Economic
Research
Working Paper
No. 75/2023

Digitization and Availability of Artworks in Online Museum Collections

Alexander Cuntz, Paul J. Heald, Matthias Sahli

WIPO

Digitization and Availability of Artworks in Online Museum Collections *

Alexander Cuntz † Paul J. Heald ‡ Matthias Sahli §

August 16, 2023

Abstract

We provide quantitative evidence from museum collections about how copyright status affects the availability of digital images of artworks. The paper applies a regression discontinuity and differences-in-differences design to estimate online availability of artworks from U.S. collections on digital platforms. We find a strong increase in the availability of digital surrogates when copyright is perceived to expire and original artworks are likely to transition to the public domain. Moreover, artworks and surrogates made available see a large number of *downstream* reuses based on google image search data, which indicates online availability is of commercial and public value independent of right status. Notably, we show that *upstream* surrogates of public domain artworks made available by museums are positively correlated with higher image resolution quality as compared to digitized artworks still protected under copyright laws. At the same time, it seems expressed industry norms can help encourage U.S. museums to also make low-resolution surrogates of copyrighted artworks available.

Keywords: copyright; museum; digitization; creative industries; availability; public domain; paintings; images; empirical
JEL Codes: L17; O34

1 Motivation

How does the copyright or public domain status of a creative work affect its distribution in the market? In theory, a story or a song might need an owner with the incentive provided by an exclusive copyright to make an adequate supply of new copies of the work available to the public (Landes and Posner, 2003; Zemer, 2006). The assumption that copyright law increases the availability of new copies of

*The authors would like to thank Imke Reimers, Guy Pessach, Yaniv Benhamou, Andrea Wallace, Alessio Muscarnera and Prince Ogoguguo as well as seminar and conference participants at IRENE UniNE (June 2021), WIPO (April 2022), the Munich Summer Institute (June 2022), The Society for Economic Research on Copyright Issues Annual Congress (SERCI, Sept 2022), European Policy for Intellectual Property Conference (EPIP, Sept 2022) for their comments on previous versions of the work. The views expressed are those of the authors, and do not reflect the views of the World Intellectual Property Organization or its member states.

†World Intellectual Property Organization, Department for Economics and Data Analytics, 34, chemin des Colombettes, CH-1211 Geneva. email: alexander.cuntz@wipo.int

‡Albert J. Harno Edward W. Cleary Chair in Law, University of Illinois College of Law. 504 East Pennsylvania Avenue Champaign, Illinois 61820 (706) 372-2567. Associated Researcher, CREATe, RCUK Centre for Copyright, University of Glasgow. email: heald@illinois.edu

§World Intellectual Property Organization, Department for Economics and Data Analytics, 34, chemin des Colombettes, CH-1211 Geneva. email: matthias.sahli@wipo.int; University of Neuchâtel, Institute of Economic Research, A.-L. Breguet 2, CH-2000 Neuchâtel. email: matthias.sahli@unine.ch

older works has been adopted by the U.S. Congress,[1] legal commentators (Bitton, 2011; Ginsburg et al., 2000) and the U.S. Supreme Court in *Eldred v. Ashcroft*.[2] Creative works are quasi-public goods, and most economic theorists of intellectual property law agree that some level of exclusivity may be needed to incentivize the creation of inexhaustible, non-rivalrous intangibles like stories and songs (Landes et al., 2003). However, once the creation of a work is adequately incentivized by a period of exclusivity, the need for additional protection to assure long-term commercialization and distribution is an open question subject to empirical investigation (Buccafusco and Heald, 2013). In this paper, by collecting large-scale data from museums and online collections, we examine the question of how copyright status affects the availability of digital images of artworks.

We touch upon several notable economic mechanisms around the digitization of artworks and museum efforts to make collections available online. First, arguably, by engaging in new digital channels, museums might cannibalize some of their existing sources of revenue (Greenstein et al., 2013). For example, while digital consumption often stimulates onsite visits, it can also replace some ticket sales by the museum (Fernandez-Blanco and Prieto-Rodríguez, 2020). At the same time, making collections available online might have complementary and promotional effects and generate new demand and new sources of income to museums (e.g. revenue from licensing digital images). In turn, this can positively affect museum sales, reputation and turnover (Borowiecki and Navarrete, 2017; Bakhshi and Throsby, 2014). Second, by creating new digital goods and services and museums being agents of change, museums have substantial control over new digital uses and act as gatekeepers of cultural goods with increasing importance that, arguably, are of both of public and private value. Third, the transaction cost incurred by museums for copyright clearance as well as the attached risk and expected litigation cost might lower the overall economic incentives to make digital images and artworks available to online audiences, independent of the actual right status. In this way, our research contributes to the long-standing debate in the law and economics literature on resource allocation and transaction cost problems around the provisioning of exclusive property rights (Coase, 1960; Calabresi, 1960; Posner, 1972; Depoorter and Parisi, 2002; Marciano, 2012).

[1] H.R. REP. NO. 105-452, at 4 (1998) (extending term of copyright protection to existing works "would provide copyright owners generally with the incentive to restore older works and further disseminate them to the public."); see also Excerpts of [former Patent and Trademark Commissioner] Bruce Lehman's Statement Before Congress, September 20, 1995, available here ("[T]here is ample evidence that shows that once a work falls into the public domain it is neither cheaper nor more widely available than most works protected by copyright. One reason quality copies of public domain works are not widely available may be because publishers will not publish a work that is in the public domain for fear that they will not be able to recoup their investment or earn enough profit.").

[2] 537 U.S. 186, 207 (2003).

Our study of the digitization of thousands of images of paintings held by hundreds of museums around the world increases our understanding of the relationship between legal status (copyright or public domain) and the distribution of new copies of older works. By observing the distribution of images of public domain and copyrighted art works on the 'democratically-curated virtual museum' (Useum, 2022) Useum, we leverage a quasi-natural experiment made possible by a significant change in the U.S. copyright law in 1998. This allows us to casually identify a positive distributional effect of perceived public domain status on the availability of digital artwork images. On average, public domain status of artworks increases their online availability by 34 to 42%, as compared to younger artworks still most likely protected under U.S. copyright laws. Moreover, we find preliminary evidence that average image resolution quality of public domain works is higher. At the same time, it seems expressed industry norms in the U.S. can help encourage museums to also make low-resolution surrogates of copyrighted artworks available.

It is common practice among U.S. museums to approximate and calculate the term of copyright protection based on the date of creation of the underlying artwork. According to the 'Tech Tutorial: Digital Copyright and Privacy' of the American Alliance of Museums, "works created before 1923 should be considered in the public domain", while "for works created after 1923, museums may need to seek permission to use the image" (American Alliance of Museums, 6 26). Arguably, while we cannot directly observe the actual right status of a work based on the date of first publication as stipulated in 'technical' copyright laws, our results nevertheless speak to the causal effect of 'perceived' public domain status on digital availability. Here, from an economic point of view, results show how most museums effectively interpret the legal environment and how that translate to economic behavior on markets, including the possibility that, in some individual cases, technical laws may be misperceived from a legal standpoint.

We further note that while we are able to approximate the copyright status of the underlying physical artworks, the copyright treatment of digital copies of public domain works varies from jurisdiction to jurisdiction in two ways.[3] First, a jurisdiction could potentially recognize a new copyright in a digital surrogate. Most major jurisdictions like the US, UK, and EU find that digital copies of public domain works lack sufficient originality to be protected; nonetheless, the question is unsettled in some countries. More importantly, cultural institutions, independent of legal precedent, sometimes

[3] c.f. section 4 for more details.

might claim ownership of digital surrogates and seek to control access or license digital copies made from public domain works in their collections. This is a major difference, for example, between U.S. and UK institutions, with UK museums being much more likely to lay claim to digital surrogates (Wallace, 2022). For this reason, some digital surrogates may 'behave' like copyright works in the market, dampening the size of the positive public domain effects that we find. Given the difficulties in identifying two legal classes of digital surrogates, however, we assume in the remainder analysis that all digital copies of public domain works are in the public domain.

The paper is structured as follows. Section 2 reviews the related literature. Section 3 describes the data and section 4 outlines the empirical strategy. Sections 5 presents the main empirical findings. Section 6 concludes.

2 Literature Review

A growing number of studies, most of them looking exclusively at book markets, conclude that when works fall into the public domain, the supply of new copies increases (Biasi and Moser, 2021; Heald, 2020, 2014, 2007; Reimers, 2019; Flynn et al., 2019; Li et al., 2018; Buccafusco and Heald, 2013). As we do, several studies take advantage of changes in copyright protection to measure distributional effects. Biasi and Moser (2021) measure increases in the use and availability of German science texts before and after a UK WWII declaration that they were no longer protected by copyright. Heald (2007) and Reimers (2019) found that U.S. books from the same approximate era (1910s-1930s) moved from out-of-print to in-print significantly more quickly after they fell into the public domain. These results have been replicated for UK and Canadian book markets (Heald, 2020). Li et al. (2018) observed a similar positive public domain effect after a change in the UK copyright term length was made in the 19th century. The only study taking a true random sample of titles in-print on the Amazon.com virtual bookshelf found a dramatic positive effect in availability associated with public domain status (Heald, 2014).

Relatedly, Flynn et al. (2019) found that public domain books were more likely to be offered as digital rentals by libraries than similarly situated copyrighted books. And in one of the few studies not involving books, Heald (2009); Heald et al. (2012) found that public domain musical works were

just as likely to be used in movie soundtracks as their copyrighted counterparts.

Only two studies that we are aware of even tangentially touch on the distribution of images. While researching the value of public domain images on Wikipedia, Heald et al. (2015) searched biographical entries of bestselling novelists for photographic images of the authors. They found that between 80-90% of the Wiki pages of authors born before 1880 had photographs while between 40-50% of the pages for authors born after 1920 contained a photograph of the writer. Given that (at the time of the study) works published before 1923 were in the public domain, the results suggested that the existence of a public domain photograph increased the likelihood that an image of the author would be found on his or her Wiki page. The finding is particularly interesting because presumably fewer photographs of oldest authors are likely to exist. Nagaraj (2018) uses variation in copyright status of issues of Baseball Digest magazine by Google Books to measure the reuse on Wikipedia. The author found, among others, that copyright has a larger negative impact on image reuse compared to text.

Although several studies measure the distributional effects of public domain status, only one study has explored the potential effect of legal status on quality. In a human subject experiment, Buccafusco and Heald (2013) compared the quality of audiobooks made from bestselling copyrighted and public domain titles, finding that copyrighted and public domain titles attracted readers of equal quality. Although the authors suggest that copyright ownership may not be necessary to assure quality control, the relationship between the legal status of a work and the quality of a new version of the work remains an interesting and wide-open question.

Our research takes the first in-depth look at what is likely to happen - in both distributional and quality terms - when a painting or other artistic image transitions to the public domain. The studies cited above suggest that digital images of art work should become more widely distributed when they fall into the public domain. However, this hypothesis is weakened by two aspects relevant to the availability of visual artworks that are somewhat different to purely commercially-driven book or music markets. First, museums in physical possession of a public domain work are under no obligation to produce a digital image for public consumption. They act as gatekeepers with control over access to the only tangible copy of the work. Nonetheless, beyond commercial interest, many museums also have a public (dual) mission to preserve and make available the cultural heritage to the general public. Second, from a business perspective, museums might strategically choose to only

release lower quality images in order to maintain consumer interest in visiting the museum and the physical work exhibited on site, or earn additional income from the licensing of higher resolution images to recover digitization costs. This again calls for an empirical investigation in this industry.

3 Data

3.1 Raw data and estimation sample

To address our empirical question, we extract data from two large online platforms in the visual art market Useum and Artnet. The raw data collected from *Useum* contains detailed information on the individual artwork and artist-level as well as providing for rich information on user ratings (e.g. page views, followers and likes). The data includes information on the artwork title and year of creation, as well as information on the place of birth and years of birth and death of the artists, while short biographies are available for more popular artists. Furthermore, the *Useum* data provides links to original image sources such as the museum collection urls or *Wikidata* pages.[4] As a second step, we webscrape and collect additional information from Artnet, as artist information on *Useum* is sometimes incomplete or hard to extract.

In general, *Useum* is a large online provider and important source of art information with more than one million followers. It considers itself the 'world's art museum that brings together art museums, galleries, artists and art lovers'. The service operates an integrated mapping of artworks hosted on the *Wikimedia Commons*, which automatically collects artwork information and digital surrogates from the latter platform. Moreover, as 'crowd-sourced art museum' (Useum, 2022), users on the *Useum* platform also contribute new artworks to the online catalogue of digital surrogates. First, artists themselves (or their agents) contribute surrogates to the platform by uploading their own artworks or by selling ancillary products on the *Useums'* webshop against a small commission fee to the service (e.g., mugs or t-shirts showing their artworks)[5]. Second, the *Useum* service also engages with art historians that voluntarily contribute new works to the collection, mostly for educative purposes. And, third, general users also contribute and discover new digital surrogates by engaging with like-minded peers and 'art-lovers' online, and by rating, liking or tagging artworks hosted on the service.

[4] Examples include 'Vincent van Gogh - The Starry Night' from the Museum of Modern Art, or 'Pablo Picasso - Les Demoiselles d'Avignon' originally hosted on Wikidata pages.

[5] More information on the commercial service can be found here.

Most artworks in the *Useum* collection provide for a link to the original source of digital surrogates, which redirects users to original museum websites or *Wikidata* pages. More details on the data collection process can be found in the Appendix 6.

The raw *Useum* data has close-to-global coverage and represents artworks from 1'286 museums located in 59 countries. Most institutions are resident in the United States (22 %), the Netherlands (21 %) and the United Kingdom (13 %). For a full list of countries represented in the data and for the top-20 list of museum venues, please refer to table 14 and table 15, respectively. Notably, the raw data consists of more than 130'000 thousand artworks created by 15'780 artists.[6] In sum, digital artwork images have received more than 56,6 mio. page views and 59'827 likes from service users. Analogously, museums on the platform have received a total of 46,6 mio. page views and 56'829 likes since the launch of the *Useum* platform in 2012 to the point of data collection (2021). Figure 1 provides examples of the digital surrogates of artworks by Wassily Kandisky and Mark Rothko posted on *Useum*. Most prominent artists in the raw data and in terms of the total number of digitized artwork (N) include Peter Paul Rubens (692), Edvard Munch (657), Anthony van Dyck (605), Auguste Renoir (543), Claude Monet (517) and Vinent van Gogh (510) (see table 15 for the top-20 list of artists). Artists are distributed and represented across different museums around the world. For instance, the 657 artworks attributed to Dutch artist Peter Paul Rubens are associated with 126 different museums in our raw data. On average, artists were born in 1760 and died in 1827 (mean). In terms of their popularity on the platform, the average artist has received a total of 3'662 page views and 18 likes (mean). Overall, the average digital artwork has received (standard deviation in brackets) 0.41 likes (2.53) and 365 page views (1173.8) since the launch of the *Useum* platform.

For the estimation, we narrow down the *Useum* data to include only artworks coming from museum collections in the US, UK and the EU. And, we further limit the sample to artworks created between 1910 to 1940. Table 1 gives descriptive statistics for the estimation sample used in the baseline regression discontinuity design and differences-in-differences estimations. Between 1910 and 1940, the average artwork creation year is 1918. Given the relevant perceived public domain cutoff year is in 1926 (for artworks in US museum collections), we observe that roughly 84 percent of artworks are

[6] We collect additional information on the year of artist death and the year of creation of the artwork. This is required to calculate the copyright term and approximate the copyright status of artworks. Data extraction from *Artnet* results in a total sample of 353'739 unique artists where information is complete. Based on an exact, non-fuzzy matching procedure which minimizes the problem of potential false-positives, we complement missing information on artists for a total of 72'709 artworks in our inital *Useum* data. We cannot attribute artworks to an artist (name) for around 12 percent of the total sample, i.e. most artists are recorded as 'unknown' or 'anonymous' in the data.

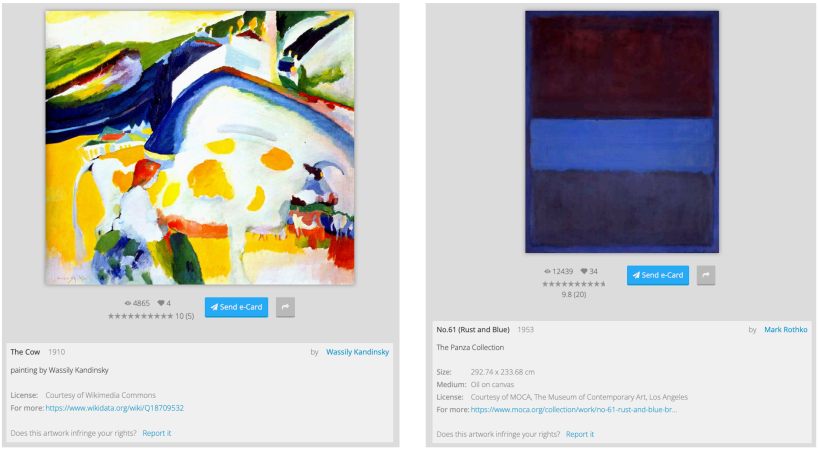

Figure 1: Useum Example - Museum of Contemporary Art Los Angeles and Wikimedia Commons

Note: This figure illustrates two digital surrogate examples from the Useum Webpage originating from different sources: the painting by Wassily Kandinsky (The Cow, 1910) via Wikimedia Commons (left) and a painting by Mark Rothko (No. 61 Rust and Blue, 1953) via the Museum of Contemporary Art Los Angeles (right). The examples further illustrate the artwork-level information (likes, views, ratings) and Useum-page information (buy or download options). The bottom panel lists important artwork- and artist-level information such as size, medium, © or licensing information.

created before this year. Overall, the average artwork has received 431 page views, average artist has received around 54 thousand page views, and the average museum has received 706 thousand page views since the launch of the *Useum* platform in 2012 until the point of data collection (2021). Roughly 70 percent of the curated observations in the sample are associated with museums in the EU and the UK, the remainder is from US museums by definition of the sample. The average number of observed artworks in a given country and year of creation (log) is around 44 (3.35) artworks. Finally, the average number of observed artworks per artist in a given year of creation (log) is around 3.5 (0.73) artworks. The next section introduces the complementary data on the downstream and upstream reuse of digital surrogates.

3.2 Data on Downstream and Upstream Reuses of Digital Images

We rely on an approach pioneered by Erickson et al. (2018). The authors track online reuse of images found on Wikimedia Commons and apply an automated reverse-image search on *Google*, systematically recording and categorizing by type of source and reuse. Drawing on a random sample of ten thousand images on Wikimedia, they find more than 54 thousand downstream uses of the images and estimate a total value of the Wikimedia Commons Images universe based on an average (non-commercial and commercial) royalty license fees of around overall 28.9 billion USD (see also Heald et al. (2015) for a valuation approach).

Table 1: Summary Statistics

	Before 1926		After 1926	
Variable	Mean	SD	Mean	SD
Artwork date	1915.97	4.467477	1932.369	4.269737
Public domain	0.99	.0884647	.6801909	.466681
Death year	1933.306	12.24018	1944.278	11.26506
Artwork views	426.59	1512.669	457.3365	1519.991
Artist views	54993.69	113451.6	53313.43	97136.39
Museum views	614344	1362341	1195887	2218405
EU UK dummy	.6949076	.4604984	.7004773	.4583225
N country	49.68905	37.18562	11.45107	8.736929
log(N country)	3.478483	1.088615	2.06098	.962152
N artist	3.531278	4.924565	3.367397	6.007613
log(N artist)	.759522	.8917698	.6073791	.8935353

Note: Summary statistics for the $n = 5,276$ observed artworks in the EU, UK an U.S. between artwork creation year 1910 to 1940. The dependent variable 'N country/ artist' and 'N country/ artist (log) are the yearly number of digitized artworks in a given country/ of an artist, the logarithmic transformation respectively.

Thus, similar to the approach in Erickson et al. (2018), we can identify online reuses based on our data. Our initial sample of artworks on *Useum* includes, in most cases, an image reference URL source and we therefore can apply a *reverse Google image search* to identify possible up- and downstream reuses of an image in our data. This step requires extensive data management and manual inspection of the data quality. We focus on the U.S. estimation sample of artworks created between 1910 to 1940. Limitations and data caveats are discussed in greater detail in the appendix. We gather the data in two consecutive steps. First, as the *reverse Google image search* requires a concise link to the image (i.e. a $.JPG - link$), we construct an automated web-scraper to collect image links on the initial webpage of the artwork image, using $python$'s $seleniumwebdriver$ package. Second, using the detailed image links, we query and simulate a human-search in the *reverse Google image search* engine to collect search results, i.e. all links listed on 'pages with the matching images'. It is important to note that we exclude results (links) obtained from searches of visually similar images.

We define an 'upstream' reuse of an original artwork when the digital surrogate is associated with the museums' web page and assume the museum is the original source of the digital image. This is because the museum acts as a gatekeeper owning the physical artwork and has some discretion over quality and image resolution of the digital surrogate derived from the artwork. Put differently, we identify online reuses in the total sample of reuses where the surrogate of the artwork is likely the digitized image made available by the museum on its website. We consider all other reuses as 'downstream' when surrogates are created and sourced from another service, entity or individual such as

a commercial platform or another internet source, and there is no indication the museum owns the underlying artwork. Practically, we approximate and distinguish upstream from downstream reuses based on the top-level domain information provided for related webpages (cf. the pie chart in figure 9), which works well as a common denominator for base-URL of museum websites in our data (e.g. .lacma.org; .dia.org; .clevelandart.org; and so on). Lastly, we perform a non-fuzzy text-match with outcomes of the google searches and reuses of digital images. We find that out of the 88'607 links obtained from Google, 1'789 were upstream uses directly linked on the museum website. 1'494 upstream uses are obtained from artworks created before 1926 while 292 are from artworks created $\geq$ 1926. [7]

Overall, starting from an initial list of 1594 artworks hosted from *Useum*, and a subsample of 1349 artworks where the direct and functioning .JPG URL can be verified,[8] we obtain a total set of 88'607 upstream and downstream reuses (cf. table for these works 2). An average (median) artwork sees reuses on 66 (42) pages, with 191 artworks creation year $\geq$ 1926 and 1158 artworks created before 1926.[9] Finally, for each reuse of a digitized artwork, we further enrich data on reuses and collect detailed information on image quality, i.e. pixel $\times$ pixel information on the resolution of images.

Table 2: Summary Statistics: Google Downstream/ Upstream Use

	Before 1926		After 1926	
Variable	Mean	SD	Mean	SD
Artwork date	1916.024	4.202358	1932.605	4.642898
Quality (pixel $\times pixel$)	698941.6	1816344	790340.9	1793988
log(Quality)	12.52377	1.316987	12.59962	1.351164
Low quality dummy	.6514375	.4765184	.6407169	.4798069
Domain (.org) Dummy	.1267598	.3327059	.1217909	.3270554
Domain (.com) Dummy	.5242732	.4994138	.5048786	.4999935
Upstream dummy	.0201872	.1406411	.0202062	.14071

Note: This table shows the descriptive statistics of the upstream- and downstream-uses obtained from a google reverse image search with artwork creation year between 1910 and 1940. The dummies 'domain' and 'upstream' are derived from semi-fuzzy text-analysis methods on the base-URLs. $N = 88'607$.

[7] These 1'789 upstream uses derive from 103 (567) unique artworks created after (before) 1926. There are several reasons why we lose artworks when we filter for upstream uses only. On the one hand, the google reverse image search might does not show the overall 'universe' of reuses for technical/ algorithmic reasons (e.g. IP addresses) or misleading/ incorrect metadata behind the JPG URLs or on the websites. On the other hand, not all museums feature artworks on their website, and some changed upload policies or use other image libraries. We have reached out to several museums that confirmed this; e.g. the METMuseum updated how they identify public domain images and some MET images on Wikimedia were added before the updated policy. As a consequence, there "is a difference between what is available on the MET website and what is on Wikimedia" (Mail correspondence April 2022). The Cleveland Museum of Art also continuously updates works to Wikimedia, and: "public domain works might have multiple entries in Wikimedia due to AMICO, Artstor and any other image library from 15 years ago that tried to aggregate museum collections" (Mail correspondence April 2022).

[8] The loss of the remaining 245 artworks during this procedure is likely due to dis-functional direct .JPG URLs, 'hard-to-scrape' pages, no initial direct URL or other errors and we expect no systematic error.

[9] Which is true by construction as our initial source of data *Useum* is a downstream-use itself and thus we know artworks were reused at least once and as a consequence should appear in a reverse google image search.

4 Empirical Framework

In the empirical strategy, we implement a regression discontinuity design (RDD) as pioneered in the literature (Imbens and Lemieux, 2008; Angrist and Pischke, 2008; Lee and Lemieux, 2010). We follow and build on previous work in copyright economics evaluating the impact of the U.S. copyright on the reuse of music (Watson, 2017), as well as the impact on prices and the availability of books once works transition to the public domain (Reimers, 2019; Heald, 2007). To provide causal evidence, we exploit exogenous variation in copyright status introduced by the 1998 copyright term extension in U.S. laws. The design allows us to causally identify treatment effects in a non-experimental setting. The '1998 Copyright Term Extension Act' extended the copyright protection for an additional 20 years for works that were still under protection at that time, from 75 years to 95 years of copyright protection from the date of publication. Given that the data is collected in 2021, the copyright reform creates a sharp, exogenous discontinuity in the copyright status of artworks copyrighted before and after 1926 that were still in copyright at the time of the reform and assuming all formalities in laws were met (e.g. initial registration and potential renewal of copyright at the U.S. Copyright Office). It is important to note that under pre-1978 U.S. law, the length of a copyright is calculated from the moment of 'publication'. However, as we only observe the year of artwork creation, we use this point in time as an imperfect cutoff between (perceived) public domain and copyright protection in the present analysis. Given the data at hand, this seems the best possible approximation for an econometric design exploiting the copyright and public domain status of artworks.[10] Because territorial copyright reform mostly affected creative works located in the United States, we limit the main data sample for the RDD results to artworks created between 1910-1940 and artworks originating from U.S. museums. [11].

[10] 'Publication', an extremely complicated term of art. The Copyright Office has stated that mere display of a work of visual art in a gallery does not constitute publication, but sale at auction or inclusion of an image of the work in a printed book or catalog (or a post card or magazine?) does constitute publication (U.S. Copyright Office, "Publication," please see here). Early 20th century case law left the matter unsettled, with at least one prominent case finding display in a gallery could constitute publication of a painting (Gerhardt, 2011). Interestingly, museums seem to consider the date of creation to be the relevant moment for calculating the 95-year copyright term, perhaps because that date is far easier to discern than the date of first display, sale, or inclusion in a catalog, book, or museum- or gallery-printed publication. Although there is no support in the law for calculating the U.S. copyright term from the date of creation, the institutions in our study seem to use it as a rule of thumb, agreeing on a synthetic copyright term based on norms rather than technical law.

[11] It is worth noting that the 1998 Extension Act also applied to the protection of foreign works (see, for instance, Heald (2007)). Permissible within the framework of the *Berne Convention*, the U.S. (as other contracting parties) will not discriminate between the protection granted to foreign vis-à-vis domestic works. For example, this means that some foreign works might have transitioned to the public domain in their 'home country', but they will continue to be protected under U.S. copyright laws.

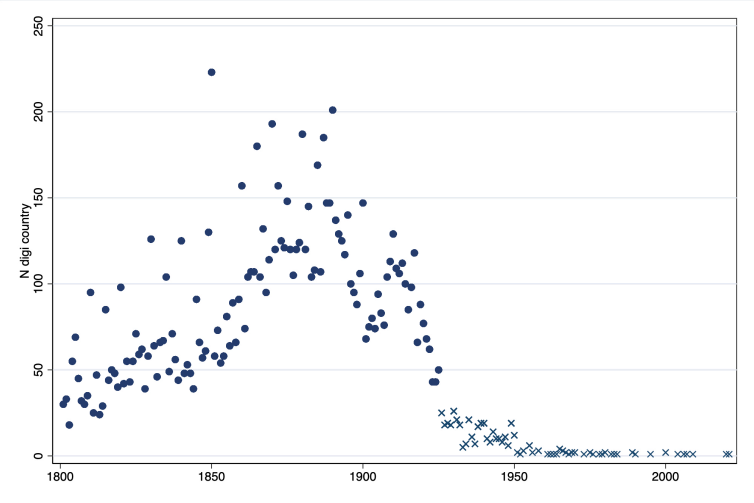

Figure 2: Number of Digitized Artworks: Raw Data

Note: This figure shows the 'raw data' for the number of yearly digital artworks found on *Useum*. The data includes digital artwork surrogates from museums in the U.S. for artworks created between 1800 and 2021. The crosses symbolize the cut-off creation year in 1926 and later where artworks are under copyright protection while the blue dotes represent numbers of artworks in public domain.

Most jurisdictions in our sample take the artist's year of death as the point of departure to calculate copyright terms and grant artists between 50 to 95 post mortem years of protection. However, in the RD design, we focus on the U.S. legal framework and, accordingly, the date of publication of the original artwork as approximated by the date of creation. Notably, this generates within-artists variation in perceived copyright status for original artworks in U.S. museum collections. So, original artworks created by the same artist before 1926 could potentially be in public domain, while original artworks created after this cutoff will likely remain under copyright, assuming formalities in laws were also met. In the RD design, the assignment to the treatment (public domain) follows a known rule. As shown in Figure 2, we observe a discontinuity in the raw data for creation dates of original artworks. Thus, we formally construct the 'forcing' variable (Lee and Lemieux, 2010) of $artwork$ indexed by artist i in creation year t as:

$$PublicDomain_{it} = \begin{cases} 0 & \text{if } artwork_{it} \geq 1926 \\ 1 & \text{if } artwork_{it} < 1926 \end{cases}$$

where 1926 is the known threshold-year for copyright protection in the U.S. (in 2021 for works that were likely in copyright and subject to the 98 term extension) of the discontinuous function of the creation year of an original $artwork_{it}$. This leads to the regression equation we used to compute

the results:

$$Y_{it} = \alpha + \delta PublicDomain_{it} + k(creationyear_t) + \rho_I + \mu_M + \boldsymbol{X'}_{it} + \epsilon_{it} \quad (1)$$

with the effect of interest captured in δ, the effect of a perceived public domain status (of the original artwork) on its digital availability, and $k(creationyear_t)$ is the artwork creation year. We add artist fixed effects ρ and museum fixed effects μ to the regression. We restrict this regression to available artworks of museums located in the United States that potentially could be affected by the copyright reform and tighten the observation window for artworks created around the cutoff in 1926. In our setting, the so-called forcing (or running) variable (Lee and Lemieux, 2010) is the year of original artwork creation and our dependent variable (Y_{it}) is the log-transformed number of digitized artworks from a given creation-year (on the aggregated country- or artist-level) that we observe online from U.S. museums.

Lee and Lemieux (2010) summarize important checkpoints for the causal identification to hold. As noted in their paper, estimates can be biased if individuals are able to manipulate the 'assignment variable'. This is not a concern in our setting as the reform in 1998 was implemented retrospectively for artworks created in 1923 and later that were still in copyright at the time of the term extension, and so it did not affect the incentives to create new artworks back then. [12] We provide robustness checks in section 5.3 and further inspect in section 5.4 if we can observe discontinuities in other observable artwork characteristics around the cutoff date.

In a second empirical framework we exploit variation in copyright status between the previous results on artworks availability in U.S. museums compared to EU and UK museums availability [13]. As the above described empirical framework suffers from the limitation that we don't know the true number of available artworks created before/ after 1926, we leverage the fact that we observe artworks of museums in several countries around this time. In a differences-in-differences design we compare the differences in number of available artworks (artist-/ country-level) closely around 1926 of U.S. artworks from U.S. museums compared to EU/ UK countries, i.e., museums. Although this

[12] It is important to note that there were no other relevant changes in copyright laws in this time period (Reimers, 2019).

[13] With the data at hand, the comparison of U.S. to EU and UK availability seems to be the most adequate counterfactual, based on numbers of available artworks in this time-period, that the EU+UK serves several cross-country counterfactual-trends (22 in our data), that the EU+U.K. rely on post-mortem copyright protection term, and the fact that the Useum platform is based in the UK.

framework relies on a cross-country control group (i.e. the countries need to have similar growth paths in dimensions, which is a stronger assumption and there is variation in post-mortem term protection in EU/ UK countries), we can learn and approximate by how much perceived public domain status increases (or decreases) the availability of artworks in U.S. museums. We therefore run the following type of differences-in-differences equation:

$$Y_{itc} = \alpha + \delta PublicDomain(US)_{it} \times US + \phi_I + \gamma_T + \beta_C + \boldsymbol{X'}_{it} + \epsilon_{itc} \tag{2}$$

where we again look at the number of available artworks Y_{itc} on the country- and artist-level, but in several museums in several countries. Y_{itc} is thus the number of observed artworks (log) for artist i, in creation year t and in museum country c. The coefficient of interest is captured in δ which compares pre- 1926 availability (perceived public domain in the US) of artworks in U.S. museums compared to museums in the EU and the UK. We again add museum, year, and country fixed effects (ϕ, γ, β) to the regressions.

5 Estimation and Results

5.1 Availability of Digital Artworks from Museums in the US

We now turn to the results of the identification strategy outlined in section 4. With Table 3 we introduce our preferred econometric specification, with results based on the dependent variable on the artist panel-level, i.e. the artists- 'yearly number of digitized artworks'. These results point towards a positive coefficient around 0.294 to 0.355 for the (log-transformed) dependent variable with statistically significant results in all model specifications with additional covariates. This structure allows us to also add to the regressions artists' fixed effects (cf. model (4) to (6) of table 3), coefficients are based on the within-estimator that controls for unobserved artist-specific and time-invariant heterogeneity. In model (6) we introduce multi-level fixed effects (museum and artist) to control additionally for museum-level differences of making artworks digitally available. In the appendix 6 of this thesis we report additional results based on the country-level and adopt a Poisson pseudo-maximum likelihood estimator.

Figure 3 further illustrates the above discussed results. The graphical approach is able to show the discontinuity in the function of artwork creation years as an approximation to the actual year of

Table 3: RDD: Availability (Artist-Level) (log)

	DV: log(N artworks Artist)					
	(1) RDD	(2) RDD Covs	(3) RDD FE	(4) RDD FE	(5) RDD FE	(6) RDD MLFE
Public Domain	0.325*	0.355**	0.338*	0.325*	0.331*	0.294*
	(2.48)	(2.74)	(2.40)	(2.53)	(2.56)	(2.19)
Artwork creation year	2.051	1.959	1.737	-0.333	-0.418	-0.119
	(0.91)	(1.07)	(0.97)	(-0.18)	(-0.23)	(-0.07)
(Artwork creation year)^2	-0.000535	-0.000510	-0.000452	0.0000842	0.000106	0.0000283
	(-0.92)	(-1.07)	(-0.97)	(0.18)	(0.22)	(0.06)
artist views		0.00000239***	0.00000216***			
		(5.29)	(4.26)			
museum views		6.22e-08	0.0000159		3.00e-08	0.0000150
		(1.26)	(1.04)		(1.47)	(1.27)
artwork views		0.00000998	0.0000151		-0.00000109	-0.00000442
		(0.59)	(0.92)		(-0.09)	(-0.37)
N	1592	1566	1566	1335	1313	1290
Cluster SE	Artist	Artist	Artist	Artist	Artist	Artist
R^2	0.0445	0.234	0.365	0.633	0.631	0.658
Artist FE	No	No	No	√	√	√
Museum FE	No	No	√	No	No	√

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: This table shows the RDD results based on the formula specified in chapter 4 and a DV as the (log) year-artist-level number of digitized artworks in model 1-6, restricted to museums in the United States and artworks created between 1910 to 1940. Model 4-6 are calculated based on artist fixed effects, and model 6 multi-way fixed effects model.

first publication. The portfolio of available artworks based on the artist-level increases sharply for those created before 1926 and presumably are in public domain in the U.S. as of 2021.

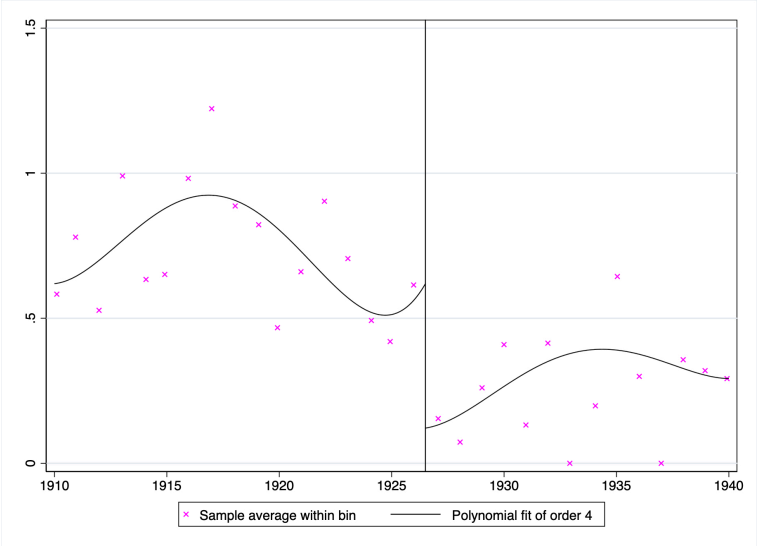

Figure 3: RDD: Digitized Artworks (log) (Artist-Level)

Note: This figure shows the (log) yearly number of digitized artworks on the artist-level and year of artwork creation date. The sample is restricted to years 1910-1940 and museums in the United States. The crosses represent sample average within bin and the line a polynomial fit of order 4. Standard error are clustered at the artist-level. The x-axis line denotes the copyright cut-off point in 1926.5.

In table 4 we present results based on the counter of number of yearly available artworks on the artist-level. Using the non-log transformation of the dependent variable can eventually help us to better understand the magnitudes of the coefficients, that we will discuss afterwards. Overall, the perceived public domain status of artworks (i.e. being created and likely published just before 1926) stays also here significant (at the 5%-level) positive of around 0.987 and 1.313 in the same model specifications as previous (1-5). We only lose statistical significance after including in model (6) multi-way fixed effects (museum and artist).

Table 4: RDD: Availability (Artist-Level)

	DV: N artworks Artist					
	(1) RDD	(2) RDD Covs	(3) RDD FE	(4) RDD FE	(5) RDD FE	(6) RDD MLFE
Public Domain	1.258*	1.402*	1.313*	1.078*	1.125*	0.987
	(2.35)	(2.53)	(2.36)	(2.23)	(2.27)	(1.91)
Artwork creation year	14.20	14.02	13.78	8.398	8.264	9.412
	(1.45)	(1.74)	(1.65)	(1.58)	(1.55)	(1.80)
(Artwork creation year)^2	-0.00369	-0.00364	-0.00358	-0.00219	-0.00215	-0.00245
	(-1.46)	(-1.74)	(-1.65)	(-1.58)	(-1.56)	(-1.81)
artist views		0.00000838***	0.00000770**			
		(3.49)	(2.95)			
museum views		0.000000281	0.0000422		0.000000111	0.0000485
		(1.39)	(0.85)		(1.21)	(1.23)
artwork views		0.0000865	0.000110		0.0000696	0.0000490
		(0.81)	(1.09)		(0.85)	(0.65)
N	1592	1566	1566	1335	1313	1290
Cluster SE	Artist	Artist	Artist	Artist	Artist	Artist
R^2	0.0429	0.229	0.338	0.665	0.666	0.694
Artist FE	No	No	No	√	√	√
Museum FE	No	No	√	No	No	√

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: This table shows the RDD results based on the formula specified in chapter 4 and a DV as the year-artist-level number of digitized artworks in model 1-6, restricted to museums in the United States and artworks created between 1910 to 1940. Model 4-6 are calculated based on artist-panel fixed effects, and model 6 multi-way fixed effects model.

Lastly, we balance our artist-level data on the right-hand- and left-hand-side around the cutoff year in 1926 around observed artists. One could argue that we estimate with sample-selection bias of different artists entering/ exiting the pre-/ post 1926 sample and as a consequence observe a biased, perceived public domain effect. [14] Given the structure of our data, we can address this point as follows: To construct the sample, we exclude artists who did not have at least one observation post 1926 in our U.S. sample. We observe 112 artists with artworks created and likely published after 1926. In the new (balanced) panel, artists had on average 2.4 artworks (median 2), with a standard deviation of 1.84 artworks, and we observe 247 artworks created in or after 1926 while 385 artworks

[14] This topic is also discussed in section 5.4 'random assignment around the discontinuity and placebo tests.

of the same artists were found prior the cutoff year.

Table 5 presents estimates based on the balanced artist-panel. We again look at the counter of created artworks (artist-year-level) using the same model-specifications (Poisson models) as before. We continue to estimate also in the artist-balanced panel positive significant coefficients. This sample increases the estimated dummy of available public domain artworks from around 0.382 (model 6) to 0.68 (model 2). This demonstrates that our sample, the estimated, perceived public domain coefficient, is also robust against using the balanced artist-panel compared to the previous results (cf. Table 12 model (6) 0.424). We visualize this result in figure 4. Also for the artist-balanced panel, results demonstrate a sharp discontinuity in the artwork creation year around 1926.

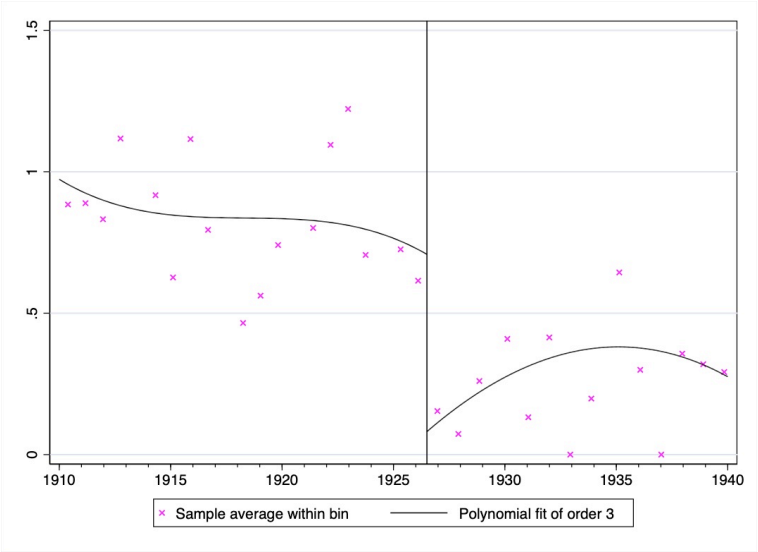

Figure 4: RDD: Digitized Artworks (log) (Artist-Level Balanced-Panel)

Note: This figure shows the yearly number of digitized artworks (log) on the artist-level and year of artwork creation date. The panel is balanced around artists who had at least one observation post 1926. The sample is restricted to years 1910-1940 and museums in the United States. The crosses represent sample average within bin and the line a polynomial fit of order 3. Standard error are clustered at the artist-level. The x-axis line denotes the copyright cut-off point in 1926.5.

5.2 Effect Size and Magnitude

Overall, we identify a positive effect of perceived public domain years on available artworks by year of artwork creation. Our preferred estimates on the artist-level in table 3 estimate a coefficient of observed artworks (log) in creation years 1926 or earlier of 0.355 (model 2) to 0.294 (model 6), all statistically significant at the 5%-level. Results indicate an increase of about 34 to 42% artworks (by artist) with creation years 1910-1926 compared to creation years 1927-1940. In table 4, we estimate

Table 5: RDD: Availability (Artist-Level) (Balanced)

	DV: N artworks Artist					
	(1) Poisson	(2) Poisson Covs	(3) Poisson FE	(4) Poisson FE	(5) Poisson FE	(6) Poisson MLFE
Public Domain	0.645	0.686*	0.587	0.418*	0.429*	0.382*
	(1.83)	(2.26)	(1.77)	(2.22)	(2.47)	(2.02)
Artwork creation year	-0.000534	0.00425	0.00365	-1.063	-0.976	-0.0141
	(-0.02)	(0.24)	(0.21)	(-0.39)	(-0.36)	(-0.01)
(Artwork creation year)^2	0	0	0	0.000274	0.000251	0.00000105
	(.)	(.)	(.)	(0.39)	(0.36)	(0.00)
artist views		0.00000179***	0.00000193***		0	0
		(4.55)	(6.43)		(.)	(.)
museum views		9.01e-08	0.000198***		2.57e-08	0
		(1.77)	(4.04)		(0.61)	(.)
artwork views		-0.0000325	-0.0000293		-0.0000214	-0.0000324
		(-1.57)	(-1.54)		(-1.08)	(-1.85)
N	632	628	628	580	577	556
Cluster SE	Artist	Artist	Artist	Artist	Artist	Artist
Pseudo R^2	0.0566	0.0917	0.1253	0.1721	0.1721	0.1846
Artist FE	No	No	No	√	√	√
Museum FE	No	No	√	No	No	√

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: This table shows the RDD results based on the formula specified in chapter 4 and a DV as the year-artist-level number of digitized artworks in model 1-6, restricted to museums in the United States and artworks created between 1910 to 1940. Model 4-6 are calculated based on artist fixed effects, and model 6 multi-way fixed effects model.

an increase of around 1 to 1.4 artworks by artist that are more available by these creation years (cf. table 1 with a yearly average of artworks by an artist that lies around 3.5 artworks in the overall estimation sample). We emphasize that we do not observe the entire catalogue of crated artworks of observed artists (or even more desirable: a sample of artists with the entire catalogue of existing (original) artworks, information on digital surrogates, and on artworks *not* digitized). Present numbers represent yearly averages calculated by artists and by year of creation. Previous results also point towards a clear cut positive effect of public domain years with the attempt to balance the panel on the artist level (cf. table 5). We challenge the effect size of our results with robustness checks, and test for random assignment around the discontinuity (to identify potential selection biases).

5.3 Robustness

In this section, we carry out a number of robustness checks and refine specifications from baseline models. A later section 5.4 will deal with identifying assumptions imposed by the RD design such as the random assignment of covariates around the discontinuity. In addition, the later section introduces separate placebo tests for a set of countries that do not see a change in copyright laws around the cutoff.

We recognize pre- and post-trends in figure 11 that could potentially bias our results. While the data shows a clear cut-off point around the year 1926, the treatment effects we estimate could nevertheless be driven by changes in digitization practices, changes in the compositions of museums in our sample, or other sources of potential bias. To address this issue, we run another set of model specifications and test for robustness of our results.

A robustness check is shown in figure 5 and table 6. We rerun and estimate regression from baseline models now further narrowing down time windows. For the panel on the left-hand side of figure 5 and models (1) and (3) in table 6, windows and samples are tightened further to artworks created between 1923 and 1927. Estimates give us the 'local treatment effect'. This shows that we can isolate out the copyright effect as artworks created between 1923 to 1926 continuously move to the public domain from 2018 onwards (see, for example, Heald (2007)). To be precise, an artwork created and likely published in 1923 had a maximum potential copyright protection of 95 years (once the 1998 copyright reform went into force, granting some artists 75+20 years of protection). So, it has moved, at the latest, in the public domain on January 1, 2019, while artworks created in 1924 and likely published shortly after entered the public domain at the latest in 2020, and so on. For this narrower time window, we should not observe any discontinuity in available artworks created between 1923 to 1925 compared to artworks created in 1926 or 1927 (which is not the case), even when, hypothetically speaking, other factors than copyright protection would explain the sharp drop in availability at the cut-off point in 1923.[15] The immediate sharp increase of perceived public domain works in figure 5 is scrutinized by the estimates. The effect size of the log-transformed yearly number of digital surrogates is 0.54 (which compares to the 0.8 in the baseline using the wider 1910-1940 window) and the effect continues to be significant in this specification (cf. model (1)). In terms of the total number of digitized works (log) on the artist-level in model (3), we cannot estimate significant results. Given the low number of observations using the multi-level fixed effects models at the cutoff, this can be expected.

To further verify our findings, we look at the longer time window of artworks created between 1800 and 1960 and re-estimate models (2) and (4) in table 6. Notably, the right-hand panel of figure 5 shows a smooth function in the number of digitized public domain artworks created and likely

[15] For example, one could argue that museums never digitized artworks created after 1923 for whatever reason, or that the copyright term for these artworks ended much earlier as their copyright was not renewed after an initial period of protection and so formalities were not met. Similarly, art historic reasons may play a role, or data providers may have systematically excluded artworks created after 1923. However, none of the above arguments seems very convincing.

Table 6: RDD: Availability (Bandwidth)

	DV: log(N artworks Country)		DV: log(N artworks Artist)	
	(1) at cutoff	(2) large sample	(3) at cutoff	(4) large sample
Public Domain	0.536***	1.724***	-0.139	0.360**
	(20.42)	(24.18)	(-0.41)	(2.76)
Artwork creation year	253.4***	0.00120	142.6	0.000907
	(16.43)	(1.84)	(0.79)	(1.31)
(Artwork creation year)^2	-0.0658***	0.000000515	-0.0371	-0.000000349
	(-16.43)	(1.76)	(-0.79)	(-0.99)
Artist views	0	0	0	0
	(.)	(.)	(.)	(.)
Museum views	0.00000617	-0.00000234	0.0000918	0.0000130
	(1.14)	(-0.42)	(1.63)	(1.62)
Artwork views	0.000000367	-0.00000705*	-0.0000395***	-0.00000406
N	92	15128	92	14987
R^2	0.979	0.730	0.860	0.586
Cluster SE	Artist	Artist	Artist	Artist
Artist FE	√	√	√	√
Museum FE	√	√	√	√

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: This table shows the RDD results based on the formula specified in chapter 4. Model (1) and (2) use the log-transformed yearly number of artworks in the U.S. as the dependent variable while model (3) and (4) continue with the log-transformed dependent variable on the artist-level. Model (1) and (3) are based on a sample with artwork creation dates between 1923 to 1927 and model (2) and (4) based on the overall creation year sample.

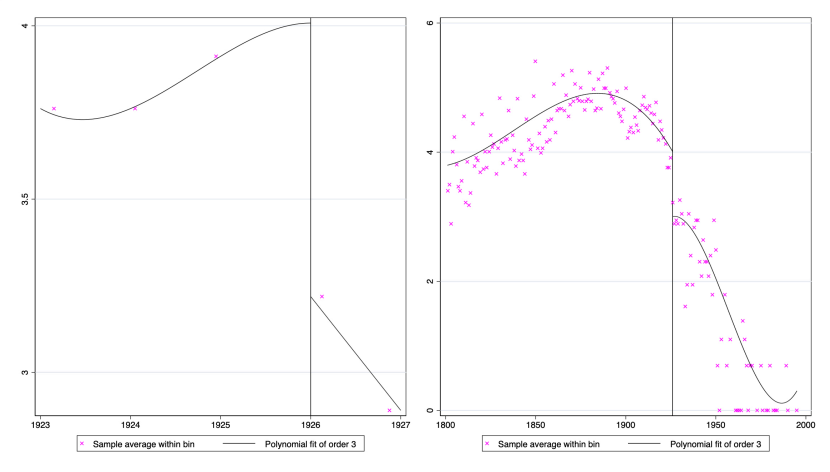

Figure 5: RDD Timeframe: Digitized Artworks (log)

Note: This figure shows the yearly number of digitized artworks (log) and year of artwork creation date (country-level). The sample is restricted to years 1923-1926 (left) and 1800-2000 (right) and museums in the United States. The crosses represent sample average within bin and the line a polynomial fit of order 3. Standard error are clustered at the artist-level. The x-axis line denotes the copyright cut-off point in 1926.

published between 1800 and 1926 and again a sharp drop at the cutoff. The wider window strongly accentuates size and direction of the estimated effects, as the yearly, log-transformed number of digital surrogates on the country-level increases to 1.7 in model (2) and stays statistically significant at the highest level. All results point towards a clear-cut positive effect and causal impact of a perceived public domain status on the availability of digitized artworks in the US. Our results continue to hold when we run the same estimates in model (4) on the artist-level. The effect of a perceived public domain status on availability (i.e. number of digitized artworks artists (log) observed) increases to 0.36 and becomes statistically significant. Finally, in the Appendix 6, we conduct additional robustness checks based on alternative placebo cut-off years, which strongly support the validity of our results and approximation of the copyright term via the date of artwork creation (cf. Figure 12).

5.4 Random assignment around the discontinuity and placebo tests

To further establish causal identification, we also visualize RDD effects for all covariates and observable characteristics we used to compute main results. The causal path is supported if, and only if, covariates do not show a discontinuity in their function. Hence, we plot each covariate in a separate graph as shown in figure 6, using a polynomial fit of order three and a triangular kernel function to weigh observations (Cattaneo et al., 2019). The distribution of likes (left) and page views (right) for artists in the top panels shows a heterogeneous pattern in terms of their popularity, but, notably, there is no discontinuity at the cutoff year in 1926. For museums (mid panels) and artworks (bottom panels), likes and page views show, if anything, a weakly increasing trend in popularity. So, at large, more recently created artworks seem more popular and they are also more often part of collections held by more popular museums. Despite these correlations, in none of the above panels do we observe a sharp discontinuity at the cutoff with regard to observable characteristics and the popularity of artworks, artists and museums, which alternatively could explain the drop in availability. This finding clearly supports our identification strategy.

As a next step, we provide a series of placebo tests by applying the same empirical strategy on digitized artworks for museums located in countries other than the U.S. We select and reduce the sample to artworks and museums located in France and the United Kingdom which do not see a change in copyright laws at the cutoff year in 1926. Copyright laws in France and the United Kingdom provide copyright protection for 70 years after the death of the original artist. At the time of

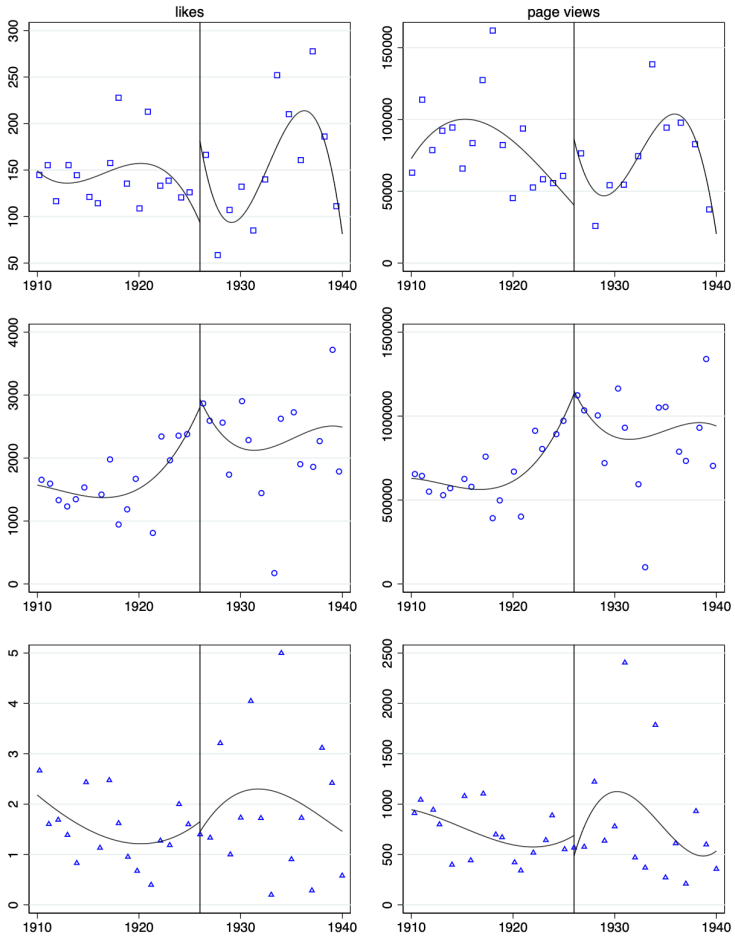

Figure 6: RDD: Covariates Artwork Characteristics

Note: This figure shows the characteristics of artists (top, square), museums (mid, circle) and artworks (bottom, triangle) of likes (left) and page-views(right) from U.S. museums. The time frame is of artworks creation years between 1910 and 1940. The x-axis line denotes the copyright cut-off point in 1926. The symbols represent sample average within bins and the lines a polynomial fit of order 3. Standard errors are clustered at the artist-level.

data collection (2021), artworks by artists that $died$ before 1951 were thus considered to be in the public domain in both countries. Placebo tests do not require that artworks created around the cut-off point in 1926 are all in public domain or protected under copyright. Rather, placebo tests might confirm that there is no discontinuity in the function around the cliff date, and we therefore should not observe a sudden jump in number of digitized artworks in museum collections located either in the United Kingdom or France.

Table 7 shows the results for the placebo timing test in the UK (left) and France (right). The models are constructed as before, with (2 and 4) and without (1 and 3) inserting covariates to specifications.

The dependent variables are yearly number of digitized artworks (log), and we stay on the country-level as we observe a relatively small number of observation in these samples. The time window is again tightened around the cutoff date (1910-1940). For the UK sample, we estimate an effect of 1.18 on the available digitized public domain artworks. Although the size of the effect is as large (or even larger) than the predicted effect using the U.S. sample, it renders statistically insignificant. Moreover, the result for the U.K builds on a relatively small sample of 137 artworks only (compared to the main results with a max of $n = 1605$). The opposite is true for artworks held by French museums, with estimates shown in models (3) and (4). When we add controls, there is a negative effect of minus 0.3 in the number of digitized artworks created before 1926 and becoming available online. Again, this effect is statistically insignificant, i.e. there is no statistically traceable discontinuity in the artwork creation-year function. So, placebo sample regressions further corroborate the validity and robustness of our empirical strategy.

Table 7: RDD: Availability (Artist-Level) (Placebo sample)

	UK sample		France sample	
	(1) No controls	(2) Controls	(3) No Controls	(4) Controls
Public Domain (US)	0.882	0.806	1.377	1.096
	(1.32)	(1.14)	(1.52)	(1.25)
Artwork creation year	-7.831	-11.90	-13.43	-17.75
	(-0.65)	(-0.81)	(-0.90)	(-1.15)
(Artwork creation year)^2	0.00203	0.00309	0.00347	0.00459
	(0.65)	(0.81)	(0.89)	(1.15)
museum views		-0.00000112		-0.000000943**
		(-0.51)		(-2.83)
artwork views		-0.0000496		0.0000693
		(-1.91)		(0.30)
N	281	280	384	360
R^2	0.014	0.021	0.107	0.122
Cluster SE	Artist	Artist	Artist	Artist
Artist FE	√	√	√	√

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: This table shows the RDD results for the placebo sample including museums from the United Kingdom (left) and France (right) with the threshold year as specified in chapter 4, and a artwork creation time-frame around 1910-1940. The dependent variable is constructed as the yearly number of available artworks on the artist-level. Artist fixed effects are included in all models.

Regression results are illustrated in figure 7. The left-hand panel shows results calculated based on the UK sample, the right-hand panel based on the French sample. Yearly sample average bins are plotted as grey dotes and we overlay a line of polynomial fit of order 3. The figure confirms estimated results from table 7 as we cannot detect a discontinuity at the 1926 cutoff upon visual inspection.

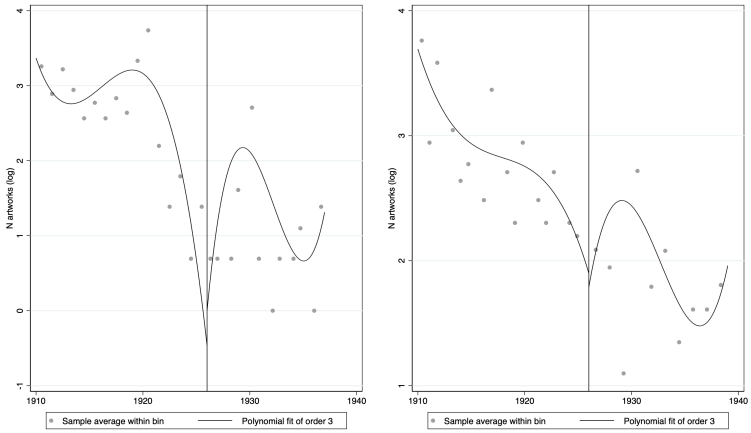

Figure 7: RDD: Digitized Artworks Placebo Sample

Note: This figure shows the yearly number of digitized artworks (log) and year of artwork creation date. The sample is restricted to years 1910-1940 and museums in the United Kingdom (left) and France (right). The dotes represent sample average within bin and the line a polynomial fit of order 3. Standard error are clustered at the artist-level.The x-axis line denotes the placebo cut-off point in 1926.

5.5 Differences-in-Differences

We present in this section the differences-in-differences regression results, calculated based on equations explained in chapter 4. The underlying idea is to construct a counterfactual trend of other museum and country available artworks compared to the discontinuity in perceived public domain status of artworks in U.S. museums. The interaction of $PublicDomain(US) \times US$ captures pre- and post-public domain years (again around artwork creation years 1910 to 1940) with a counterfactual trend of UK and EU museums and countries that do not have a sudden change in copyright status (calculated on the artwork creation year). The bench-marked jump measures by how much availability changes once artworks fall into public domain relative to other museums in the EU and the UK.

In table 8 we run the set of regressions on the artist-level (model 1-5), and include from model 1 on wards artist fixed effects, using (multi-level) fixed effects models. The dependent variable is constructed as the number of artworks of artist i, in year t and in country c. Our results continue to hold and stay robust and statistically significant at the 1%-level (except for model 1). The differences-in-differences coefficient is 2.059 (standard error 0.9) in model 2 and accentuates in model (5) using all fixed effects and control variables to 2.182 (standard error 0.9).

Finally, compared to previous sections, we address potential sample selection biases by balancing the panel on the artist-level. In order to calculate the difference-in-differences we balance the panel

Table 8: Differences-in-Differences: Availability (Artist-Level)

	DV: N artworks (artist)				
	(1)	(2)	(3)	(4)	(5)
Public Domain (US) $\times US$	1.460	2.059**	2.224**	2.155**	2.182**
	(0.950)	(0.988)	(0.962)	(0.911)	(0.913)
museum views					0.0000592
					(0.0000590)
artwork views					0.0000733
					(0.0000658)
N	4032	4032	4032	3956	3902
adj. R^2	0.567	0.635	0.648	0.656	0.657
Cluster SE	Artist	Artist	Artist	Artist	Artist
Artist FE	√	√	√	√	√
Year FE	No	√	√	√	√
Country FE	No	No	√	√	√
Museum FE	No	No	No	√	√

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: This table shows the differences-in-differences results, specified in section 4, for artworks created between 1910-1940 from U.S. and EU + UK museums.

around artists in the counterfactual trend (EU and UK) by limiting to artists we observe with at least one observation post 1926. We estimate in table 9 the same model specification as before, using the new artist-balanced panel. We note that the estimated coefficient is throughout statistically significant and positive regarding the upwards jumps of available artworks in U.S. museums compared to EU and UK museums around the 1926 cutoff. The effect accentuate from 2.96 (model 1) artworks per artist per year in U.S. museums to roughly 3.6 (in model 2-5) (compared to previous results in table 8 of around 1.46 to 2.22). To illustrate these results, we run an event study design, again using the restricted sample of -/+ 15 years to-/from public domain in the US, i.e. 1926= time treat 0, and to artworks in the U.S. (treated) and EU and UK (counterfactual) using the balanced-panel. In figure 8 we present the event study plot. Note that effect direction of this event study design in the figure is reversed given the normalization around 0 and coefficients of $post$ creation year 1926 compared to pre-years (previous results compared coefficients of public domain years, i.e. pre 1926 compared to the baseline post 1926 years). We observe common pre-1926 years availability of artworks in U.S. museums and artworks hosted in EU plus UK museums, with a sudden drop for non-public domain years in U.S. museums. The results are computed on multi-level fixed effects and with the (log) artists yearly number of digitized artworks as dependent variable. For artists in the US, we observe a clear-cut decrease in available digital artworks (log-transformed depended variable) of artworks not in public domain, compared to available artworks in the EU and UK around the same time-periods, with statistically significant negative coefficients for years following the public domain cut-off year. Results are also robust against using country specific artwork creation year trends (not reported).

Table 9: Differences-in-Differences: Availability (Artist-Level) (Balanced)

	DV: N artworks (artist)				
	(1)	(2)	(3)	(4)	(5)
Public Domain (US) $\times US$	2.962***	3.619***	3.601***	3.554***	3.594***
	(1.086)	(1.078)	(1.051)	(0.925)	(0.909)
museum views					0.000964***
					(0.000290)
artwork views					0.0000949
					(0.0000779)
N	1640	1640	1638	1582	1549
adj. R^2	0.469	0.633	0.638	0.648	0.650
Cluster SE	Artist	Artist	Artist	Artist	Artist
Artist FE	√	√	√	√	√
Year FE	No	√	√	√	√
Country FE	No	No	√	√	√
Museum FE	No	No	No	√	√

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: This table shows the differences-in-differences results, specified in section 4, for artworks created between 1910-1940 from U.S. and EU + UK museums, based on a balanced panel that only includes artists who had at least one observation in the counterfactual group post 1926.

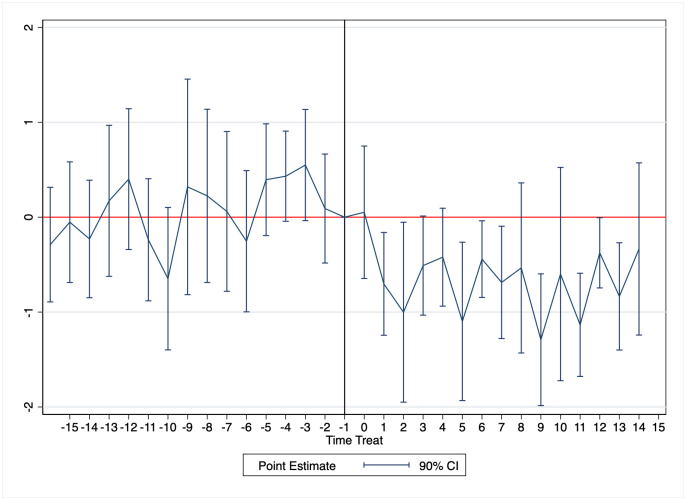

Figure 8: Dynamic Differences-in-Differences

Note: This figure shows the differences-in-differences event study plot with the yearly number of artworks (log) on the artist-level. The sample is restricted to -/+ 15 years around the year 1926. Time treat = 0 in year artwork creation year 1926. The sample is restricted to artworks in U.S. museums (treated) and EU and UK museums (counterfactual). The counterfactual trend is balanced around artists who had at least one observation post 1926. The coefficients are normalized around t-1. The event study based on a MWFE model with artist, creation year, museum, and country fixed effects and controls for artwork-, and museum-likes. N= 1'641.

5.6 External Validity

As an interim conclusion, we note that all our estimates point toward a clear-cut positive effect of artworks which recently fell into public domain vis-à-vis artworks with creation years subject to possible copyright protection. One major question is, to what extend these results generalize to global availability. We thus address the external validity of our results in a systematic matter. First, we man-

ually inspect the data set and compare to the available information directly accessed on the museum webpages, we contact selected museums worldwide, and we benchmark our data set with alternative data sources. Second, we hypothesize and discuss what other factors could potentially influence and restrict external validity. Third, we tentatively discuss and disentangle 'availability' from 'digitization' effects. Finally, section 3.2 provides additional evidence on image quality using an alternative source of data.

As preparatory work, we conduct extensive desk research on the available data sources around digital museum collections to better understand the representativeness of our data. Examples of sources we have considered are listed on Wikipedia, lists on Github covering U.S. museums (such as larger public data sets from MET or MoMA), and data collected by the Openartdata initiative. Initially, we also evaluate several alternative digital artwork sources such as Artstor (via JSTOR) and image licensing sources such as Gettyimages, or Scala Archives. These data sources, however, are biased towards pure commercial interests (in terms of artwork content and copyrights on the digital surrogates), or they focus on non-contemporary art (i.e. they contain historic art collections with artworks which are in the public domain for a long time). Taken together, all of the above data sources would not allow us to go beyond descriptive and case-study evidence. Moreover, they seem to suffer from severe bias as they do not provide for a representative sample of museums and do not well reflect museum efforts globally to digitize and make available artworks and images. In contrast, our *Useum* data sample is fairly comprehensive in terms of coverage and variety of museums and artists, even though certain data caveats apply. With the data at hand, we can only observe reuses of, arguably, more popular digital artwork surrogates that the *Useum* platform directly sources from museum websites, *Wikidata* repositories, or images crowd-sourced and uploaded to the platform by individual users. Accordingly, we cannot describe the entire universe and overall population of digitized artworks in museum collections. In other words, digital surrogates of, arguably, less popular original artworks might still exist online or offline, but are not made available on the platform. Similarly, the museum might consider alternative digital channels to make digitized artworks available. Eventually, it might also be the case that sometimes globally operating platforms and alternative channels such as *Wikimedia Commons* are challenged in many copyright jurisdictions to comply and interpret legal rules in a particular way, and policies around them may have changed over time [16]. To add yet another level of complexity, imagine an upload of a photograph that has been directly

[16] Some Wikidata policies to be found here or there.

taken from a copyrighted artwork hosted in a U.S. museum or from a book. This image could be considered elsewhere as in the public domain, or vice versa. Such digital artwork surrogates are often made available online, but they are not digitized and made available by the museum itself. Therefore, these images should not be considered for our research purposes.

In the appendix 6 and based on more extensive desk research, we furthermore provide and summarize information on all available and digitized artwork images by selected U.S. museums in the estimation sample (used to compute baseline results), some of which are images we cannot observe on the *Useum* platform. We manually inspect online collections and search GitHub repositories for selected museums, essentially, to better understand potential bias in our data sample, in particular as concerns the availability of digital artwork surrogates that are still under copyright protection. The exercise largely supports the general validity of our data. By comparing artworks in the estimation sample to the larger and arguably, more complete data universe from selected museum collections, we can show that our data displayed on *Useum* represent approximately 1 to 35 percent of the total online collections held by selected museums. Where our data only identifies very low numbers or no artwork images created after 1926, we are indeed able to replicate this finding directly through the museum website, as access to copyrighted works is often restricted, no images are available, there is no download option to the image (i.e. not shareable), or the image is displayed in thumbnails only (i.e. image resolution is very low). We are therefore confident that our data constitutes a representative sample for the online availability of artwork images hosted in U.S. museums.

Notably, however, while we can clearly identify the status effect on online availability, our approach can only approximate the effect of status changes on *digitization*, since we cannot observe the overall population of digitized artworks from museum collections with the present data. In other words, a digital surrogate of an original artwork might still exist somewhere online, but, for some reason, it does not become available on the *Useum* platform. In addition, any substantive claim on digitization effects would need to account for the selection of artworks into digitization. For example, such a comparison would require even richer information on the *entire* physical collection held by the museum and the dynamics and costs associated with the digitization process and artwork reproduction.

Finally, one could argue that the validity of our results is limited due to certain geo-blocking prac-

tices, given that IP-addresses from different countries might qualify for different digital access and territorial image availability. We tackle this question by inspecting the data on three different levels. First, geo-blocking seems not to be an issue for the Useum data and related results. Based on a random sample of artworks and IP-addresses hosted in different countries, we found no differences in image access and availability. Second, we are aware of some geo-blocking practices and potential bias regarding Google reverse image searches and their downstream reuses. In some cases, the search results were restricted or deviated when using different IP-addresses. While we cannot fully rule out this source of bias for downstream searches, this is less of a concern for our estimates as we further distinguish between upstream reuses using direct museum-URLs to compute results. Hence, the third level of potential geo-blocking and territorial availability bias could originate directly from the curation of museums. To address this issue, we directly reached out to selected museums and collected qualitative evidence.[17] Despite the fact that US-style fair use and the low-resolution/high-resolution distinction would not protect the U.S. museums from liability in the EU, none of the museums engaged in any geo-blocking of European digital visitors. This lack of concern about facing a negative judgment in an EU court apparently comes from reliance on the Second Circuit decision in *Sarl Louis Feraud, Int'l v. Viewfinder, Inc.* (2007) which refused to enforce a French copyright judgment on first amendment grounds (Guehenno, 2009). After *Sarl Louis*, museums, perhaps wrongly, appear unconcerned about potential liability for illegal distribution in the EU.

In sum, our estimates for the effect on right status on 'digital availability' should generalize well as we can build the core analysis on a representative sample of online artwork surrogates hosted by U.S. museums. At a minimum, the external validity is limited to online availability on the *Useum* service only (based on their own criteria of artwork selection), but this seems less likely. Based on the above discussion and additional verification of the sample using other sources, it seems we meaningfully capture the perceived public domain effect as applicable for all U.S. museums. In the next section, we provide additional context and shed further light on our findings by looking at image reuse and upstream quality of digital surrogates made available online, which adds another novel view on our data.

[17] In section 5.7 we describe the qualitative evidence and information collected from a number of U.S. museums, among others, the Cleveland Museum of Art, the Cincinnati Art Museum, the Museum of Fine Arts (Boston), the Guggenheim, and The Art Institute of Chicago.

5.7 Image Reuse and Upstream Quality

In this section, we first follow a descriptive and visual approach discussing the sample obtained from the 'google reverse image searches', as described in subsection 3.2 and in the Appendix 6, before we turn to estimation models.

5.8 Upstream and Downstream Reuses

Again, we start with a list of 1349 unique artworks from *Useum* and that hold a direct and functioning .JPG URL. This results in a total of 88'607 upstream and downstream reuses of digitized artworks (cf. table 2). For each reuse, we also observe detailed pixel $\times$ pixel information which gives us a good proxy for image quality. On the artwork-level, we find substantial online reuses for all artworks in our sample, with an average (median) reuse on 66 (42) webpages, 191 unique artworks created $\geq$ 1926 were found, and 1158 unique artworks created before 1926.[18] On the level of museums, our sample contains a variety of 45 museums in total, while 17 museums contribute to the results of artworks created after 1926.

With an average reuse rate on 66 pages, artworks show an impressive amount of reuses on the web (compared to the downstream reuse rate of Wikimedia images on 5.4 pages found by Erickson et al. (2018)). If we compare reuses by artwork creation year, we do not detect significant reuse differences (averages) for artworks created before or after 1926. Recall that we started with an unbalanced distribution of artwork creation years, and, as a result, more reuses should appear for older artworks by construction. Hence, a comparison based on an unbalanced sample might not be helpful. Similarly, at first sight, we cannot detect systematic differences in image quality for reuses of artworks around the cutoff year in 1926, based on their average quality (pixel $\times$ pixel). This can be explained after a closer look at the sample. The data reveals that 52 percent of the 'downstream' use is commercially driven (cf. the pie chart in figure 9 with the domain ending text-analysis) and it is noteworthy that these downstream uses often report misleading (image quality) information[19]. For instance, many commercial vendors, such as poster-sellers or interior design pages offer down-

[18] It is true that, by construction, our sample and initial source of data *Useum* is in itself a 'downstream' reuse of a digital image. Thus it appears artworks were reused at least once, and as a consequence should always appear in a reverse google image search.

[19] Which was the case for more than 80 percent of the commercial downstream uses, based on a random sample

stream reuses of digitized artworks in high-resolution. [20] This information, however, is misleading as the artwork itself (e.g. sold on a poster) is not shown in high resolution, while the product on sale shows the high-quality image. In other instances, misleading downstream image quality shows a high-resolution image that is protected by a watermark, or a high-resolution image scan (e.g. from a book), which can, strictly seen, not be considered a high-quality version of digitized artworks.

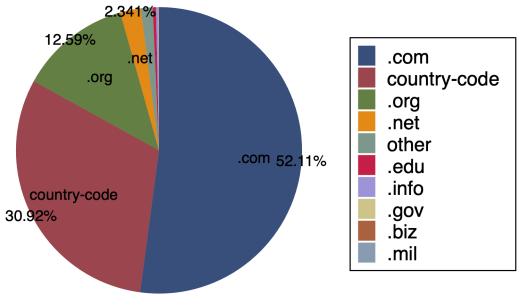

Figure 9: Number of Reuses by Artwork Creation Year

Note: This figure shows the percentage of google downstream use links found by a reverse google image search for U.S. artworks created between 1910 to 1940 by domain-addresses based on a link-text-analysis. Legend in descending order of occurrence. N=88'607.

So, upon visual inspection of the google search sample, if any, we observe a slightly higher reuse rate for younger artworks, where reused images are of equal resolution quality. Results are interesting per se, as digital artworks are highly reused, on a commercial and non-commercial basis, with a more detailed analysis being required to conclude on the overall downstream activities. But it is worth recalling, again, that we observe initially fewer artworks in our sample being created after 1926 than before.

5.9 Image Quality of Upstream Works

In order to better describe the image resolution quality, we now filter the google-image search results by using 'upstream reuses' only, which generates another useful subsample of the data, as described in subsection 3.2 and in the online Appendix 6. So here, again, we define an upstream use as artworks made available directly on the museum website. Put differently, we assume that the museum owns and digitizes the artwork, and therefore also controls and governs the initial image resolution (i.e. quality) of the digital surrogate.

[20] Examples for such commercial downstream activities, to name two out of many, are Redbubble or Alamy.

Interestingly, the 'Association of Art Museum Directors (AAMD) defines/ recommends in their 'Guidelines for the use of copyrighted materials and works of art by art museums' what could be considered as low-/ high-resolution images in terms of fair use: "While the one-quarter screen and 560 x 843 pixels dimensions should be well within accepted norms of fair use for online collections, the application of the law of fair use to digital images as well as technology itself is constantly evolving" (AAMD (2017), p.12). Based on this information we define a threshold line of high-/ low-quality images below or above the pixel $\times$ pixel quality line of 472'080. This allows us to combine and research on availability *and* quality of upstream works.

We now visually inspect the upstream reuse sample and the detailed information on pixel $\times$ pixel image quality. Figure 10 plots the image resolution of reusing digitized artworks and the creation year of original artworks, with an horizontal red line indicating the threshold line for high resolution images (pixel × pixel $\geq$ 472'080) and the vertical red line is plotted at the year 1926. Since average quality is driven by a few outliers with a very high image resolution, we exclude a total of 80 upstream images in this figure. The descriptive evidence now shows a clear pattern and distribution for image resolution and original artwork creation years. While perceived public domain artworks show overall a broader (i.e. higher) distribution of image quality, the clear cut line of low-/high-resolution images is more clearly identified for digital artwork surrogates and reuses, where the original artwork is still protected under copyright. In total, only around 32 images on the right-hand side of the cut are available in high-quality in the upstream sample, and at the same time, many museums seem to make these artworks available in low-quality only. We interpret this finding as evidence for a fair use practice of museums for making artworks available that remain copyright protected.

This interpretation of results finds further support from qualitative research. We contacted several larger museums in the US, namely, the Cleveland Museum of Art, the Cincinnati Art Museum, the Museum of Fine Arts (Boston), the Guggenheim, and The Art Institute of Chicago. They all report to rely heavily on the U.S. fair use doctrine in making a distinction between the uploading of low-resolution thumbnails without permission and the making available of high resolution images. For example, the Cincinnati Art Museum stated that it will sometimes contact a copyright owner, but it was "often able to reproduce images for educational purposes without seeking permission," while The Art Institute of Chicago "relies heavily on fair use." The Cleveland Museum of Art digitizes ev-

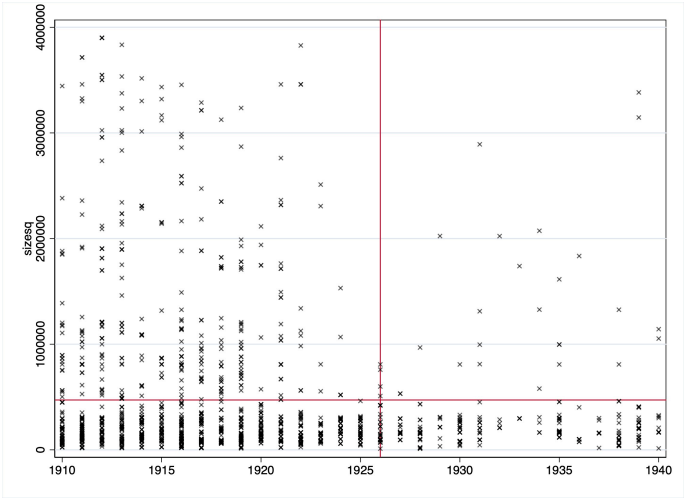

Figure 10: Upstream Use and Image Resolution

Note: This figure shows the distribution of museum upstream use links found by a reverse google image search for U.S. artworks created between 1910 to 1940. A cross represents an upstream image (opacity set at = 60 %). The horizontal red line indicates the threshold line for low-/high resolution images (pixel $\times$ pixel $\geq 472'080$) and the vertical red line is plotted at 1926.5. The figure is restricted to observations with image resolution $<$ 4'000'000. $N = 1'709$.

erything without permission and makes its entire digitized collection available in low resolution form without seeking permissions. Moreover, the Museum of Fine Arts in Boston provided the foundation for the distinction made between high and low resolution images, explaining that it "releases some thumbnail-size images of copyrighted works under fair use, in accordance with the guidelines recommended by the Code of Best Practices in Fair Use for the Visual Arts, published by the College Art Association in February 2015." This Code of Best Practices draws the same line between low resolution (no permission needed) and high resolution images (permission needed). None of the museums contacted would upload high resolution protected images without permission of the copyright owner. Interestingly, when asked, none expressed concern that online access would result in reduced traffic and fewer visitors to their facilities. Accordingly, it seems that cannibalization of sales through digital channels and the potential negative effect on commercial activities of the museum was not a consideration.

Next, we turn to estimations and address the econometric question whether perceived public domain works are generally available in higher image resolution than digital surrogates of artworks still under copyright. In table 10, we regress a public domain dummy variable (i.e. one if the artwork is created before 1926, zero otherwise) on image quality as measured by image resolution, using the upstream sample. Model 1 reports our baseline, model 2 introduces year fixed effects and clustered

standard errors at the museum-level (our relevant level of upstream analysis), model 3 adds museum fixed effects and control variables, model 4 clusters standard errors at the artist creation-year, and model 5 reports estimates using a log-transformed outcome variable that is robust to outliers in the sample. The coefficient of interest $PublicDomain$ shows a positive sign in all model specification, and introducing clustered standard errors (at the museum-level) makes the coefficient statistically significant (model 2), and stays robust (but insignificant) also for the most demanding multi-level fixed effects model (3). Clustered standard errors at the artwork creation date (model 4) provides in the same model specification significant positive effects. Using the log-transformed quality outcome in model 5, we lose the statistically significance, but the direction of the effect does not change. We conclude that, at the minimum, we have weak evidence that perceived public domain works are made available in higher pixel resolution by museums (i.e. upstream users) than digital surrogates of artworks still protected under copyright.

Table 10: Upstream Use and Image Resolution

	Quality				log(Quality)
	(1)	(2)	(3)	(4)	(5)
	No Controls	Year FE	All		All
Public Domain	279892.1	403952.0*	312118.1	312118.1***	0.122
	(1.37)	(2.25)	(1.48)	(4.38)	(0.28)
artwork views			35.04	35.04	0.0000751***
			(1.70)	(0.95)	(5.59)
artist views			-1.950***	-1.950**	-0.000000631*
			(-3.86)	(-2.81)	(-2.37)
museum views			0.0160	0.0160	5.27e-08***
			(0.67)	(0.54)	(4.24)
N	1734	1734	1734	1734	1734
r2	0.00108	0.0225	0.582	0.582	0.324
Year FE	No	√	√	√	√
Museum FE	No	No	√	√	√
Cluster SE	No	Museum	Museum	Artwork C Date	Museum

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: This table shows the upstream use estimates on image resolution (pixel $\times$ pixel) from U.S. museums between 1910-1940. The public domain dummy denotes artworks created before 1926.

6 Summary

In this paper, based on a novel and rich dataset of digitized museum collections, we provide quantitative evidence for a causal relationship between the online availability of digital images and the right status of the underlying artwork using a regression discontinuity design. To the best of our knowledge, this research is among the first studies that provides quantitative evidence in this area. We exploit the recent extension of copyright terms in the U.S. as a quasi-natural experiment to show

that, on average, perceived public domain status of an artwork in the U.S. increases image availability online by 34 to 42 percent in artworks by a given artist in the sample. This translates to an effective increase by 1 to 1.4 artworks made available online. Our results on museums largely confirm findings on public domain effects from previous research on music and books.

Core findings are robust against a series of placebo tests, random assignment around the discontinuity, and using an alternative differences-in-differences design. While, arguably, our results might not generalize to universal availability of artwork images on the internet, they, at minimum, continue to hold and are relevant for the increase in availability observed on Useum, Wikimedia and similar large digital platforms (for popular artists and artworks).

We can further show that digital images made available online see a large total number of reuses upstream and downstream which indicates their availability is of high public and private value, independent of the status of rights. Moreover, we are first to also provide empirical evidence on the average quality of images (resolution) made available upstream by museums (as part of their digitized collections) and depending on the right status. Here, findings suggest that industry practices and expressed norms might help to also make images available in lower resolution when artworks are still protected under copyright. We hope our results will help inform the current debate on copyrights and the preservation of cultural heritage in the digital age.

References

AAMD (2017). Guidelines for the use of copyrighted materials and works of art by art museums, url: https://aamd.org/sites/default/files/document/guidelines.

American Alliance of Museums (Accessed: 2023-06-26). Tech tutorial digital copyright and privacy. `https://www.aam-us.org/wp-content/uploads/2018/01/digital-copyright-and-privacy.pdf`.

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly harmless econometrics*. Princeton university press.

Bakhshi, H. and Throsby, D. (2014). Digital complements or substitutes? a quasi-field experiment from the royal national theatre. *Journal of Cultural Economics*, 38(1):1–8.

Biasi, B. and Moser, P. (2021). Effects of copyrights on science: Evidence from the wwii book republication program. *American Economic Journal: Microeconomics*, 13(4):218–60.

Bitton, M. (2011). Modernizing copyright law. *Tex. Intell. Prop. LJ*, 20:65.

Borowiecki, K. J. and Navarrete, T. (2017). Digitization of heritage collections as indicator of innovation. *Economics of Innovation and New Technology*, 26(3):227–246.

Buccafusco, C. and Heald, P. J. (2013). Do bad things happen when works enter the public domain?: Empirical tests of copyright term extension. *Berkeley Technology Law Journal*, pages 1–43.

Calabresi, G. (1960). Some thoughts on risk distribution and the law of torts. *Yale Lj*, 70:499.

Cattaneo, M. D., Idrobo, N., and Titiunik, R. (2019). *A practical introduction to regression discontinuity designs: Foundations*. Cambridge University Press.

Coase, R. H. (1960). The problem of social cost. *Journal Law and Economic)*, 3:144.

Correia, S., Guimarães, P., and Zylkin, T. (2020). Fast poisson estimation with high-dimensional fixed effects. *The Stata Journal*, 20(1):95–115.

Depoorter, B. and Parisi, F. (2002). Fair use and copyright protection: a price theory explanation. *International Review of Law and Economics*, 21(4):453–473. 17th Annual Conference of the European Association of Law Economics, Ghent, Belgium, September 2000.

Erickson, K., Perez, F. R., and Perez, J. R. (2018). What is the commons worth? estimating the value of wikimedia imagery by observing downstream use. In *Proceedings of the 14th International Symposium on Open Collaboration*, pages 1–6.

Fernandez-Blanco, V. and Prieto-Rodríguez, J. (2020). Museums. In *Handbook of Cultural Economics, Third Edition*. Edward Elgar Publishing.

Flynn, J., Giblin, R., and Petitjean, F. (2019). What happens when books enter the public domain?: Testing copyright's under use hypothesis across australia, new zealand, the united states and canada. *University of New South Wales Law Journal, The*, 42(4):1215–1253.

Gerhardt, D. R. (2011). Copyright publication: An empirical study. *Notre Dame L. Rev.*, 87:135.

Ginsburg, J. C., Gordon, W. J., Miller, A. R., and Patry, W. F. (2000). The constitutionality of copyright term extension: How long is too long? *Cardozo Arts & Ent. LJ*, 18:651–701.

Greenstein, S., Goldfarb, A., and Tucker, C., editors (2013). *The Economics of Digitization*. Edward Elgar Publishing.

Guehenno, C. (2009). Recent development online. *FORDHAM L. REV*, 2537:2546.

Heald, P., Erickson, K., and Kretschmer, M. (2015). The valuation of unprotected works: a case study of public domain images on wikipedia. *Harv. JL & Tech.*, 29:1.

Heald, P. J. (2007). Property rights and the efficient exploitation of copyrighted works: An empirical analysis of public domain and copyrighted fiction bestsellers. *Minn. L. Rev.*, 92:1031.

Heald, P. J. (2009). Does the song remain the same-an empirical study of bestselling musical compositions (1913-1932) and their use in cinema (1968-2007). *Case W. Res. L. Rev.*, 60:1.

Heald, P. J. (2014). How copyright keeps works disappeared. *Journal of Empirical Legal Studies*, 11(4):829–866.

Heald, P. J. (2020). The impact of implementing a 25-year reversion/termination right in canada. *J. Intell. Prop. L.*, 28:63.

Heald, P. J., Shi, P., Stoiber, J., and Zheng, Q. (2012). More music in movies: what box office data reveals about the availability of public domain songs in movies from 1968-2008. *Review of Economic Research on Copyright Issues*, 9(2):31–54.

Imbens, G. W. and Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of econometrics*, 142(2):615–635.

Landes, W. M. and Posner, R. A. (2003). Indefinitely renewable copyright. *U. Chi. l. Rev.*, 70:471.

Landes, W. M., Posner, R. A., et al. (2003). *The economic structure of intellectual property law*. Harvard university press.

Lee, D. S. and Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of economic literature*, 48(2):281–355.

Li, X., MacGarvie, M., and Moser, P. (2018). Dead poets' property—how does copyright influence price? *The RAND Journal of Economics*, 49(1):181–205.

Marciano, A. (2012). Guido calabresi's economic analysis of law, coase and the coase theorem. *International Review of Law and Economics*, 32(1):110–118. The Economics of Efficiency and the Judicial System.

Nagaraj, A. (2018). Does copyright affect reuse? evidence from google books and wikipedia. *Management Science*, 64(7):3091–3107.

Posner, R. A. (1972). Economic analysis of law (boston. *MA: Little*.

Reimers, I. (2019). Copyright and generic entry in book publishing. *American Economic Journal: Microeconomics*, 11(3):257–84.

Useum (2022). Useum webpage. url: https://useum.org/content/about-useummission. retrieved 25.04.22.

Wallace, A. (2022). A culture of copyright.

Watson, J. (2017). Copyright and the production of hip-hop music. Technical report, Working Paper.

Zemer, L. (2006). The making of a new copyright lockean. *Harv. JL & Pub. Pol'y*, 29:891.

Appendix

Additional regression results

Table 11 shows the results of the regression discontinuity estimates using the dependent variable on the country-level (log). The table reports 6 different models. In the first model, we present a RDD regression that estimates the effect of artworks created before/ after the threshold year on the number of digital available artworks in a given year. The second column introduces covariates (2) and clustered standard errors at the museum-levels (3). Model (4) estimates the RDD model with artist fixed-effects, model (5) with museum fixed-effects and model (6) a multi-way fixed-effects model respectively.

The effect of perceived public domain status is statistically significant and positive throughout RDD models. The size of the effect ranges from 0.751 to 0.819. The models suggest a significant positive impact on the digital availability of artworks hosted on the platform when works are likely to fall into public domain. This applies for artworks in U.S. museums created before 1926. Model (1) is based on RDD model and shows a positive effect of around 0.814 for the coefficient of interest. This coefficient is a dummy variable that is one if the artwork was created before 1926, i.e.

'$PublicDomain$' holds, zero otherwise. The same coefficient stays robust and highly statistically significant at the 99%-level in Model (2), once we include the covariates that aim to control for artwork popularity. Coefficient estimates are robust to the inclusion museum-level clustered SEs, artist-, and museum fixed-effects. While the very aggregated level of the dependent variable hints towards a more refined econometric specification, we can interpret the results as an aggregated and positive correlation of public domain artworks available.

Table 11: RDD: Availability (Country-Level)

	DV: log(N artworks Country)					
	(1) RDD	(2) RDD Covs	(3) RDD	(4) RDD FE	(5) RDD FE	(6) RDD MWFE
Public Domain	0.814***	0.819***	0.819***	0.776***	0.792***	0.751***
	(28.47)	(28.13)	(30.32)	(25.48)	(32.69)	(22.05)
Artwork creation year	-2.904***	-2.889***	-2.889***	-2.572***	-2.766***	-2.317***
	(-6.46)	(-6.34)	(-7.12)	(-3.89)	(-6.19)	(-4.91)
(Artwork creation year)^2	0.000742***	0.000738***	0.000738***	0.000655***	0.000706***	0.000589***
	(6.34)	(6.22)	(7.00)	(3.80)	(6.07)	(4.79)
artist views		1.01e-08	1.01e-08		8.54e-09	
		(0.18)	(0.24)		(0.24)	
museum views		6.75e-09	6.75e-09	1.27e-08	-0.00000190***	0.00000123
		(1.08)	(0.99)	(1.93)	(-10.33)	(1.07)
artwork views		-0.00000141	-0.00000141	-0.000000717	-0.00000118	-0.00000185
		(-0.33)	(-0.53)	(-0.13)	(-0.44)	(-0.78)
N	1605	1566	1566	1313	1539	1290
R^2	0.914	0.914	0.914	0.924	0.919	0.928
Cluster SE	Artist	Artist	Museum	Artist	Museum	Museum
Artist FE	No	No	No	√	No	√
Museum FE	No	No	No	No	√	√

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: This table shows the RDD results based on the formula specified in chapter 4 and a DV as the year-country number of digitized artworks (log), restricted to museums in the United States and artworks created between 1910 to 1940.

Figure 11 visualizes the results from the corresponding model specification in table 11. It shows a sharp discontinuity in the availability of digital artworks around the cliff year in 1926, with the yearly sample average bins illustrated as pink crosses and plotting a black line with polynomial fit of order 3. Upon visual inspection, figure 11 largely supports our identification strategy.

As our dependent variable from the previous results is a counter, we adopt furthermore in table 12 Poisson (1-3) and a Poisson pseudo-maximum likelihood model (4-6) with multi-way fixed effects (model 6) (Correia et al., 2020). We continue to estimate throughout significant positive effects of public domain years in otherwise the same model specifications. Including artist-fixed effects halves the coefficient to around 0.85 (model 3) to 0.445 in model (4) on-wards but stays robust against including museum-fixed-effects in model (5-6).

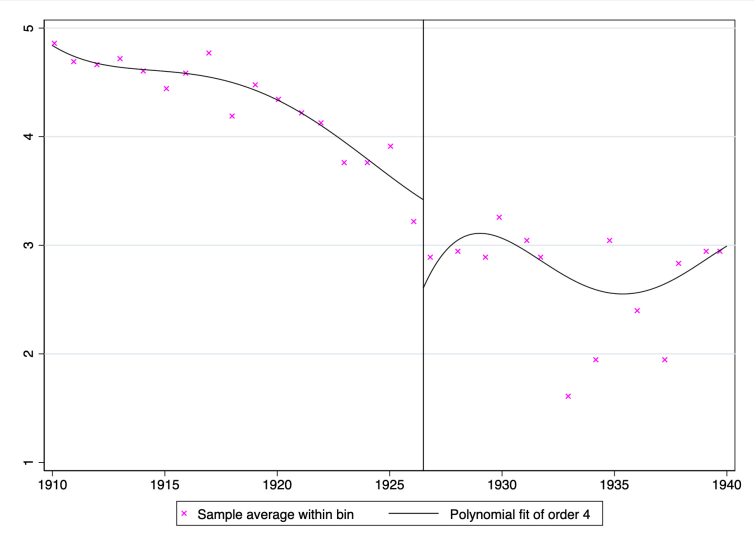

Figure 11: RDD: Digitized Artworks (log) (Country-Level)

Note: This figure shows the yearly number of digitized artworks (log) and year of artwork creation date. The sample is restricted to years 1910-1940 and museums in the United States. The crosses represent sample average within bin and the line a polynomial fit of order 4. Standard error are clustered at the artist-level. The x-axis line denotes the copyright cut-off point in 1926.5.

Table 12: RDD: Availability (Artist-Level) (poisson)

	DV: N artworks Artist					
	(1) Poisson	(2) Poisson Covs	(3) Poisson FE	(4) Poisson FE	(5) Poisson FE	(6) Poisson MLFE
Public Domain	0.797**	0.847**	0.851**	0.446**	0.455**	0.424**
	(2.76)	(3.16)	(3.21)	(2.86)	(2.92)	(2.71)
Artwork creation year	0.00818	0.0130	0.0172	2.886	2.848	2.777
	(0.63)	(1.03)	(1.36)	(1.49)	(1.47)	(1.56)
(Artwork creation year)^2				-0.000752	-0.000742	-0.000724
				(-1.50)	(-1.48)	(-1.57)
artist views		0.00000220***	0.00000196***			
		(5.01)	(4.05)			
museum views		9.80e-08	0.0000158		2.85e-08	
		(1.59)	(0.91)		(1.26)	
artwork views		0.0000243	0.0000260		0.00000952	0.00000248
		(1.31)	(1.91)		(1.04)	(0.28)
N	1592	1566	1566	1335	1313	1290
Cluster SE	Artist	Artist	Artist	Artist	Artist	Artist
Pseudo R^2	0.0258	0.1157	0.1724	0.3052	0.3049	0.3172
Artist FE	No	No	No	√	√	√
Museum FE	No	No	√	No	No	√

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: This table shows the RDD results based on the formula specified in chapter 4 and a DV as the year-artist-level number of digitized artworks in model 1-6, restricted to museums in the United States and artworks created between 1910 to 1940. Model 4-6 are calculated based on artist-panel fixed effects, and model 6 multi-way fixed effects model.

Moreover, in table 13 we present additional differences-in-differences regressions based on a straight forward setting of yearly available artworks in the U.S. compared to the EU and UK museums,

pre and post 1926 (i.e. perceived public domain status in the U.S.). The interaction term continues to replicate our positive public domain coefficient from the previous results, with a yearly upwards jump of around plus 50-54 available artworks of U.S. museums in post 1926 years. We control in this table again for year-fixed-effects (model 2), add country-fixed-effects (model 3) or combine multiway-fixed-effects including museum-fixed-effects in model (4).

Table 13: Differences-in-Differences: Availability (Country-Level)

	DV: N artworks (country)				
	(1)	(2)	(3)	(4)	(5)
Public Domain (U.S.) $\times U.S.$	49.80***	50.30***	54.35***	53.44***	53.48***
	(4.918)	(4.597)	(4.181)	(4.149)	(4.238)
artist views					0.00000219
					(0.00000356)
museum views					0.000170
					(0.000213)
artwork views					0.000196*
					(0.000111)
N	4762	4762	4762	4762	4660
adj. R^2	0.661	0.780	0.883	0.887	0.886
Cluster SE	Country	Country	Country	Country	Country
Year FE	No	√	√	√	√
Country FE	No	No	√	√	√
Museum FE	No	No	No	√	√

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: This table shows the differences-in-differences results, specified in section 4, for artworks created between 1910-1940 from U.S. and EU + UK museums.

Finally, in the empirical framework section, we discuss potential issues relevant to the approximation of the copyright status of the underlying artwork and conclude that by looking at artwork creation years, we are only able to approximate the public domain status (instead of artwork publication year). In additional robustness checks, we re-run the main estimations based on artworks from U.S. museums, but now choose alternative placebo cut-off *years*. Our results continue to hold if (and only if) a significant jump only appears in 1926, while estimations the hypothetical, alternative cut-off years remain statistically insignificant. As shown in Figure 12, this is indeed the case. In both specifications (*placebo 1921* with a different time-frame (left) and *placebo 1925* with the same time-frame (right)), we estimate a clear continuous function of (log) digitized artworks and creation years. Based on these results, we conclude that our approximated public domain year is not biased against a generic status 'misclassification' problem, and museums (and platforms) are very likely to consider the year 1926 as the relevant approximation year.

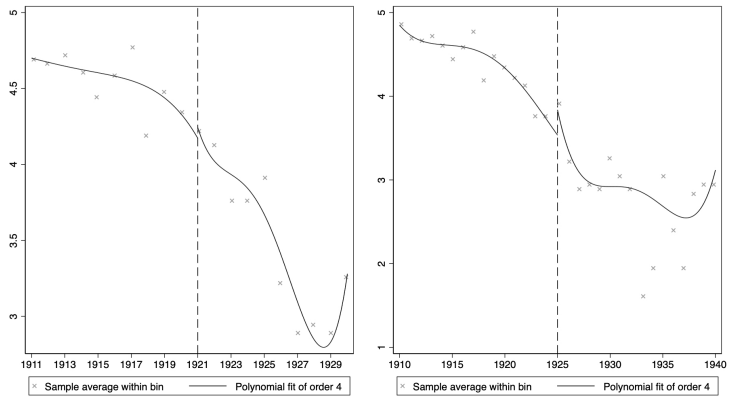

Figure 12: RDD: Digitized Artworks (log) Placebo Years

Note: This figure shows the (log) yearly number of digitized artworks on the U.S.-level and year of artwork creation date. The sample is restricted to years 1910-1930 (left) and 1910 to 1940 (right) and museums in the United States. The placebo cut-off year is set at 1921 (left) and 1925 (right). The crosses represent sample average within bin and the line a polynomial fit of order 4. Standard error are clustered at the artist-level.

Useum

We briefly explain the web-scraping process hereafter that was built in order to obtain data from the provider Useum. We first created 'element scroll down' and 'parallel link' selectors in order to address the thousands of museums' websites with a delay of 2 seconds each. In a next step, each of these links were processed and museum-level meta-information (country, followers ...) were stored in a 'csv' file. Subsequently, a second 'element scroll down' selector collected all links of the museums' artworks. Similarly, each of these links contained artwork-level information (artist, title, creation year...). The selectors of museum-artwork-links are shown in figure 13 to illustrate the idea of the scraping logic.

The final baseline scraped data set contained three levels of information beginning with the web-scraperstarturl and artworklinkhref; on the museum-level: museum_name; museum_country; museum_views; museum_likes; museum_follower; museum_about museum_url; on the artist-level: artist_link; artist_linkhref; artist_country; artist_bio; artist_views; artist_likes; and on the artwork-level: artwork_name; artwork_date; artist_name; artwork_description; artwork_likes; artwork_views; artwork_license; artwork_link; artwork_download; artwork_buy.

6.1 Artnet

More than 340 thousand artists are listed on Artnet. Here, we extract only information relevant to the artist'-level to complement our main data set. The data allow for the construction of more com-

Table 14: Country Table (overall)

Country	N Artworks	Country	N Artworks
United States	27,934	Greece	145
Netherlands	27,050	Taiwan	95
United Kingdom	17,433	Vatican City	85
Germany	11,970	Serbia	64
France	8,457	Belarus	62
Spain	5,325	Romania	50
Austria	4,462	Ukraine	49
Poland	4,100	Peru	44
Denmark	3,999	Latvia	35
Russia	3,573	Cuba	33
Norway	2,782	China	30
Italy	2,121	Slovenia	20
Belgium	1,766	Azerbaijan	19
Sweden	1,291	South Africa	17
Ireland	1,063	Israel	13
Australia	689	Bulgaria	12
Canada	595	Egypt	10
Argentina	567	Malta	9
Finland	526	Iran	8
Brazil	501	Lithuania	8
Portugal	495	Colombia	5
New Zealand	490	Philippines	5
Hungary	435	Croatia	3
Japan	373	India	3
Estonia	321	Monaco	3
Switzerland	311	South Korea	3
Slovakia	247	Costa Rica	1
Mexico	207	Liechtenstein	1
Czech Republic	156	Malaysia	1
Armenia	151		

Note: This table shows the overall distribution of digitized artworks sorted by countries.

Table 15: Top 20 Venue and Artist Table (overall)

Top Museum Venue	N	Top Artist	N
jksmuseum.nl/en	17,731	Peter Paul Rubens	692
pinakothek.de/en	7,749	Edvard Munch	657
metmuseum.org	6,188	Anthony van Dyck	605
nationaltrust.org.uk	5,240	Auguste Renoir	543
rct.uk	3,513	Claude Monet	517
smk.dk	3,255	Vincent van Gogh	510
museodelprado.es/en	3,019	Rembrandt	503
mnw.art.pl/en	2,934	Jean-Baptiste-Camille Corot	458
culturalheritageagency.nl/en/cultural-h	2,759	James Tissot	455
hermitagemuseum.org	2,641	Leo Gestel	430
belvedere.at	2,247	Paul Cézanne	403
khm.at/en	1,976	David Teniers the Younger	400
npg.org.uk	1,864	Camille Pissarro	363
nasjonalmuseet.no/en/	1,844	John Singer Sargent	325
tate.org.uk/visit/tate-britain	1,836	Jan Ciągliński	324
dia.org	1,746	Jozef Israëls	323
mfa.org	1,741	J M W Turner	294
amsterdammuseum.nl/en	1,739	Lucas Cranach the Elder	287
britishart.yale.edu	1,727	Jan van Goyen	269
louvre.fr/en	1,380	Jan Brandes	264

Note: This table shows the overall top-20 museum venues (link) and artist names.

prehensive information on artists' nationality and year of death as, in some cases, this information was not available from the original data source. If available, we match the list of artists from Artnet

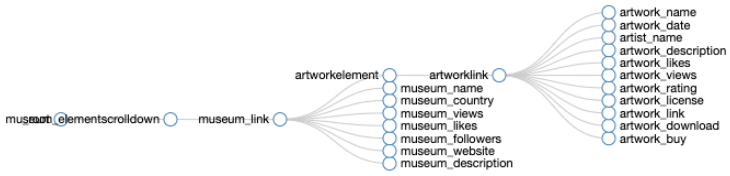

Figure 13: Webscraper Graph

Note: This figure shows the graph of the scraping-process of the webpage Useum (starting from August 18. 2021).

Table 16: External Validity of Data: Manual inspection for selected museums and GitHub repositories by U.S. museums, creation year 1910-1940

Museum	**Observed Data on Useum**	**Information for selected museums websites and GitHub repositories**	**≈ %**
metmuseum.org	358 artworks, 27% post 1926	Filter by 1900-present on museum website: 4.773 results (paintings), with 2.9% open access (139 artworks). Their collection data on GitHub brings up a total of 1,021 paintings for the exact period 1910 to 1940, with 43% created after 1926. For 56% of the total paintings in the overall period the data also provides for a wiki data url. Only 10(2) paintings in the pre(post) 1926 period are classified as perceived public domain.	35%
imamuseum.org	117 artworks, 30% post 1926	1088 objects 20th CE (all objects, including e.g. textile art) on the museum website. Sample of post 1926: "This image is not available online larger than a thumbnail to protect the copyright of its creator(s)". Their collection data on GitHub brings up a total of 364 paintings for the exact period 1910 to 1940, with 39% created after 1926. For 55% of the total paintings an image url is provided. Only 77(9) paintings in the pre(post) 1926 period are classified as perceived public domain, all other relate rights to artists and artist estates.	32%
barnesfoundation .org	87 artworks, 0 post 1926	358 Results paintings found between 1910-1940. 16% post 1926. 3 in perceived public domain, all other copyright restricted access, not downloadable or not shareable. Mid-image resolution.	18%
guggenheim.org	81 artworks, 23% post 1926	For artworks from the 1920ies, 115 artworks are available. Low resolution, potentially downloadable.	nA
mfa.org	72 artworks, 3% post 1926	466 artworks online between 1910-1940. Download option also for post 1926 artworks possible in low-resolution. High resolution request $50.	15%
nga.gov	61 artworks, 10% post 1926	452 paintings between 1910-1940 (408 with images). Of which 198 post 1926, 3 paintings with download option.	13%
moma.org	60 artworks, 20% post 1926	597 (527) paintings between 1910-1940 listed on museum website (GitHub). 331 (283) post 1926 on museum website (GitHub). Mid-level resolution, no download option on museum website. 91% of total period with thumbnail url on GitHub, 92% for post 1926.	11%
slam.org	59 artworks, 15% post 1926	170 paintings 1900-present. Potential download option in low-resolution, also for post-1926 artworks.	min. 34%
artgallery.yale .edu	51 artworks, 6% post 1926	2740 paintings 20th century. Post 1926 paintings available. Selected sample all thumbnails only.	nA
artic.edu	44 artworks, 20% post 1926	423 paintings between 1910-1940. 28 paintings are in perceived public domain.	10%
harvardart museums.org	43 artworks, 0% post 1926	2087 paintings 20th century. All post 1926 creation date samples non-downloadable thumbnails or "no image available".	nA
dma.org	36 artworks, 3% post 1926	3228 objects on museum website (all objects 20th CE). All samples of post 1926 objects in low-resolution and no download option. Their collection data on GitHub brings up a total of 171 paintings for the exact period 1910 to 1940, with 60% created after 1926. For 96% of the total paintings an thumbnail image url is provided. Only 40(13) paintings in the pre(post) 1926 period are classified as perceived public domain, 3(11) are held under non-exclusive license, all other relate rights to artists and artist estates.	21%
npg.si.edu	34 artworks, none post 1926	Their collection data on GitHub brings up a total of 239 paintings for the exact period 1910 to 1940, with 50% created after 1926. For 96% of the total paintings an image url is provided, some of which are thumbnail or low resolution only.	14%
crystalbridges .org	19 artworks, none post 1926	292 paintings on museum website (1910-present). Low-mid image resolution. Their collection data on GitHub brings up a total of 62 paintings for the exact period 1910 to 1940, with 58% created after 1926.	30%
...	...	...	...

based on a non-fuzzy text-matching process (i.e. we only used the exact name and surname match) and identify 86'519 artists' nationalities and 72'709 death years from the unbalanced *Useum* data set

External Validity of Data: Manual inspection for selected museums and GitHub repositories by U.S. museums, creation year 1910-1940 (continued)

Museum	Observed Data on Useum	Information for selected museums websites and GitHub repositories	≈ %
...	...	...	...
clevelandart.org	19 artworks, 10% post 1926	234 paintings between 1910-1940 as listed in GitHub collection, 45% are post 1926. 28(2) paintings from the pre(post) 1926 periods are available under a CC0 license and provide for an image/jpg url. Of those created in the pre(post) 1926 periods, classified as under copyright and that hold a data entry on right ownership, 17(34)% refer to the Artists Right Society (ARS) as the owner/representative, all others refer to artist estates. 72(69)% of copyrighted works for each period do not have an ownership entry on GitHub.	8%
americanart.si.edu	8 artworks, 25% post 1926	Unknown # artworks on website. Sample post 1926 available in high-resolution, download option. Non-commercial reuse permitted, up to $200 fees for other types of use here. Their collection data on GitHub brings up a total of 1,303 paintings for the exact period 1910 to 1940, with 71% created after 1926. For 96% of the total paintings an image url is provided.	0-1%

Note: The summary table lists notes on observed Useum data and manual inspection of selected U.S. museum online collections and GitHub repositories (e.g. https://github.com/american-art/), searched for 'paintings' and artwork creation years 1910-1940, where applicable. Last column indicates the ≈ percentage of the online collection we observe. Last website accessed: September 2022.

($n = 130'223$).

6.2 Google reverse image search

In this subsection, we explain the data collection process for the 'reverse google image search'. This required several steps, each of it with some error-potential for search results. First of all, in order to search for google-images one need a detailed URL link to the image (i.e. the .JPG URL and not the corresponding Wikimedia-page). Based on the artwork-links listed on *Useum*, we therefore constructed an automated web-scraper in *Python* and the selenium web-driver that opened each of the artwork image webpages and searched for the '.JPG' link of the artwork. We also manually checked for the quality of the search results by inspecting the .JPG image and if it corresponded to the URL page. This was of high quality for the Wikipedia pages (where 93 percent of original link-source originated from). Unfortunately, as some of the remaining webpages (e.g. a museum webpage) also cross-linked other (JPG) images on the artwork-page, we manually curated the .JPG ending to the corresponding artwork of interest. In some cases, the browser automated framework opened an old, broken or non-correct artwork-link. In most of the cases, this lead to an 'error'- link and we excluded the results from this query.

Next, we made use of the Google image search (by URL) function which allows users to find related images. For our research purpose, we only focused on 'Pages that include matching images', and not 'Visually similar images' results, see for an example this search query. More information of a reverse

image search can be found here or there. This step was automated with a cloud-based webscraper, emulating a human search, that extracted each of the search result domains. In order to not trigger a DDOS attack pattern, one can randomly change time intervals between the search queries, and the webscraper is capable of detecting CAPTCHA is served by a website and automatically resumed scraping using different IP addresses. Also here, we manually checked the search results to better understand if we extracted all possible links of the search queries.

To further illustrate the data collection process, we provide an example for an individual artwork. We observe on an artwork by Henry Ossawa Tanner created in 1913 on Useum, which gives us the reference to the artwork as hosted on Wikimedia. Next, we search for the URL to the .JPG Image, and query the URL on the 'Google reverse image search engine'. We collect all links from search results obtained, and can track reuses of the same exact image on various commercial or non-commercial websites, as shown here. This process is automated and repeated for all the artworks contained in our U.S. sample.

This data collection has limitations and bears potential risks for errors. For example, the .JPG link of the digital image we find might not reveal a direct link to the original artwork. So, in some instances, the link refers to a thumbnail of another artwork appearing on the same webpage, which eventually links to another artwork, and so Google searches would return URLs of downstream reuse of that artwork. However, for the vast majority of images, we can observe a Wikimedia link of 'good quality', with relatively 'clean' underlying metadata on artworks and direct links. This makes us confident that the data obtained does not suffer from systematic errors. Applying a more conservative rule, we conclude that in *at least* 95 percent of all search queries, the correct links and unique downstream reuses of artworks were returned.

Eventually, to check the quality of our google reverse image search, we investigated for the duplicates of the google results in our data. If we find many duplicates, this could hint towards a wrong JPG URL source, and as a result, the google image searched for will provide incorrect results. On the one hand, an example for such a incorrect results (that we excluded) was that the original artwork link led (incorrectly) to the main museum page. As a result, the JPG URL search found an incorrect image on the main page, and the google image search was repeated several times for the very same (incorrect) image. On the other hand, in most of the cases, duplicates of google search results can

also be correct, as results directed to e.g. an 'overview' page that listed several artwork images we searched for. An example is the Wikidata sum of all paintings of the Yale University Art Gallery colleciton, the Wikimedia list of Painitngs by Paul Klee, and a myriad number of commercial vendors of art-prints. Overall, we observe the large share of 77 percent of unique Google links, and 95 percent of links appear at a maximum three times in our data set. If we apply a *very* conservative standard of uniqueness of data, we can be confident that at least in 95 percent of our downstream data we can be sure that we found the correct start-image URL and the corresponding 'correct' google reverse image search results.

Overall, we started with a list of 1594 artworks from *Useum*, and obtained for 1349 unique artworks functioning direct .JPG URLs with a total of 88'607 upstream- and downstream-uses and artwork average (median) uses of 66 (42) pages.

World Intellectual Property Organization
34, chemin des Colombettes, P.O. Box 18
CH-1211 Geneva 20, Switzerland

Suggested citation: Cuntz, A, P J Heald and M Sahli (2023), "Digitization and Availability of Artworks in Online Museum Collections", WIPO Economic Research Working Paper No. 75, Geneva: World Intellectual Property Organization.

Cover: WIPO Design