357)	0.398 (0.366)	0.180 (0.345)	9.732 (27.252)
Probably allowed	1.460*** (0.161)	1.453*** (0.168)	1.520*** (0.184)	1.134*** (0.187)	0.880*** (0.195)
Probably not allowed	0.073 (0.137)	0.296** (0.130)	0.350** (0.145)	0.081 (0.150)	-0.022 (0.156)
GDP/capita ($1,000)		0.010** (0.004)	0.007* (0.004)	-0.019*** (0.005)	-0.006 (0.006)
Population (log)		-0.021 (0.047)	-0.036 (0.045)	-0.457*** (0.067)	-0.440*** (0.067)
Rule of Law		-0.001* (0.0004)	-0.001* (0.0004)	-0.001* (0.0004)	-0.418*** (0.120)
EU Member (Yes)			-0.103 (0.145)	-0.193 (0.151)	0.127 (0.182)
Total research output (log)				0.551*** (0.061)	0.505*** (0.063)
*Definitely allowed**Rule of Law					-6.958 (20.726)
*Probably allowed**Rule of Law					0.418*** (0.120)
*Probably not allowed**Rule of Law					0.416*** (0.120)
R^2	0.165	0.168	0.168	0.275	0.305
F	50.079*** (df = 3; 763)	18.704*** (df = 6; 557)	16.008*** (df=7; 556)	26.338*** (df = 8; 555)	21.970*** (df = 11; 552)
N	767	564	564	564	564

*p<0.1; **p<0.05; ***p<0.01